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 Hải

  Lớp: K57KMT

Giáo viên giảng dạy: Nguyễn Văn Huy

Link GitHub:

**THÁI NGUYÊN 2025**

|  |  |
| --- | --- |
| **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 Hải*

*Lớp*: *K57KMT*…………………….  *Ngành: Kỹ thuật máy tính*

*Giáo viên hướng dẫn: Nguyễn Văn Huy*

*Ngày giao đề:20-5-2025*  *Ngày hoàn thành:30-5-2025*

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

*Yêu cầu :* *Phân tích comment Facebook để xác định cảm xúc tích cực, tiêu cực.*

**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 và ghi rõ họ tên)*

MỤC LỤC

[CHƯƠNG 1: GIỚI THIỆU 2](#_Toc199458162)

[1.1. Mô tả bài toán 2](#_Toc199458163)

[1.2. Tính năng của chương trình 2](#_Toc199458164)

[1.3. Thách thức và kiến thức vận dụng 2](#_Toc199458165)

[1.3.1. Thách thức 2](#_Toc199458166)

[1.3.2. Kiến thức vận dụng 3](#_Toc199458167)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 4](#_Toc199458168)

[2.1. Xử lý dữ liệu với Pandas 4](#_Toc199458169)

[2.2. Biểu diễn văn bản với TF-IDF 4](#_Toc199458170)

[2.4. Trực quan hóa với Matplotlib 5](#_Toc199458171)

[CHƯƠNG 3: THIẾT KẾ VÀ TRIỂN KHAI 6](#_Toc199458172)

[3.1. Thiết kế hệ thống 6](#_Toc199458173)

[3.2. Thuật toán chính 7](#_Toc199458174)

[3.3. Cấu trúc dữ liệu 7](#_Toc199458175)

[3.4. Triển khai chương trình 7](#_Toc199458176)

[CHƯƠNG 4: KẾT QUẢ VÀ KẾT LUẬN 9](#_Toc199458177)

[4.1. Kết quả thực nghiệm 9](#_Toc199458178)

[4.1.1. Phân tích cảm xúc tổng thể 9](#_Toc199458179)

[4.1.2. Đánh giá mô hình 9](#_Toc199458180)

[4.2. Kết luận và hướng phát triển 10](#_Toc199458181)

[4.2.1. Kết quả đạt được 10](#_Toc199458182)

[4.2.2. Bài học 10](#_Toc199458183)

[4.2.3. Hướng phát triển 10](#_Toc199458184)

[TÀI LIỆU THAM KHẢO 12](#_Toc199458185)

# CHƯƠNG 1: GIỚI THIỆU

## 1.1. Mô tả bài toán

Bài toán yêu cầu xây dựng một hệ thống phân tích cảm xúc từ bình luận Facebook để xác định phản ứng của người dùng đối với một bài đăng, phân loại thành ba nhóm: tiêu cực, tích cực, hoặc trung tính. Hệ thống sử dụng tập dữ liệu từ Kaggle (*Facebook Comments Sentiment Dataset*), chứa các bình luận kèm nhãn cảm xúc.

Quy trình xử lý bao gồm:

* **Tiền xử lý văn bản:** Làm sạch bình luận bằng cách loại bỏ ký tự đặc biệt, chuyển chữ thường, và chuẩn hóa dữ liệu.
* **Phân tích cảm xúc**: Áp dụng kỹ thuật học máy với mô hình Logistic Regression và biểu diễn văn bản bằng TF-IDF để phân loại cảm xúc.
* **Thống kê và trực quan hóa:** Đếm số lượng, tính tỷ lệ phần trăm các loại cảm xúc, và hiển thị kết quả qua biểu đồ tròn.

Mục tiêu là cung cấp cái nhìn tổng quan về cảm xúc của cộng đồng mạng đối với một bài đăng cụ thể.

## 1.2. Tính năng của chương trình

Chương trình bao gồm ba tính năng chính:

* **Xử lý văn bản:** Làm sạch bình luận bằng cách chuyển chữ thường và loại bỏ ký tự đặc biệt, chuẩn bị dữ liệu cho phân tích.
* **Phân tích cảm xúc:** Phân loại bình luận thành tiêu cực (phàn nàn), tích cực (khen ngợi), hoặc trung tính (không rõ thái độ) bằng Logistic Regression và TF-IDF.
* **Trực quan hóa:** Hiển thị tỷ lệ cảm xúc qua biểu đồ tròn, giúp người dùng dễ dàng nắm bắt xu hướng cảm xúc.

## 1.3. Thách thức và kiến thức vận dụng

### 1.3.1. Thách thức

**Ngôn ngữ không chuẩn:** Bình luận có thể chứa tiếng lóng, viết tắt, lỗi chính tả, hoặc ngữ nghĩa mỉa mai, gây khó khăn cho phân loại.

**Dữ liệu nhiễu:** Một số bình luận không liên quan (spam) hoặc quá ít dữ liệu có thể làm giảm ý nghĩa thống kê.

**Lựa chọn mô hình:** Mô hình đơn giản (như Logistic Regression) có thể thiếu chính xác, trong khi mô hình phức tạp (như BERT) đòi hỏi tài nguyên lớn.

**Trực quan hóa:** Cần biểu đồ rõ ràng, không gây nhầm lẫn.

### 1.3.2. Kiến thức vận dụng

* **Xử lý ngôn ngữ tự nhiên (NLP):** Làm sạch văn bản, chuyển chữ thường, loại bỏ ký tự đặc biệt.
* **Học máy:** Sử dụng Logistic Regression và TF-IDF để phân loại cảm xúc.
* **Thư viện:**
* pandas: Xử lý dữ liệu dạng bảng.
* scikit-learn: Biểu diễn văn bản, huấn luyện và đánh giá mô hình.
* matplotlib: Trực quan hóa dữ liệu qua biểu đồ tròn.
* re: Xử lý biểu thức chính quy.

*Tóm tắt chương:*

*Chương 1 giới thiệu bài toán phân tích cảm xúc bình luận Facebook, bao gồm tiền xử lý, phân loại cảm xúc, và trực quan hóa. Hệ thống gặp thách thức về ngôn ngữ, dữ liệu, và mô hình, đòi hỏi kiến thức NLP, học máy, và các thư viện như Pandas, scikit-learn, Matplotlib.*

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Xử lý dữ liệu với Pandas

Pandas là thư viện Python dùng để thao tác dữ liệu dạng bảng (DataFrame). Trong chương trình:

* **Đọc dữ liệu:** Hàm read\_csv() đọc tệp facebook\_comments.csv chứa cột comment (nội dung bình luận) và label (nhãn cảm xúc: 0 - tiêu cực, 1 - tích cực, 2 - trung tính).
* **Kiểm tra dữ liệu:** Hàm head() hiển thị 5 dòng đầu tiên để kiểm tra cấu trúc.
* **Tiền xử lý:** Tạo cột van\_ban\_sach bằng cách áp dụng hàm làm sạch văn bản, chuẩn bị dữ liệu cho phân tích.

Pandas đảm bảo dữ liệu được quản lý hiệu quả, là bước đầu tiên trong quy trình.

## 2.2. Biểu diễn văn bản với TF-IDF

Kỹ thuật TF-IDF (Term Frequency-Inverse Document Frequency) chuyển đổi văn bản thành dạng số để mô hình học máy có thể xử lý.

* **Nguyên lý:** TF-IDF đánh giá tầm quan trọng của từ dựa trên tần suất xuất hiện (TF) và mức độ hiếm trong tập dữ liệu (IDF), làm nổi bật các từ mang ý nghĩa cảm xúc.
* **Ứng dụng:** Sử dụng TfidfVectorizer từ scikit-learn để tạo ma trận đặc trưng (X) từ cột van\_ban\_sach, với nhãn cảm xúc lưu trong y.
* **Lợi ích:** Giảm ảnh hưởng của từ phổ biến, cải thiện hiệu quả phân loại.

**2.3. Phân tích cảm xúc với Logistic Regression**

Logistic Regression là mô hình học máy phân loại cảm xúc của bình luận.

* **Nguyên lý:** Dự đoán xác suất một bình luận thuộc lớp cảm xúc (tiêu cực, tích cực, trung tính) dựa trên đặc trưng TF-IDF, sử dụng hàm mất mát log-loss và tối ưu bằng thuật toán lbfgs (max\_iter=1000).
* **Quy trình:** 
  + Chia dữ liệu thành tập huấn luyện (80%) và kiểm tra (20%) bằng train\_test\_split (random\_state=42).
  + Huấn luyện mô hình trên tập huấn luyện.
  + Đánh giá bằng classification\_report với các chỉ số precision, recall, F1-score.
* Ưu điểm: Đơn giản, hiệu quả cho bài toán phân loại đa lớp.

## 2.4. Trực quan hóa với Matplotlib

Matplotlib được sử dụng để tạo biểu đồ tròn hiển thị tỷ lệ cảm xúc.

**Chức năng:** Hiển thị phần trăm các loại cảm xúc (tiêu cực, tích cực, trung tính) với màu sắc rõ ràng (lightcoral, lightgreen, lightgray), tiêu đề “Tỷ lệ cảm xúc trong comment Facebook”, và định dạng tỷ lệ bằng autopct='%1.1f%%'.

**Định dạng:** Kích thước (6,6), sử dụng axis('equal') để đảm bảo hình tròn cân đối.

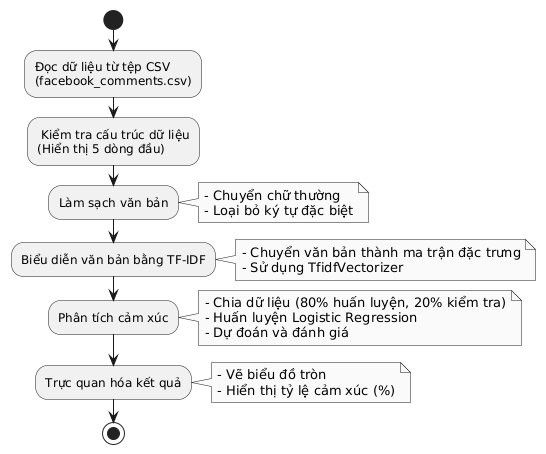
*Tóm tắt chương:*

*Chương 2 trình bày các công nghệ cốt lõi: Pandas xử lý dữ liệu, TF-IDF biểu diễn văn bản, Logistic Regression phân loại cảm xúc, và Matplotlib trực quan hóa kết quả. Các kỹ thuật này tạo nền tảng cho hệ thống phân tích cảm xúc.*

# CHƯƠNG 3: THIẾT KẾ VÀ TRIỂN KHAI

## 3.1. Thiết kế hệ thống

Hệ thống gồm các bước chính, được minh họa trong



Hình 3.1: Sơ đồ khối hệ thống phân tích cảm xúc

**Đọc dữ liệu:** Nhập tệp CSV chứa bình luận và nhãn cảm xúc.

**Làm sạch văn bản:** Chuẩn hóa bình luận bằng cách chuyển chữ thường, loại bỏ ký tự đặc biệt.

**Biểu diễn TF-IDF:** Chuyển văn bản thành ma trận đặc trưng.

**Phân loại cảm xúc:** Sử dụng Logistic Regression để dự đoán cảm xúc.

**Trực quan hóa:** Hiển thị tỷ lệ cảm xúc qua biểu đồ tròn.

## 3.2. Thuật toán chính

Làm sạch văn bản:

* Input: Chuỗi văn bản từ cột comment.
* Quy trình: Chuyển chữ thường, loại bỏ ký tự đặc biệt bằng re.sub.
* Output: Văn bản sạch, lưu trong cột van\_ban\_sach.

Phân tích cảm xúc:

* nput: Ma trận TF-IDF (X), nhãn cảm xúc (y).
* Quy trình: Chia dữ liệu, huấn luyện Logistic Regression, dự đoán và đánh giá.
* Output: Nhãn cảm xúc (0, 1, 2).

Trực quan hóa:

* Input: Số lượng bình luận theo cảm xúc.
* Quy trình: Vẽ biểu đồ tròn bằng plt.pie().
* Output: Biểu đồ hiển thị tỷ lệ cảm xúc.

## 3.3. Cấu trúc dữ liệu

Dữ liệu đầu vào:

* Tệp: facebook\_comments.csv.
* Cột: comment (chuỗi), label (số: 0, 1, 2).
* Dữ liệu sau xử lý:
* Cột mới: van\_ban\_sach (văn bản sạch).
* Ma trận: X (TF-IDF), y (nhãn cảm xúc).

## 3.4. Triển khai chương trình

Chương trình được viết bằng Python, sử dụng các thư viện:

* **pandas:** Đọc và xử lý dữ liệu.
* **scikit-learn:** Biểu diễn TF-IDF, huấn luyện Logistic Regression, đánh giá mô hình.
* **matplotlib:** Vẽ biểu đồ tròn.
* **re:** Làm sạch văn bản.

*Tóm tắt chương:*

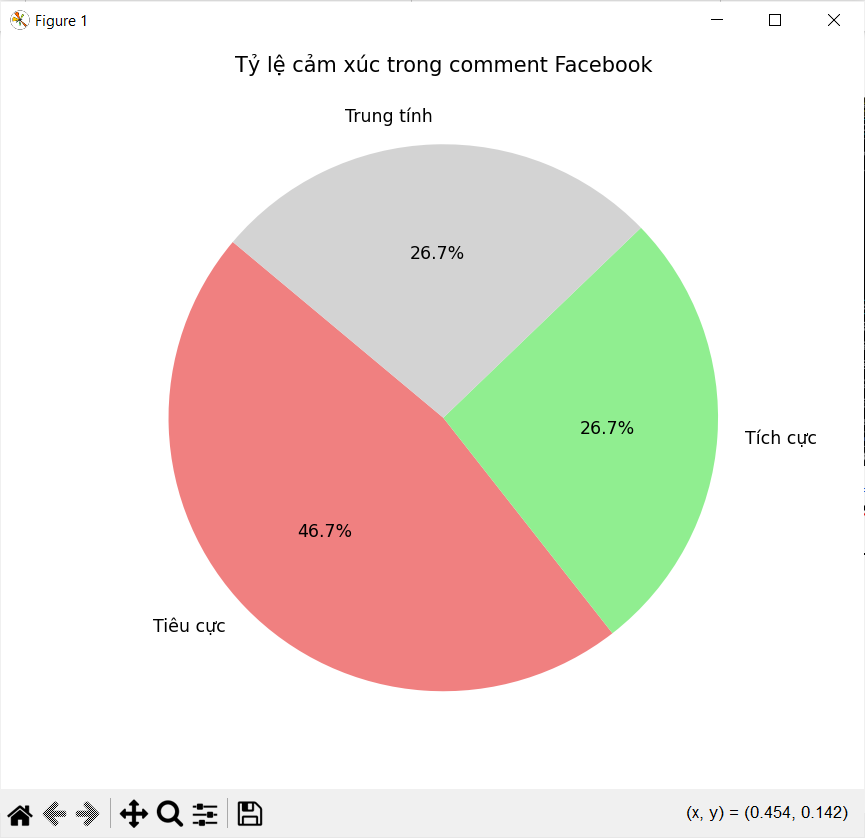
*Chương 3 trình bày thiết kế và triển khai hệ thống phân tích cảm xúc, từ đọc dữ liệu, làm sạch văn bản, biểu diễn TF-IDF, phân loại bằng Logistic Regression, đến trực quan hóa. Mã nguồn sử dụng Pandas, scikit-learn, Matplotlib, và re, đảm bảo quy trình rõ ràng và hiệu quả.*

# CHƯƠNG 4: KẾT QUẢ VÀ KẾT LUẬN

## 4.1. Kết quả thực nghiệm

### 4.1.1. Phân tích cảm xúc tổng thể

Chương trình dự đoán cảm xúc cho toàn bộ bình luận và tính tỷ lệ phần trăm.



Hình 4.1 hiển thị tỷ lệ cảm xúc.

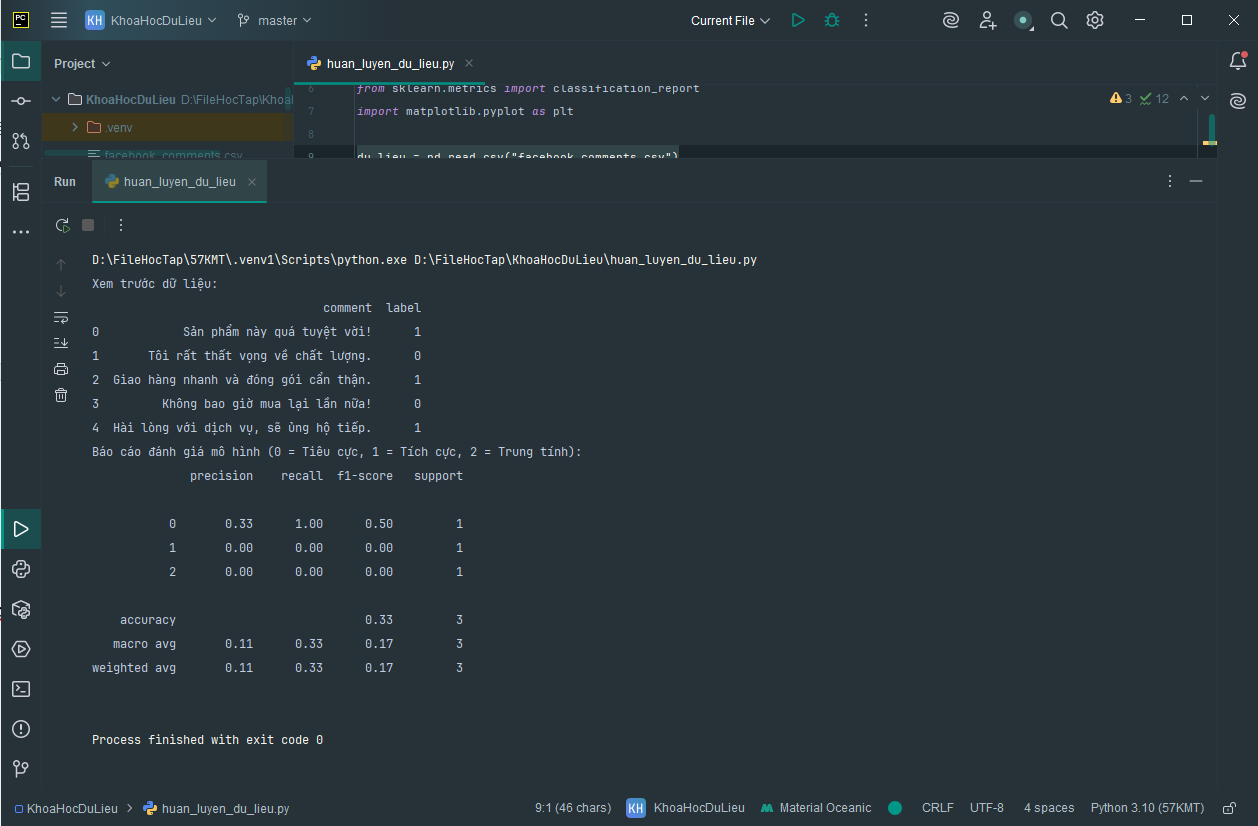
Đơn vị: Phần trăm (%).

* Tiêu cực: chiếm 46.7%
* Tích cực: chiếm 26.7%
* Trung tính:chiếm 26.7%

Màu sắc: Tiêu cực (đỏ nhạt), Tích cực (xanh nhạt), Trung tính (xám nhạt).

### 4.1.2. Đánh giá mô hình

Đánh giá Logistic Regression trên tập kiểm tra bằng classification\_report.



Bảng 4.1: Báo cáo đánh giá mô hình

## 4.2. Kết luận và hướng phát triển

### 4.2.1. Kết quả đạt được

* Xử lý dữ liệu bình luận từ tệp CSV bằng Pandas.
* Làm sạch văn bản, biểu diễn bằng TF-IDF, và phân loại cảm xúc bằng Logistic Regression.
* Trực quan hóa tỷ lệ cảm xúc qua biểu đồ tròn, dễ hiểu.
* Đánh giá mô hình với các chỉ số precision, recall, F1-score.
* Mã nguồn rõ ràng, dễ mở rộng.

### 4.2.2. Bài học

* Thành thạo xử lý dữ liệu với Pandas và biểu diễn văn bản bằng TF-IDF.
* Hiểu quy trình huấn luyện và đánh giá mô hình học máy.
* Rèn luyện kỹ năng trực quan hóa dữ liệu với Matplotlib.
* Nhận ra hạn chế của tiền xử lý văn bản cơ bản cho tiếng Việt.

### 4.2.3. Hướng phát triển

* **Kỹ thuật:** Tích hợp underthesea để tách từ và loại bỏ stopwords tiếng Việt; thử nghiệm Naive Bayes hoặc BERT.
* **Giao diện:** Xây dựng ứng dụng Streamlit để nhập bình luận và hiển thị kết quả trực tiếp.

*Tóm tắt chương:*

*Chương 4 trình bày kết quả phân tích cảm xúc (biểu đồ tròn) và đánh giá mô hình (báo cáo hiệu suất). Chương trình đáp ứng mục tiêu, nhưng cần cải thiện xử lý tiếng Việt và giao diện người dùng.*

TÀI LIỆU THAM KHẢO

https://www.kaggle.com/

https://chatgpt.com/