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: . . . . Đặng Phương Nam . . . . . . . . . . . . . . . . . .

Lớp: . . . . . . . . . 57KMT.01. . . . . . . . . . . . . . . . . . . . . . . .

Giáo viên GIẢNG DẠY: . . Ts. Nguyễn Văn Huy. . . . . . .

Link GitHub: https://github.com/Dang-Nam/btl\_KHDL

**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: Đặng Phương Nam*

*Lớp: 57KMT.01 Ngành: Kỹ thuật máy tính*

*Giáo viên hướng dẫn: Ts. Nguyễn Văn Huy*

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

*Tên đề tài : Xây dựng hệ thống phân tích comment để xác định sản phẩm được yêu thích.*

*Yêu cầu :*

*Đầu vào*

* *Dữ liệu từ* [*Facebook Comments Product Reviews*](https://www.kaggle.com/datasets/saurabhshahane/facebook-comment-volume-prediction)

*Đầu ra*

* *Biểu đồ sản phẩm theo số lượng comment tích cực/tiêu cực.*

*Các tính năng*

* *Text processing, sentiment analysis*
* *Visualization số liệu tương tác*

|  |
| --- |
| **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)*

# LỜI NÓI ĐẦU

Trong thời đại bùng nổ công nghệ thông tin và mạng xã hội, các nền tảng như Facebook trở thành nơi người dùng tự do chia sẻ cảm nhận, đánh giá về sản phẩm và dịch vụ. Những phản hồi này không chỉ phản ánh trải nghiệm của khách hàng mà còn là nguồn dữ liệu quý giá giúp doanh nghiệp hiểu rõ hơn về nhu cầu và xu hướng tiêu dùng.

Với mục tiêu vận dụng kiến thức đã học trong môn Khoa học dữ liệu, em lựa chọn đề tài “Phân tích comment Facebook về sản phẩm” để thực hiện bài tập lớn. Thông qua đề tài này, nhóm tiến hành xử lý và phân tích các bình luận về sản phẩm trên Facebook, từ đó xác định mức độ yêu thích của từng sản phẩm dựa trên cảm xúc tích cực hoặc tiêu cực được thể hiện qua nội dung comment.

Báo cáo này trình bày toàn bộ quá trình từ thu thập và xử lý dữ liệu văn bản, áp dụng kỹ thuật phân tích cảm xúc (sentiment analysis), cho đến trực quan hóa kết quả nhằm đánh giá mức độ tương tác và cảm nhận của người dùng đối với các sản phẩm.

Qua quá trình thực hiện, nhóm không chỉ củng cố kiến thức lý thuyết đã học mà còn rèn luyện kỹ năng làm việc với dữ liệu thực tế, góp phần nâng cao tư duy phân tích và khả năng áp dụng các công cụ khoa học dữ liệu trong thực tiễn.

**MỤC LỤC**

[LỜI NÓI ĐẦU 4](#_Toc199426587)

[CHƯƠNG 1 GIỚI THIỆU ĐẦU BÀI 7](#_Toc199426588)

[1.1. Bối cảnh và động lực 7](#_Toc199426589)

[1.2. Mục tiêu của đề tài 7](#_Toc199426590)

[1.3. Tính năng chính của hệ thống 7](#_Toc199426591)

[1.4. Thách thức của bài toán 8](#_Toc199426592)

[1.5. Kiến thức và công nghệ sử dụng 8](#_Toc199426593)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 9](#_Toc199426594)

[2.1. Khoa học dữ liệu (Data Science) 9](#_Toc199426595)

[2.2. Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP) 9](#_Toc199426596)

[2.3. Phân tích cảm xúc (Sentiment Analysis) 9](#_Toc199426597)

[2.4. Trực quan hóa dữ liệu (Data Visualization) 10](#_Toc199426598)

[2.5. Pandas và xử lý dữ liệu CSV 11](#_Toc199426599)

[2.6. Các thư viện bổ trợ khác 11](#_Toc199426600)

[CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH 13](#_Toc199426601)

[3.1. Sơ đồ khối hệ thống 13](#_Toc199426602)

[3.1.1. Mô tả các module chính 13](#_Toc199426603)

[3.1.2. Biểu đồ phân cấp chức năng 13](#_Toc199426604)

[3.2. Sơ đồ khối các thuật toán chính 14](#_Toc199426605)

[3.3. Cấu trúc dữ liệu 14](#_Toc199426606)

[3.4. Chương trình 15](#_Toc199426607)

[Chương 4: Thực nghiệm và Kết luận 18](#_Toc199426608)

[4.1. Thực nghiệm 18](#_Toc199426609)

[4.2. Kết luận 20](#_Toc199426610)

[KẾT LUẬN 22](#_Toc199426611)

**DANH SÁCH HÌNH ẢNH VÀ BIỂU ĐỒ**

Hình 2.1 Bảng phân loại cảm xúc theo compound

Hình 3.1 Biểu đồ phân cấp chức nắng

Hình 4.1 Đọc dữ liệu thành công

Hình 4.2 Cột sentiment được thêm vào DataFrame

Hình 4.3 Bảng thống kê sản phẩm và cảm xúc

Hình 4.4 Biểu đồ cảm xúc các sản phẩm

# CHƯƠNG 1 GIỚI THIỆU ĐẦU BÀI

## 1.1. Bối cảnh và động lực

Trong thời đại bùng nổ thông tin và thương mại điện tử hiện nay, người tiêu dùng ngày càng có xu hướng chia sẻ ý kiến, cảm nhận cá nhân về sản phẩm thông qua các nền tảng trực tuyến như Facebook, Amazon, Tiki, Shopee,... Những đánh giá và bình luận này không chỉ đóng vai trò như phản hồi dành cho nhà cung cấp mà còn là nguồn thông tin quý giá giúp những khách hàng tiềm năng đưa ra quyết định mua hàng.

Tuy nhiên, lượng bình luận quá lớn khiến cho việc đọc, lọc và đánh giá thủ công là không khả thi. Do đó, các công cụ và hệ thống phân tích cảm xúc từ văn bản đang trở nên ngày càng cần thiết và phổ biến. Việc tự động phân tích cảm xúc từ các comment không chỉ giúp xác định mức độ yêu thích của khách hàng đối với từng sản phẩm, mà còn cung cấp dữ liệu đầu vào cho các hệ thống hỗ trợ ra quyết định trong kinh doanh.

## 1.2. Mục tiêu của đề tài

Đề tài hướng đến việc xây dựng một hệ thống phân tích comment người dùng để xác định mức độ yêu thích của các sản phẩm. Cụ thể:

* Thực hiện tiền xử lý văn bản từ các bình luận sản phẩm trên mạng xã hội hoặc các nền tảng thương mại điện tử.
* Phân tích cảm xúc của từng bình luận để phân loại thành bình luận tích cực hoặc tiêu cực.
* Tổng hợp và thống kê số lượng bình luận tích cực, tiêu cực tương ứng với từng sản phẩm.
* Trực quan hóa kết quả bằng các biểu đồ, giúp dễ dàng nhận biết sản phẩm nào đang được người dùng yêu thích nhiều nhất và sản phẩm nào đang bị đánh giá tiêu cực.

## 1.3. Tính năng chính của hệ thống

Chương trình được xây dựng bao gồm các tính năng sau:

* Tiền xử lý văn bản: loại bỏ ký tự đặc biệt, chuyển về chữ thường, chuẩn hóa dữ liệu text để dễ phân tích.
* Phân tích cảm xúc (Sentiment Analysis): sử dụng thư viện ngôn ngữ tự nhiên (NLTK, TextBlob, hoặc VADER) để xác định cảm xúc của từng comment.
* Thống kê số liệu tương tác: tính toán số lượng comment tích cực và tiêu cực cho từng sản phẩm.
* Trực quan hóa dữ liệu: vẽ biểu đồ cột hoặc biểu đồ tròn thể hiện tỉ lệ cảm xúc tích cực / tiêu cực của từng sản phẩm.

## 1.4. Thách thức của bài toán

Việc xây dựng hệ thống phân tích cảm xúc từ dữ liệu bình luận gặp phải một số thách thức như sau:

* Dữ liệu bình luận không đồng nhất: Các comment có thể dùng tiếng lóng, viết tắt, sai chính tả, biểu tượng cảm xúc hoặc ngôn ngữ đa dạng, gây khó khăn cho việc phân tích tự động.
* Phân loại cảm xúc mơ hồ: Một số bình luận mang tính trung tính hoặc nhiều sắc thái pha trộn, khiến việc phân loại là tích cực hay tiêu cực trở nên không rõ ràng.
* Xử lý dữ liệu lớn: Khi số lượng comment tăng lên hàng chục nghìn dòng, hệ thống cần đảm bảo tốc độ xử lý nhanh và tiêu tốn ít tài nguyên.
* Tối ưu mô hình phân tích cảm xúc: Cần lựa chọn mô hình và kỹ thuật phù hợp để đảm bảo độ chính xác của việc phân loại cảm xúc.

## 1.5. Kiến thức và công nghệ sử dụng

Để giải quyết bài toán, nhóm đã vận dụng các kiến thức và công nghệ sau:

* Python: ngôn ngữ chính để xây dựng hệ thống, nổi bật với các thư viện hỗ trợ xử lý dữ liệu và học máy.
* Pandas: dùng để đọc, lọc, phân tích dữ liệu từ file CSV.
* Xử lý ngôn ngữ tự nhiên (NLP): sử dụng các kỹ thuật tiền xử lý văn bản (xóa stopwords, stemming, tokenization...).
* Thư viện phân tích cảm xúc: như TextBlob, VADER SentimentIntensityAnalyzer, giúp đánh giá nhanh cảm xúc của câu.
* Matplotlib / Seaborn: trực quan hóa kết quả phân tích thông qua biểu đồ, giúp người đọc dễ hiểu và đánh giá nhanh tình hình.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## 2.1. Khoa học dữ liệu (Data Science)

Khoa học dữ liệu là một lĩnh vực liên ngành kết hợp giữa lập trình, thống kê, trí tuệ nhân tạo (AI) và phân tích dữ liệu nhằm trích xuất thông tin có giá trị từ dữ liệu thô. Trong bối cảnh bài tập này, khoa học dữ liệu được ứng dụng để **phân tích** bình luận (comment) của người dùng về các sản phẩm nhằm rút ra cảm nhận của họ: yêu thích, không hài lòng hay trung lập.

Bài toán gồm các bước điển hình:

* Thu thập dữ liệu: Từ file CSV chứa các bình luận người dùng về sản phẩm.
* Tiền xử lý: Làm sạch và chuẩn hóa dữ liệu văn bản.
* Phân tích cảm xúc: Xác định cảm xúc trong từng bình luận.
* Tổng hợp và trực quan hóa: Đếm số lượng bình luận tích cực/tiêu cực và biểu diễn bằng biểu đồ.

## 2.2. Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)

Xử lý ngôn ngữ tự nhiên (NLP) là lĩnh vực của trí tuệ nhân tạo cho phép máy tính hiểu và phân tích văn bản giống như con người. Trong bài này, NLP là cốt lõi cho việc phân tích cảm xúc người dùng.

Các bước xử lý văn bản trong NLP:

* Bỏ dấu câu và ký tự đặc biệt: Nhằm đơn giản hóa văn bản.
* Chuyển về chữ thường (lowercasing): Tránh phân biệt "Tốt" và "tốt".
* Loại bỏ từ dừng (stopwords): Những từ phổ biến nhưng không mang nhiều thông tin như “là”, “và”, “của”, v.v.
* Tách từ (tokenization): Phân tách văn bản thành các đơn vị từ riêng lẻ.
* Lemmatization hoặc stemming: Đưa từ về gốc. Ví dụ: “mua”, “mua hàng”, “mua sắm” → “mua”.

## 2.3. Phân tích cảm xúc (Sentiment Analysis)

Phân tích cảm xúc là kỹ thuật xác định cảm nhận hoặc thái độ của người viết văn bản. Mỗi bình luận sẽ được gán nhãn là tích cực, tiêu cực, hoặc trung lập.

Phương pháp sử dụng: VADER Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) là một mô hình phân tích cảm xúc dựa trên từ điển. VADER rất phù hợp cho:

* Văn bản ngắn như bình luận, tweet, review sản phẩm
* Có khả năng hiểu các từ cảm xúc mạnh như “rất tốt”, “kinh khủng”, “siêu thích”

VADER trả về các giá trị sau:

* pos: Mức độ tích cực (positive)
* neu: Mức độ trung lập (neutral)
* neg: Mức độ tiêu cực (negative)
* compound: Tổng hợp toàn bộ điểm trên, từ -1 đến 1

Quy tắc phân loại cảm xúc theo compound:

|  |  |
| --- | --- |
| Giá trị compound | Nhãn cảm xúc |
| ≥ 0.05 | Tích cực |
| ≤ -0.05 | Tiêu cực |
| (-0.05, 0.05) | Trung lập |

Hình 2.1 Bảng phân loại cảm xúc theo compound

Ví dụ:

* "Sản phẩm quá tệ, không đáng tiền" → tiêu cực
* "Tôi rất hài lòng với chất lượng" → tích cực

## 2.4. Trực quan hóa dữ liệu (Data Visualization)

Trực quan hóa dữ liệu giúp người đọc dễ dàng hiểu và phân tích kết quả. Trong bài này, biểu đồ được sử dụng để:

* So sánh số lượng comment tích cực/tiêu cực giữa các sản phẩm
* Minh họa tỷ lệ cảm xúc của từng sản phẩm
* Phân tích xu hướng cảm xúc của người dùng

Các biểu đồ sẽ sử dụng:

* Bar chart (biểu đồ cột): Hiển thị tổng số comment tích cực/tiêu cực cho từng sản phẩm.
* Pie chart (biểu đồ tròn): Thể hiện tỷ lệ phần trăm các loại cảm xúc.

Các thư viện hỗ trợ:

* matplotlib: Thư viện trực quan hóa nền tảng trong Python.
* seaborn: Dễ dùng hơn matplotlib, dùng để tạo biểu đồ thống kê đẹp mắt.

## 2.5. Pandas và xử lý dữ liệu CSV

Pandas là thư viện Python phổ biến để xử lý dữ liệu bảng (DataFrame). Trong bài, pandas được dùng để:

* Đọc dữ liệu từ file .csv
* Lọc các cột cần thiết: Tên sản phẩm, nội dung comment
* Xử lý dữ liệu bị thiếu: Loại bỏ các dòng rỗng hoặc không hợp lệ
* Gán nhãn cảm xúc cho từng bình luận
* Thống kê số lượng cảm xúc theo từng sản phẩm

Một số hàm thường dùng:

* pd.read\_csv(): Đọc file CSV
* df.dropna(): Loại bỏ các dòng bị thiếu dữ liệu
* df.groupby(): Nhóm dữ liệu theo sản phẩm
* df.apply(): Áp dụng hàm cho từng dòng/cột

## 2.6. Các thư viện bổ trợ khác

Để xây dựng hệ thống phân tích comment và trực quan hóa dữ liệu một cách hiệu quả, chương trình sử dụng thêm một số thư viện Python hỗ trợ như sau:

* NLTK (Natural Language Toolkit): Là một thư viện nổi tiếng trong xử lý ngôn ngữ tự nhiên. Trong bài toán này, NLTK được dùng để tiền xử lý văn bản, bao gồm việc loại bỏ từ dừng (stopwords), tách từ (tokenization), và đặc biệt là hỗ trợ tích hợp công cụ phân tích cảm xúc VADER.
* VADER Sentiment (Valence Aware Dictionary and sEntiment Reasoner): Là một công cụ phân tích cảm xúc mạnh mẽ, phù hợp với các văn bản ngắn như comment, đánh giá sản phẩm. VADER sử dụng từ điển cảm xúc có sẵn để chấm điểm mức độ tích cực hoặc tiêu cực của câu. Thư viện này thường được dùng chung với NLTK.
* Matplotlib: Thư viện căn bản nhất để trực quan hóa dữ liệu trong Python. Nó được dùng để vẽ các biểu đồ như biểu đồ cột (bar chart), biểu đồ tròn (pie chart),... giúp trình bày kết quả phân tích một cách trực quan.
* Seaborn: Là một thư viện xây dựng dựa trên matplotlib, giúp vẽ các biểu đồ thống kê với phong cách trực quan và thẩm mỹ hơn. Trong bài này, Seaborn hỗ trợ vẽ các biểu đồ phân tích cảm xúc theo sản phẩm một cách dễ nhìn.
* re (Regular Expressions): Là thư viện xử lý biểu thức chính quy trong Python, dùng để lọc bỏ các ký tự đặc biệt, dấu câu và các thành phần không cần thiết trong văn bản trước khi phân tích.
* WordCloud *(tùy chọn)*: Thư viện này có thể được sử dụng để tạo các “đám mây từ”, trong đó những từ xuất hiện nhiều nhất trong comment sẽ có kích thước lớn hơn. Đây là một cách trực quan để thể hiện từ khóa nổi bật trong tập bình luận.

# CHƯƠNG 3: THIẾT KẾ VÀ XÂY DỰNG CHƯƠNG TRÌNH

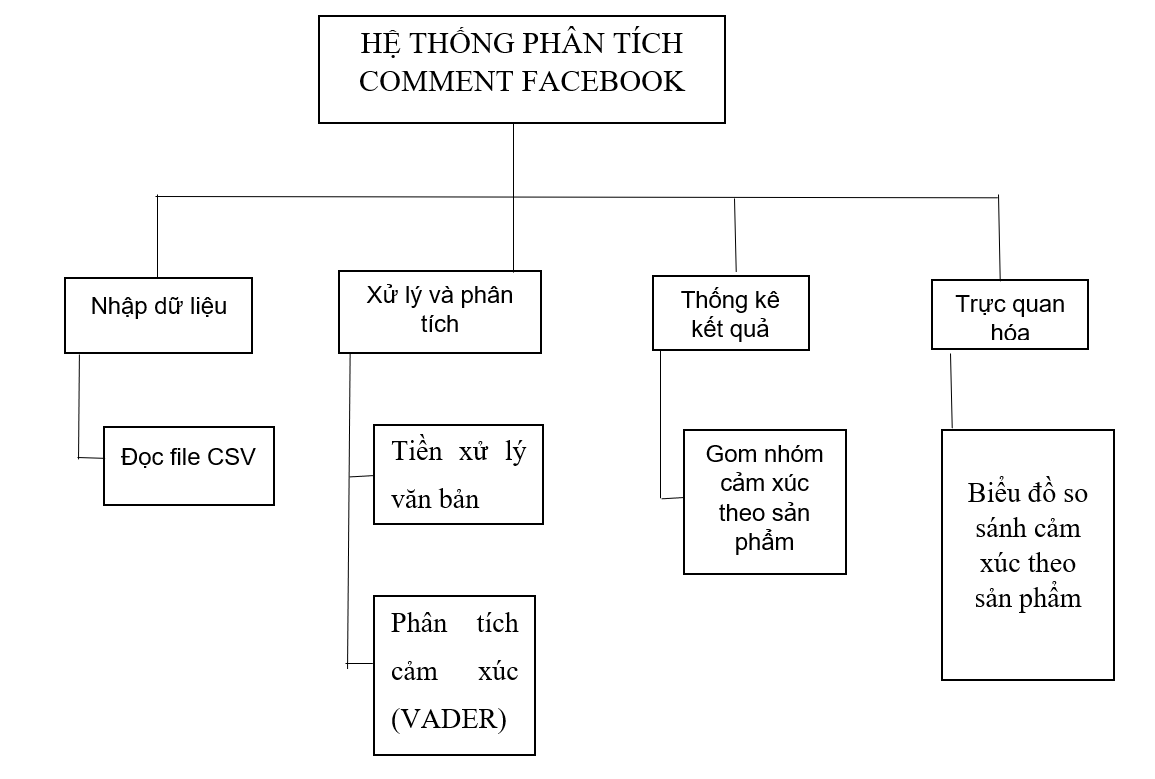
## 3.1. Sơ đồ khối hệ thống

### 3.1.1. Mô tả các module chính

Hệ thống phân tích bình luận Facebook về sản phẩm bao gồm 4 module chính:

* Module nhập dữ liệu, chức năng của module này là đọc dữ liệu từ file CSV chứa các bình luận Facebook của người dùng về các sản phẩm. Dữ liệu bao gồm tên sản phẩm và nội dung bình luận.
* Module xử lý văn bản và phân tích cảm xúc, các bình luận sẽ được xử lý tiền xử lý văn bản như kiểm tra định dạng, loại bỏ giá trị thiếu. Sau đó, mỗi bình luận sẽ được phân tích cảm xúc bằng thư viện VADER. Kết quả là mỗi bình luận sẽ được gán nhãn “Tích cực”, “Tiêu cực” hoặc “Trung tính”.
* Module tổng hợp thống kê, kết quả phân tích cảm xúc được gom nhóm theo từng sản phẩm. Hệ thống thống kê số lượng bình luận tích cực và tiêu cực cho từng sản phẩm, từ đó xác định sản phẩm được người dùng yêu thích hay không.
* Module trực quan hóa dữ liệu, kết quả phân tích được thể hiện dưới dạng biểu đồ cột giúp người dùng dễ dàng quan sát mức độ yêu thích của từng sản phẩm dựa trên cảm xúc bình luận.

### 3.1.2. Biểu đồ phân cấp chức năng



Hình 3.1 Biểu đồ phân cấp chức nắng

## 3.2. Sơ đồ khối các thuật toán chính

Hệ thống sử dụng các thuật toán đơn giản nhưng hiệu quả, sắp xếp theo dạng khối như sau:

* Khối 1: Đọc dữ liệu
  + Đầu vào: File CSV (tên sản phẩm, bình luận).
  + Xử lý: Sử dụng thư viện pandas để đọc và kiểm tra dữ liệu.
  + Đầu ra: DataFrame chứa dữ liệu ban đầu.
* Khối 2: Phân tích cảm xúc
  + Đầu vào: Cột bình luận (reviews.text).
  + Xử lý: Phân tích bằng thư viện VADER SentimentIntensityAnalyzer.
  + Kết quả: Thêm cột “sentiment” với 3 giá trị: Tích cực, Trung tính, Tiêu cực.
* Khối 3: Gom nhóm kết quả
  + Đầu vào: Dữ liệu đã gán nhãn cảm xúc.
  + Xử lý: Nhóm theo tên sản phẩm và loại cảm xúc.
  + Đầu ra: Bảng tổng hợp số lượng bình luận theo từng loại cảm xúc.
* Khối 4: Vẽ biểu đồ
  + Đầu vào: Bảng tổng hợp kết quả.
  + Xử lý: Lọc top 10 sản phẩm được bình luận nhiều nhất.
  + Kết quả: Biểu đồ cột thể hiện số lượng bình luận tích cực và tiêu cực theo từng sản phẩm.

## 3.3. Cấu trúc dữ liệu

a. Dữ liệu đầu vào

Dữ liệu được lấy từ tập tin CSV tên 1429\_1.csv. Một số trường chính:

* name: Tên sản phẩm được đánh giá.
* reviews.text: Nội dung bình luận người dùng.

b. Dữ liệu sau xử lý

Sau khi xử lý, dữ liệu sẽ có thêm một số trường như:

* sentiment: Nhãn cảm xúc sau phân tích (Tích cực, Tiêu cực, Trung tính).

## 3.4. Chương trình

Chương trình được xây dựng bằng ngôn ngữ Python với các thư viện phổ biến phục vụ xử lý dữ liệu, phân tích cảm xúc và trực quan hóa.

* pandas: Đọc và xử lý dữ liệu dạng bảng.
* matplotlib, seaborn: Vẽ biểu đồ.
* nltk.sentiment.vader: Phân tích cảm xúc văn bản tiếng Anh bằng mô hình VADER.

Chương trình gồm các bước và các hàm chính như sau:

1. Khởi tạo thư viện và đọc dữ liệu

Trước tiên, chương trình sử dụng các thư viện sau:

* pandas: để đọc và xử lý dữ liệu dạng bảng.
* matplotlib.pyplot và seaborn: để vẽ và trực quan hóa dữ liệu.
* nltk.sentiment.vader: để phân tích cảm xúc bình luận bằng mô hình VADER.

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from nltk.sentiment.vader import SentimentIntensityAnalyzer

import nltk

nltk.download("vader\_lexicon") # Tải bộ từ điển cảm xúc

sia = SentimentIntensityAnalyzer() # Khởi tạo bộ phân tích cảm xúc

Sau đó, dữ liệu bình luận được đọc từ tập tin 1429\_1.csv, và chỉ giữ lại hai cột quan trọng:

* name: tên sản phẩm.
* reviews.text: nội dung bình luận.

df = pd.read\_csv("1429\_1.csv")

df = df[["name", "reviews.text"]] # Giữ lại cột tên sản phẩm và nội dung bình luận

df = df.dropna() # Loại bỏ các dòng có giá trị bị thiếu

1. Hàm phân tích cảm xúc cho mỗi bình luận

Mỗi bình luận sẽ được gán một nhãn cảm xúc dựa trên mô hình VADER. Mô hình này sẽ tính điểm compound – là điểm tổng hợp cho biết mức độ tích cực hoặc tiêu cực của một câu:

* Nếu compound ≥ 0.05: bình luận là Tích cực.
* Nếu compound ≤ -0.05: bình luận là Tiêu cực.
* Nếu nằm giữa: bình luận là Trung tính.

Hàm xử lý như sau:

def analyze\_sentiment(text):

if isinstance(text, str):

score = sia.polarity\_scores(text)['compound']

if score >= 0.05:

return "Tích cực"

elif score <= -0.05:

return "Tiêu cực"

else:

return "Trung tính"

return "Trung tính"

Sau đó, chương trình áp dụng hàm này vào toàn bộ cột reviews.text để tạo thêm một cột mới sentiment.

df["sentiment"] = df["reviews.text"].apply(analyze\_sentiment)

c. Thống kê cảm xúc theo từng sản phẩm

Tiếp theo, chương trình nhóm các bình luận theo tên sản phẩm và loại cảm xúc, sau đó đếm số lượng cho từng loại:

sentiment\_counts = df.groupby(["name", "sentiment"]).size().unstack().fillna(0)

* Mỗi dòng tương ứng với một sản phẩm.
* Mỗi cột tương ứng với loại cảm xúc: Tích cực, Tiêu cực, Trung tính.

Chương trình tính thêm tổng số bình luận để sắp xếp:

sentiment\_counts["Total"] = sentiment\_counts.sum(axis=1)

sentiment\_counts = sentiment\_counts.sort\_values("Total", ascending=False)

top10 = sentiment\_counts.head(10) # Lấy 10 sản phẩm có nhiều bình luận nhất

d. Trực quan hóa dữ liệu

Cuối cùng, chương trình trực quan hóa số lượng bình luận Tích cực và Tiêu cực của 10 sản phẩm hàng đầu bằng biểu đồ cột chồng (stacked bar chart):

plt.figure(figsize=(12, 6))

top10[["Tích cực", "Tiêu cực"]].plot(kind="bar", stacked=True, colormap="Set2")

plt.title("Top 10 sản phẩm theo cảm xúc bình luận")

plt.ylabel("Số lượng bình luận")

plt.xlabel("Tên sản phẩm")

plt.xticks(rotation=45, ha='right')

plt.tight\_layout()

plt.show()

* Trục X: tên sản phẩm.
* Trục Y: số lượng bình luận.
* Mỗi cột thể hiện tổng số bình luận, trong đó màu sắc khác nhau thể hiện phần bình luận tích cực và tiêu cực.

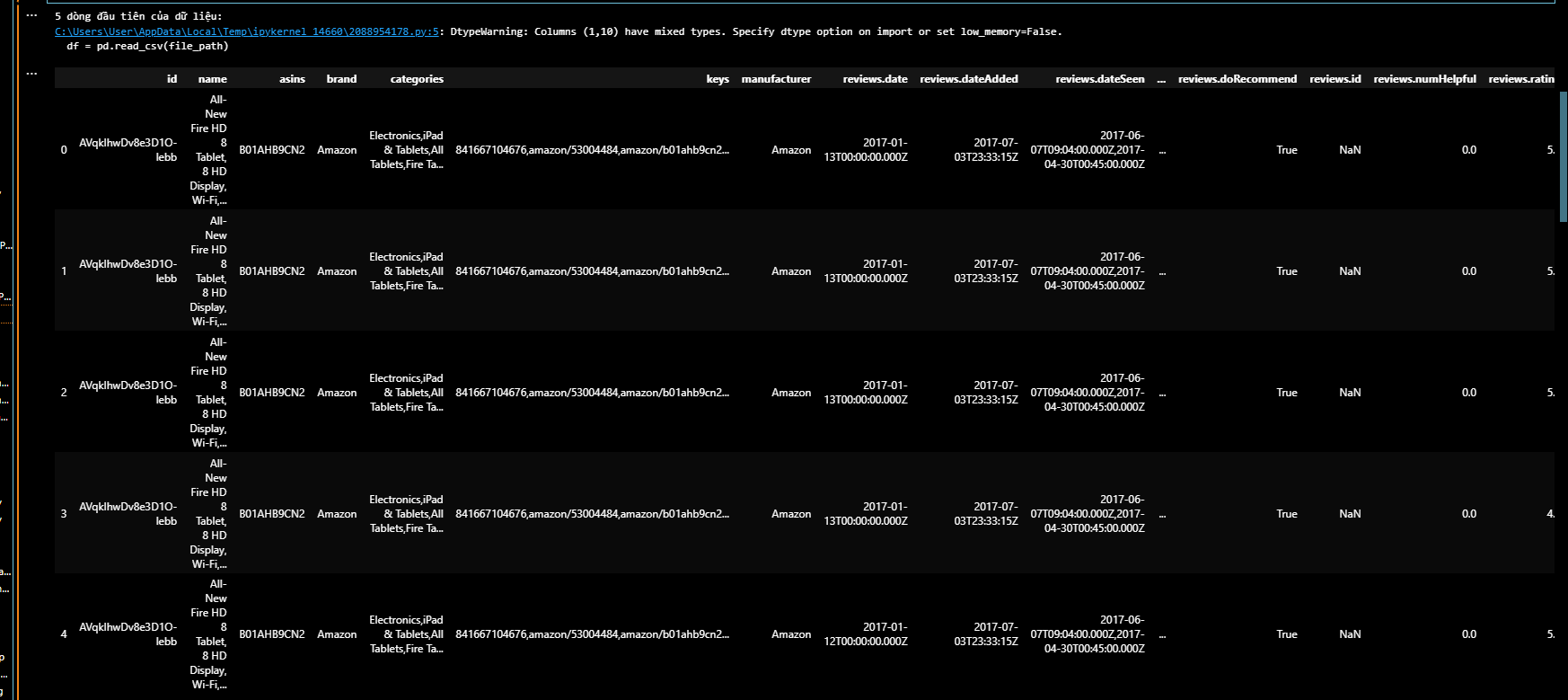
# Chương 4: Thực nghiệm và Kết luận

## 4.1. Thực nghiệm

Để đánh giá hiệu quả của chương trình phân tích cảm xúc bình luận sản phẩm từ Facebook, nhóm đã tiến hành các bước chạy thử nghiệm với tập dữ liệu thực tế (Facebook Comments Product Reviews) và ghi lại kết quả như sau:

a. Kiểm thử tính năng 1: Đọc và xử lý dữ liệu

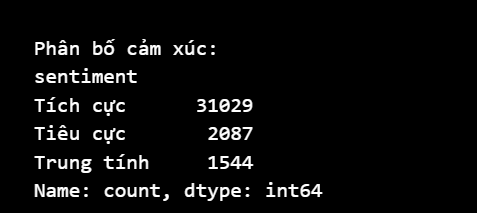
* Dữ liệu được tải thành công từ tệp CSV.
* Các dòng thiếu thông tin được xử lý hợp lý bằng phương pháp loại bỏ (dropna()).
* Đảm bảo giữ lại các cột quan trọng: tên sản phẩm và nội dung bình luận.



Hình 4.1 Đọc dữ liệu thành công

b. Kiểm thử tính năng 2: Phân tích cảm xúc bình luận

* Áp dụng mô hình VADER từ thư viện nltk.sentiment.vader cho từng dòng bình luận.
* Mỗi bình luận được phân loại thành: Tích cực, Tiêu cực, hoặc Trung tính.
* Đảm bảo không có lỗi khi xử lý các trường hợp văn bản trống, dữ liệu bị lỗi mã hóa.



Hình 4.2 *Cột sentiment được thêm vào DataFrame*

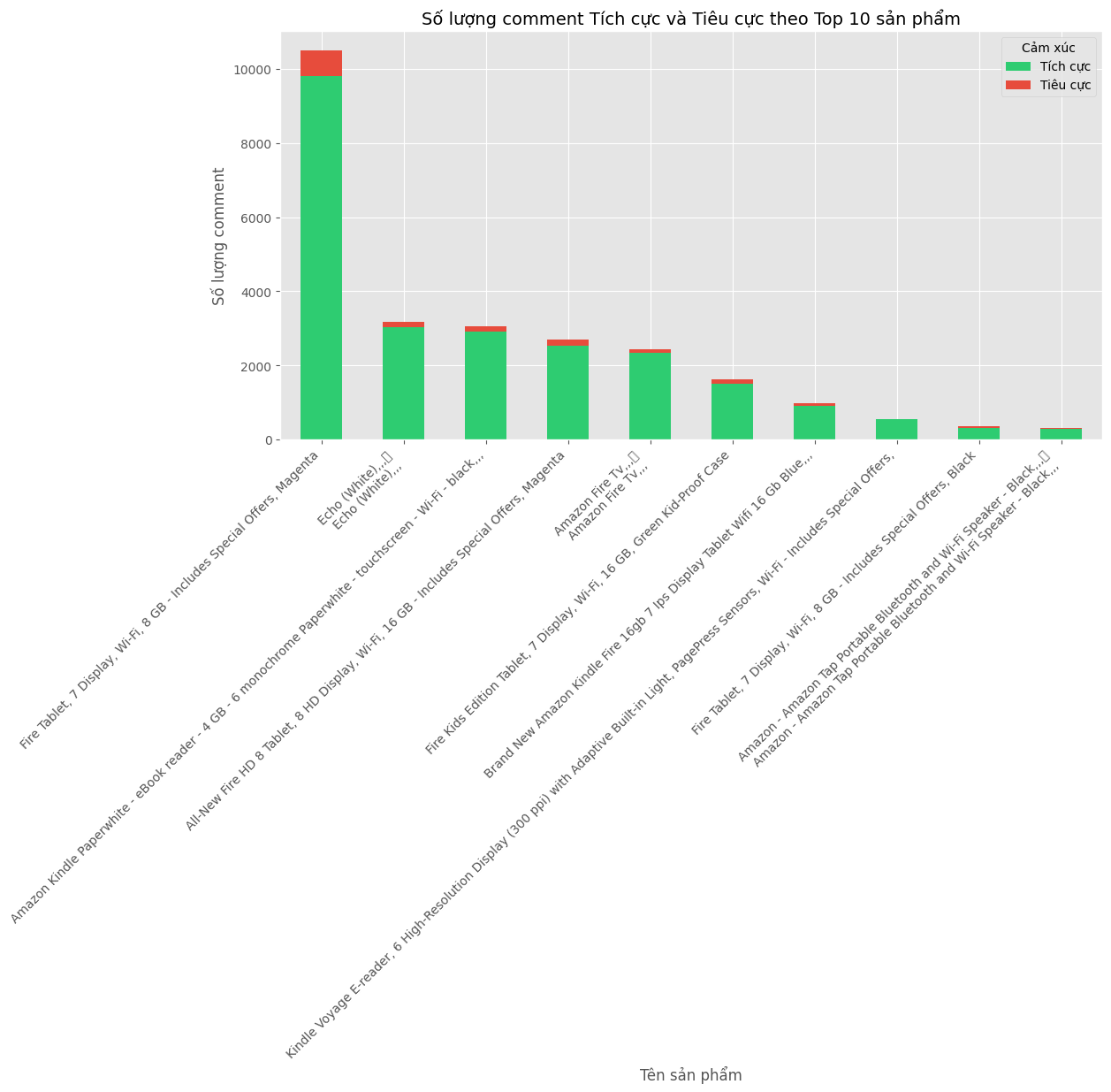
c. Kiểm thử tính năng 3: Thống kê và sắp xếp sản phẩm

* Dữ liệu được nhóm theo tên sản phẩm và loại cảm xúc.
* Tính tổng số bình luận và lọc ra 10 sản phẩm có số lượng bình luận nhiều nhất.
* Đảm bảo có thể sắp xếp đúng theo thứ tự giảm dần.

Hình 4.3 Bảng thống kê sản phẩm và cảm xúc

d. Kiểm thử tính năng 4: Trực quan hóa dữ liệu

* Biểu đồ cột chồng (stacked bar chart) thể hiện rõ ràng sự phân bố bình luận tích cực và tiêu cực của top 10 sản phẩm.
* Màu sắc khác nhau giúp dễ nhận diện mức độ yêu thích.
* Đảm bảo biểu đồ hiển thị rõ ràng, dễ đọc và không lỗi font hoặc kích thước.



Hình 4.4 *Biểu đồ cảm xúc các sản phẩm*

## 4.2. Kết luận

a. Sản phẩm đã đạt được

* Hoàn thiện hệ thống phân tích cảm xúc các bình luận về sản phẩm từ dữ liệu Facebook.
* Ứng dụng xử lý văn bản, phân tích cảm xúc bằng VADER, và trực quan hóa kết quả bằng biểu đồ.
* Có thể xác định sản phẩm được yêu thích thông qua tỷ lệ bình luận tích cực/tiêu cực.

b. Bài học kinh nghiệm

* Hiểu rõ hơn quy trình xử lý dữ liệu văn bản (text processing).
* Nắm được cách sử dụng mô hình cảm xúc VADER và các bước đánh giá một hệ thống phân tích dữ liệu.
* Cải thiện kỹ năng sử dụng thư viện Python như pandas, matplotlib, seaborn, nltk.

c. Hướng cải tiến trong tương lai

* Bổ sung giao diện người dùng (GUI hoặc web app) để người dùng không cần dùng dòng lệnh.
* Cho phép nhập dữ liệu mới hoặc live comment để phân tích real-time.
* Tích hợp mô hình học sâu (deep learning) để cải thiện độ chính xác cảm xúc.
* Cho phép phân tích cảm xúc theo thời gian hoặc xu hướng người dùng.

# KẾT LUẬN

Sau quá trình nghiên cứu và thực hiện đề tài "Xây dựng hệ thống phân tích comment để xác định sản phẩm được yêu thích", nhóm đã xây dựng thành công một chương trình có khả năng:

* Đọc và xử lý dữ liệu bình luận từ file CSV.
* Phân tích cảm xúc của từng bình luận thông qua mô hình VADER.
* Tổng hợp và thống kê cảm xúc theo từng sản phẩm.
* Trực quan hóa kết quả bằng biểu đồ để dễ dàng đánh giá sản phẩm nào được yêu thích nhiều nhất.

Hệ thống này chứng minh được tính khả thi trong việc ứng dụng các kỹ thuật khoa học dữ liệu, xử lý ngôn ngữ tự nhiên và trực quan hóa thông tin vào việc khai thác ý kiến người dùng trên mạng xã hội. Qua quá trình thực hiện, nhóm đã nâng cao kiến thức về Python, thư viện NLP như nltk, kỹ thuật phân tích dữ liệu với pandas, cũng như khả năng thiết kế một hệ thống đơn giản nhưng hiệu quả để giải quyết bài toán thực tế.

Tuy vậy, hệ thống vẫn còn một số hạn chế như chưa hỗ trợ tiếng Việt hoặc ngôn ngữ đa dạng, chưa có giao diện thân thiện cho người dùng cuối, và chưa xử lý được dữ liệu lớn. Trong tương lai, nhóm sẽ hướng đến việc:

* Xây dựng giao diện người dùng (desktop hoặc web).
* Hỗ trợ phân tích đa ngôn ngữ, đặc biệt là tiếng Việt.
* Kết hợp thêm các mô hình học sâu (deep learning) để tăng độ chính xác.
* Phân tích bình luận theo thời gian để đánh giá xu hướng cảm xúc người tiêu dùng.

Đề tài là một ví dụ tiêu biểu về việc áp dụng kiến thức Khoa học Dữ liệu vào thực tiễn, đặc biệt trong bối cảnh dữ liệu người dùng ngày càng phong phú trên các nền tảng mạng xã hội hiện nay.