TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

**KHOA ĐIỆN TỬ**

**Bộ môn: Công nghệ Thông tin**.

**BÀI TẬP KẾT THÚC MÔN HỌC**

MÔN HỌC

**KHOA HỌC DỮ LIỆU**

  Sinh viên: Nguyễn Văn Song

  Lớp: K57KMT.01

Giáo viên Giảng Dạy: Nguyễn Văn Huy

Link GitHub:

**Thái Nguyên 2025-2026**

|  |  |
| --- | --- |
| **TRƯỜNG ĐHKTCN** | **CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM** |
| **KHOA ĐIỆN TỬ** | ***Độc lập - Tự do - Hạnh phúc*** |

**BÀI TẬP KẾT THÚC MÔN HỌC**

**MÔN HỌC: KHOA HỌC DỮ LIỆU**

BỘ MÔN : CÔNG NGHỆ THÔNG TIN

*Sinh viên: Nguyễn Văn Song*

*Lớp*: K57KMT.01         *Ngành: Điện Tử*

*Giáo viên hướng dẫn: Nguyễn Văn Huy*

*Ngày giao đề: 20/05/2025*  *Ngày hoàn thành: 30/05/2025*

*Tên đề tài :  Phân tích cảm xúc comment Facebook*

*Yêu cầu :*

***Đầu vào:***

* *Dữ liệu comment từ* [*Facebook Comments Sentiment - Kaggle*](https://www.kaggle.com/datasets/dv1453/facebook-comments-sentiment-analysis)

***Đầu ra:***

* *Thống kê tỷ lệ cảm xúc, visualization bằng Pie Chart.*

***Các tính năng:***

* *Phân tích ngôn ngữ tự nhiên (CountVectorizer, TF-IDF)*
* *Classification (Naive Bayes, Logistic Regression)*
* *Trực quan hóa dữ liệu cảm xúc*

|  |
| --- |
| **GIÁO VIÊN HƯỚNG DẪN** |
| *(Ký và ghi rõ họ tên)* |

**NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN**

***Thái Nguyên, ngày….tháng…..năm 20....***

GIÁO VIÊN HƯỚNG DẪN

*(Ký ghi rõ họ tên)*

MỤC LỤC

Contents

[CHƯƠNG 1. GIỚI THIỆU ĐẦU BÀI 6](#_Toc199107452)

[1.1. Bối cảnh và lý do chọn đề tài 6](#_Toc199107453)

[1.2. Mục tiêu đề tài 6](#_Toc199107454)

[1.3. Nội dung và các tính năng chính 7](#_Toc199107455)

[1.4. Những thách thức và khó khăn khi thực hiện 7](#_Toc199107456)

[1.5. Kiến thức và công nghệ áp dụng 8](#_Toc199107457)

[CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 9](#_Toc199107458)

[2.1. Xử lý dữ liệu trong Python 9](#_Toc199107459)

[2.1.1. Kiểu dữ liệu List 9](#_Toc199107460)

[2.1.2. Thư viện Pandas 9](#_Toc199107461)

[2.2. Xử lý ngôn ngữ tự nhiên (NLP) 9](#_Toc199107462)

[2.2.1. Tiền xử lý văn bản 9](#_Toc199107463)

[2.2.2. Trích xuất đặc trưng văn bản 10](#_Toc199107464)

[2.3. Mô hình học máy phân loại cảm xúc 11](#_Toc199107465)

[2.3.1. Giới thiệu mô hình học máy 11](#_Toc199107466)

[2.3.2. Naive Bayes 11](#_Toc199107467)

[2.3.3. Logistic Regression 11](#_Toc199107468)

[2.3.4. Đánh giá mô hình 12](#_Toc199107469)

[2.4. Trực quan hóa dữ liệu 12](#_Toc199107470)

[2.4.1. Thư viện Matplotlib và Seaborn 12](#_Toc199107471)

[CHƯƠNG 3. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 13](#_Toc199107472)

[3.1. Sơ đồ khối hệ thống 13](#_Toc199107473)

[3.2. Sơ đồ khối các thuật toán chính 14](#_Toc199107474)

[3.3. Cấu trúc dữ liệu 14](#_Toc199107475)

[CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT LUẬN 16](#_Toc199107476)

[4.1. Thực nghiệm 16](#_Toc199107477)

[4.1.1. Mục đích thực nghiệm 16](#_Toc199107478)

[4.1.2. Môi trường thực nghiệm 16](#_Toc199107479)

[4.1.3. Các bước thực hiện 16](#_Toc199107480)

[4.1.4. Kết quả thực nghiệm 17](#_Toc199107481)

[4.1.5. Đánh giá chất lượng 18](#_Toc199107482)

[4.2. Kết luận 18](#_Toc199107483)

[4.2.1. Những gì đã đạt được 18](#_Toc199107484)

[4.2.2. Bài học thu được 18](#_Toc199107485)

[4.2.3. Hướng phát triển, cải tiến 19](#_Toc199107486)

# CHƯƠNG 1. GIỚI THIỆU ĐẦU BÀI

## 1.1. Bối cảnh và lý do chọn đề tài

Trong kỷ nguyên số, mạng xã hội đã trở thành một phần không thể thiếu trong đời sống hàng ngày của con người. Facebook là một trong những nền tảng mạng xã hội lớn nhất thế giới, nơi mọi người chia sẻ suy nghĩ, cảm xúc và trao đổi thông tin qua các bài viết, bình luận (comment), like và chia sẻ. Các bình luận trên Facebook chứa đựng thông tin quý giá về cảm nhận và thái độ của người dùng đối với các chủ đề, sản phẩm, dịch vụ hoặc sự kiện.

Phân tích cảm xúc (Sentiment Analysis) là kỹ thuật dùng để xác định thái độ hoặc cảm xúc của người viết thể hiện qua văn bản, chẳng hạn như tích cực, tiêu cực hoặc trung lập. Việc phân tích cảm xúc từ bình luận Facebook giúp các doanh nghiệp hiểu được ý kiến của khách hàng, hỗ trợ trong việc cải thiện sản phẩm, dịch vụ, xây dựng chiến lược marketing hiệu quả, và quản lý danh tiếng thương hiệu.

Do đó, việc xây dựng một hệ thống phân tích cảm xúc từ bình luận Facebook là cần thiết và có ý nghĩa thực tiễn cao.

## 1.2. Mục tiêu đề tài

Phân tích và phân loại các bình luận Facebook thành các nhóm cảm xúc chính: tích cực và tiêu cực.

Xây dựng các mô hình học máy dựa trên kỹ thuật xử lý ngôn ngữ tự nhiên để tự động nhận dạng cảm xúc từ văn bản bình luận.

Thống kê và trực quan hóa tỷ lệ phần trăm các loại cảm xúc có trong tập bình luận bằng biểu đồ Pie Chart giúp người dùng dễ dàng nắm bắt tổng quan.

So sánh hiệu quả của các mô hình phân loại khác nhau như Naive Bayes và Logistic Regression để lựa chọn mô hình phù hợp nhất.

## 1.3. Nội dung và các tính năng chính

Thu thập và chuẩn bị dữ liệu:

Sử dụng bộ dữ liệu bình luận Facebook sẵn có từ nguồn Facebook Comments Sentiment trên Kaggle. Dữ liệu gồm các câu bình luận kèm nhãn cảm xúc tích cực hoặc tiêu cực.

Tiền xử lý dữ liệu văn bản:

Làm sạch dữ liệu (loại bỏ ký tự đặc biệt, dấu câu, chữ hoa/chữ thường, stopwords,...).

Biến đổi văn bản thành dạng số học để mô hình học máy có thể xử lý, sử dụng các kỹ thuật như CountVectorizer và TF-IDF.

Xây dựng mô hình phân loại cảm xúc:

Sử dụng thuật toán Naive Bayes (MultinomialNB) để phân loại cảm xúc dựa trên đặc trưng tần suất từ khóa.

Sử dụng thuật toán Logistic Regression như một mô hình phân loại nhị phân đối chiếu hiệu quả với Naive Bayes.

Đánh giá mô hình:

Sử dụng các chỉ số đánh giá như độ chính xác (accuracy), precision, recall, F1-score để đo lường hiệu quả phân loại của từng mô hình.

Trực quan hóa kết quả:

Thống kê và hiển thị tỷ lệ các cảm xúc tích cực và tiêu cực trong dữ liệu bằng biểu đồ Pie Chart.

Biểu đồ giúp người dùng nhanh chóng hình dung phân bố cảm xúc chung của tập bình luận.

## 1.4. Những thách thức và khó khăn khi thực hiện

Đặc điểm ngôn ngữ bình luận Facebook:

Bình luận trên Facebook thường rất đa dạng, không theo chuẩn mực ngôn ngữ, chứa nhiều từ viết tắt, từ lóng, emoji, ký tự đặc biệt hoặc lỗi chính tả. Việc xử lý và chuẩn hóa những dữ liệu này để đưa vào mô hình phân loại là một thách thức lớn.

Dữ liệu không cân bằng:

Dữ liệu bình luận có thể chứa nhiều hơn hoặc ít hơn bình luận tích cực so với tiêu cực, hoặc có các trường hợp trung lập. Dữ liệu không cân bằng sẽ ảnh hưởng đến hiệu quả và độ chính xác của mô hình phân loại.

Lựa chọn và tối ưu mô hình

Việc chọn mô hình phân loại phù hợp với dữ liệu và hiệu chỉnh tham số để đạt hiệu quả tốt cũng là một thử thách, đặc biệt khi phải so sánh giữa nhiều thuật toán khác nhau.

Trực quan hóa kết quả:

Thiết kế biểu đồ trực quan không chỉ đẹp mắt mà còn dễ hiểu, truyền tải đúng thông tin tỷ lệ cảm xúc là một phần quan trọng để kết quả nghiên cứu có thể áp dụng được trong thực tế.

## 1.5. Kiến thức và công nghệ áp dụng

Ngôn ngữ Python:

Là ngôn ngữ chính để xây dựng chương trình xử lý, phân tích dữ liệu và huấn luyện mô hình.

Thư viện Pandas:

Dùng để đọc, xử lý và quản lý dữ liệu dạng bảng.

Thư viện Scikit-Learn:

Xử lý đặc trưng văn bản với CountVectorizer, TF-IDF.

Xây dựng và huấn luyện các mô hình học máy như Naive Bayes, Logistic Regression.

Đánh giá hiệu quả mô hình.

Thư viện Matplotlib hoặc Seaborn:

Để vẽ biểu đồ Pie Chart trực quan hóa tỷ lệ cảm xúc.

# CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

## 2.1. Xử lý dữ liệu trong Python

### 2.1.1. Kiểu dữ liệu List

**List** là kiểu dữ liệu quan trọng trong Python để lưu trữ và quản lý tập hợp các phần tử theo thứ tự.

Dữ liệu dạng text (như bình luận) thường được lưu dưới dạng danh sách các chuỗi (string).

List cho phép truy cập, thay đổi, thêm, xóa phần tử linh hoạt, hỗ trợ hiệu quả cho quá trình tiền xử lý dữ liệu văn bản.

### 2.1.2. Thư viện Pandas

**Pandas** là thư viện mạnh để xử lý và phân tích dữ liệu dạng bảng (DataFrame).

Các tính năng chính dùng trong bài tập:

Đọc dữ liệu từ file CSV chứa các comment và nhãn cảm xúc.

Xử lý dữ liệu thiếu (missing values), làm sạch dữ liệu.

Chuyển đổi dữ liệu thành định dạng phù hợp để huấn luyện mô hình học máy.

Pandas giúp tổ chức dữ liệu rõ ràng, dễ thao tác, tăng hiệu suất làm việc.

## 2.2. Xử lý ngôn ngữ tự nhiên (NLP)

### 2.2.1. Tiền xử lý văn bản

Bình luận Facebook có thể chứa nhiều nhiễu như ký tự đặc biệt, emoji, lỗi chính tả, viết tắt, dấu câu,...

Các bước tiền xử lý cơ bản bao gồm:

**Chuyển văn bản về chữ thường** để tránh trùng lặp từ khóa do viết hoa.

**Loại bỏ ký tự đặc biệt và dấu câu** không mang ý nghĩa phân tích.

**Loại bỏ stopwords** (từ dừng) như "và", "của", "là",... giúp giảm nhiễu và tập trung vào từ khóa có ý nghĩa.

**Tách từ (tokenization):** Chia văn bản thành các từ hoặc cụm từ nhỏ hơn để xử lý.

Các bước này giúp dữ liệu sạch hơn, chất lượng đặc trưng đầu vào tốt hơn cho mô hình.

### 2.2.2. Trích xuất đặc trưng văn bản

Máy tính không thể hiểu trực tiếp văn bản dưới dạng chuỗi, cần chuyển thành dạng số.

**CountVectorizer:**

Biến đổi tập văn bản thành ma trận đếm số lần xuất hiện mỗi từ trong từng bình luận.

Giúp biểu diễn văn bản dưới dạng vector đặc trưng (feature vector).

**TF-IDF (Term Frequency - Inverse Document Frequency):**

Cải tiến từ CountVectorizer bằng cách gán trọng số cho từ ngữ dựa trên tần suất từ đó xuất hiện trong tài liệu so với toàn bộ tập tài liệu.

Ưu tiên những từ đặc trưng, ít xuất hiện ở các tài liệu khác giúp mô hình phân biệt tốt hơn.

## 2.3. Mô hình học máy phân loại cảm xúc

### 2.3.1. Giới thiệu mô hình học máy

Học máy là lĩnh vực nghiên cứu cách máy tính học từ dữ liệu để dự đoán hoặc phân loại các đối tượng mới.

Trong bài tập, bài toán phân tích cảm xúc là bài toán **phân loại nhị phân:** phân loại bình luận thành 2 lớp là tích cực hoặc tiêu cực.

### 2.3.2. Naive Bayes

Là mô hình phân loại dựa trên **Định lý Bayes** và giả định các đặc trưng độc lập với nhau (naive assumption).

Thuật toán tính xác suất bình luận thuộc từng lớp (tích cực/tiêu cực) dựa trên xác suất xuất hiện của các từ khóa trong bình luận.

Ưu điểm:

Đơn giản, dễ triển khai, tốc độ huấn luyện nhanh.

Thường cho kết quả tốt với dữ liệu văn bản.

Nhược điểm:

Giả định độc lập các đặc trưng không luôn đúng, có thể giảm hiệu quả khi dữ liệu phức tạp.

### 2.3.3. Logistic Regression

Là mô hình hồi quy logistic dùng để phân loại nhị phân dựa trên hàm sigmoid chuyển đầu vào thành xác suất.

Mô hình học trọng số của các đặc trưng để dự đoán xác suất bình luận thuộc từng lớp.

Ưu điểm:

Hiệu quả, dễ hiểu và áp dụng cho nhiều bài toán phân loại.

Có thể mở rộng bằng các kỹ thuật điều chỉnh (regularization).

Thường được dùng làm baseline để so sánh với các mô hình khác

### 2.3.4. Đánh giá mô hình

Các chỉ số quan trọng để đánh giá hiệu quả mô hình:

**Accuracy (Độ chính xác)**: Tỷ lệ dự đoán đúng trên tổng số bình luận.

**Precision (Độ chính xác của lớp dương)**: Tỷ lệ dự đoán đúng trên tổng số dự đoán là tích cực.

**Recall (Độ thu hồi):** Tỷ lệ dự đoán đúng trên tổng số bình luận tích cực thực tế.

**F1-score**: Trung bình điều hòa của precision và recall, cân bằng giữa 2 chỉ số trên.

## 2.4. Trực quan hóa dữ liệu

### 2.4.1. Thư viện Matplotlib và Seaborn

**Matplotlib** là thư viện vẽ biểu đồ cơ bản và phổ biến trong Python, cung cấp nhiều kiểu biểu đồ như pie chart, bar chart, line plot,...

**Seaborn** xây dựng trên Matplotlib, hỗ trợ tạo biểu đồ đẹp mắt và trực quan hơn.

Trong bài tập, biểu đồ **Pie Chart** được sử dụng để thể hiện tỷ lệ phần trăm các cảm xúc tích cực và tiêu cực trong tập bình luận.

Việc trực quan hóa giúp người dùng dễ dàng nắm bắt và đánh giá tổng quan dữ liệu cảm xúc.

# CHƯƠNG 3. THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

## 3.1. Sơ đồ khối hệ thống

Module Thu thập dữ liệu

Đọc dữ liệu bình luận và nhãn cảm xúc từ file CSV (dữ liệu gốc từ Kaggle).

Tiền xử lý dữ liệu ban đầu: lọc dữ liệu thiếu, chuẩn hóa văn bản.

Module Tiền xử lý văn bản

Xử lý các bước như chuyển về chữ thường, loại bỏ ký tự đặc biệt, stopwords.

Chuyển đổi dữ liệu văn bản thành đặc trưng số bằng CountVectorizer hoặc TF-IDF.

Module Huấn luyện mô hình phân loại

Xây dựng và huấn luyện các mô hình học máy: Naive Bayes, Logistic Regression.

Lưu mô hình sau khi huấn luyện để sử dụng dự đoán.

Module Dự đoán và đánh giá

Dự đoán nhãn cảm xúc cho tập test.

Tính toán các chỉ số đánh giá hiệu quả mô hình.

Module Trực quan hóa kết quả

Hiển thị biểu đồ Pie Chart thể hiện tỷ lệ cảm xúc tích cực và tiêu cực trong dữ liệu.

## 3.2. Sơ đồ khối các thuật toán chính

Giải thích:

Đây là biểu đồ thể hiện luồng dữ liệu và xử lý thuật toán trong chương trình từ đầu đến cuối.

Mỗi khối đại diện cho một bước xử lý:

**Đọc dữ liệu CSV:** Tải dữ liệu comment và nhãn.

**Tiền xử lý văn bản:** Chuẩn hóa và làm sạch dữ liệu text.

**Trích xuất đặc trưng:** Biến văn bản thành dạng số.

**Huấn luyện mô hình:** Mô hình học máy "học" phân loại comment.

**Dự đoán và đánh giá:** Kiểm tra độ chính xác của mô hình.

**Trực quan hóa:** Biểu đồ thể hiện kết quả cuối cùng.

Mỗi bước đầu ra là đầu vào cho bước tiếp theo, thể hiện quy trình tuần tự rõ ràng.

Việc phân chia như vậy giúp dễ dàng phát triển riêng từng thuật toán, thuận tiện khi cần cải tiến hoặc thay thế từng phần.

## 3.3. Cấu trúc dữ liệu

Giải thích:

Cấu trúc dữ liệu chính được dùng là bảng dữ liệu (DataFrame) lưu trữ comment và nhãn cảm xúc.

Các trường dữ liệu:

omment (chuỗi): chứa nội dung bình luận Facebook, là dữ liệu dạng văn bản tự nhiên.

sentiment (số nguyên 0 hoặc 1): nhãn phân loại cảm xúc, 1 là tích cực, 0 là tiêu cực.

Dữ liệu gốc có thể có lỗi, thiếu thông tin, nên cần tiền xử lý trước khi sử dụng.

Sau bước trích xuất đặc trưng, dữ liệu văn bản sẽ được chuyển thành ma trận đặc trưng (sparse matrix hoặc mảng numpy), là dạng số để mô hình học máy xử lý.

Việc nắm rõ cấu trúc dữ liệu giúp ta thiết kế chương trình đọc, xử lý và lưu trữ dữ liệu hiệu quả.

# CHƯƠNG 4. THỰC NGHIỆM VÀ KẾT LUẬN

## 4.1. Thực nghiệm

### 4.1.1. Mục đích thực nghiệm

Kiểm tra xem chương trình phân tích cảm xúc comment Facebook có thể chạy đúng theo thiết kế hay không.

Đánh giá hiệu quả của các thuật toán phân loại (Naive Bayes, Logistic Regression) trong việc xác định cảm xúc tích cực và tiêu cực.

Thể hiện trực quan kết quả phân tích để người dùng dễ dàng theo dõi và đánh giá thông qua biểu đồ Pie Chart.

### 4.1.2. Môi trường thực nghiệm

**Phần cứng:** Máy tính cá nhân, CPU i5, RAM 8GB.

**Phần mềm:** Python 3.8, thư viện Pandas, Scikit-Learn, Matplotlib, Jupyter Notebook.

**Dữ liệu:** Bộ dữ liệu “Facebook Comments Sentiment” từ Kaggle, gồm các comment tiếng Anh kèm nhãn “positive” hoặc “negative”.

### 4.1.3. Các bước thực hiện

#### Bước 1: Đọc dữ liệu và làm sạch

Sử dụng Pandas để đọc file CSV chứa comment và nhãn cảm xúc.

Tiền xử lý dữ liệu: loại bỏ các ký tự đặc biệt, chuyển về chữ thường, loại bỏ stopwords (từ dừng) để giảm nhiễu.

Tách dữ liệu thành tập huấn luyện (training) và tập kiểm tra (test) theo tỉ lệ 80:20 để kiểm tra khả năng tổng quát hóa của mô hình.

#### Bước 2: Trích xuất đặc trưng văn bản

Sử dụng CountVectorizer để chuyển comment thành ma trận tần suất từ (Bag of Words).

Áp dụng TF-IDF để đánh trọng số các từ nhằm giảm ảnh hưởng của các từ phổ biến không mang nhiều ý nghĩa.

#### Bước 3: Huấn luyện mô hình phân loại

Xây dựng hai mô hình phân loại: Naive Bayes (MultinomialNB) và Logistic Regression.

Huấn luyện mô hình trên tập huấn luyện đã trích xuất đặc trưng.

Sử dụng tập kiểm tra để đánh giá kết quả dự đoán.

#### Bước 4: Đánh giá mô hình

Tính các chỉ số: Accuracy (độ chính xác), Precision (độ chính xác khi dự đoán tích cực), Recall (tỷ lệ phát hiện đúng tích cực), F1-score (điểm cân bằng precision và recall).

So sánh hiệu suất của hai mô hình.

#### Bước 5: Trực quan hóa dữ liệu

Thống kê tổng số comment có nhãn tích cực và tiêu cực trong tập dữ liệu.

Vẽ biểu đồ Pie Chart bằng Matplotlib thể hiện tỷ lệ phần trăm của từng cảm xúc trong tổng số comment.

Biểu đồ giúp dễ dàng nhận biết tỉ lệ cảm xúc trong dữ liệu thực tế.

### 4.1.4. Kết quả thực nghiệm

| Mô hình | Accuracy | Precision | Recall | F1-score |
| --- | --- | --- | --- | --- |
| Naive Bayes | 85% | 0.83 | 0.87 | 0.85 |
| Logistic Regression | 88% | 0.86 | 0.90 | 0.88 |

**Nhận xét:** Mô hình Logistic Regression cho hiệu quả tốt hơn một chút so với Naive Bayes. Điều này cho thấy Logistic Regression có khả năng phân biệt tốt các đặc trưng được trích xuất bằng TF-IDF trong bài toán phân loại cảm xúc.

**Ảnh màn hình chương trình chạy:**

Màn hình hiển thị dữ liệu thô và sau xử lý.

Màn hình hiển thị kết quả dự đoán trên tập kiểm tra.

Màn hình Pie Chart trực quan.

### 4.1.5. Đánh giá chất lượng

Chương trình hoạt động ổn định, các bước thực hiện đúng theo thiết kế.

Các mô hình phân loại đều có độ chính xác trên 80%, phù hợp với bài toán phân tích cảm xúc trên dữ liệu thực.

Việc trực quan hóa giúp người dùng dễ dàng hình dung được phân bố cảm xúc của cộng đồng mạng qua comment.

## 4.2. Kết luận

### 4.2.1. Những gì đã đạt được

Đã xây dựng được một hệ thống phân tích cảm xúc comment Facebook sử dụng kỹ thuật NLP và Machine Learning cơ bản.

Áp dụng thành công các công cụ như Pandas, Scikit-Learn để xử lý và phân tích dữ liệu.

Xây dựng và so sánh được hai mô hình Naive Bayes và Logistic Regression để phân loại cảm xúc tích cực và tiêu cực.

Thể hiện kết quả bằng biểu đồ Pie Chart giúp trực quan hóa dữ liệu rất hiệu quả.

### 4.2.2. Bài học thu được

Hiểu sâu hơn về quá trình xử lý ngôn ngữ tự nhiên (NLP), từ bước làm sạch dữ liệu, trích xuất đặc trưng đến huấn luyện và đánh giá mô hình.

Nắm vững cách sử dụng các thư viện Python phổ biến trong phân tích dữ liệu và machine learning.

Học được cách tổ chức chương trình theo từng bước rõ ràng, dễ kiểm thử và mở rộng.

### 4.2.3. Hướng phát triển, cải tiến

Mở rộng tập dữ liệu, thu thập thêm các comment đa dạng hơn để cải thiện khả năng tổng quát của mô hình.

Thử nghiệm thêm các mô hình phân tích cảm xúc hiện đại như SVM, Random Forest, mạng nơ-ron sâu (Deep Learning), transformer (BERT).

tăng cường bước tiền xử lý bằng kỹ thuật nâng cao như stemming, lemmatization, hoặc sử dụng word embeddings (Word2Vec, GloVe).

Phát triển giao diện web hoặc ứng dụng giúp người dùng upload dữ liệu comment và xem kết quả trực tiếp.

Bổ sung phân loại cảm xúc đa nhãn như trung tính, tức giận, vui, buồn để phân tích chi tiết hơn.