



Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh  
**TRUNG TÂM TIN HỌC**

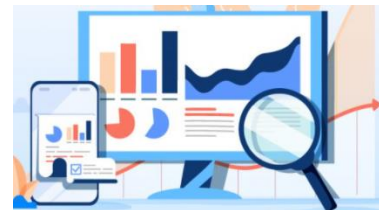
# Đồ án tốt nghiệp Data Science

## Topic: *Sentiment Analysis*

Phòng LT & Mạng

[https://csc.edu.vn/data-science-machine-learning/Do-An-Tot-Nghiep-Data-Science---Machine-Learning\\_229](https://csc.edu.vn/data-science-machine-learning/Do-An-Tot-Nghiep-Data-Science---Machine-Learning_229)

2024



# Nội dung

---



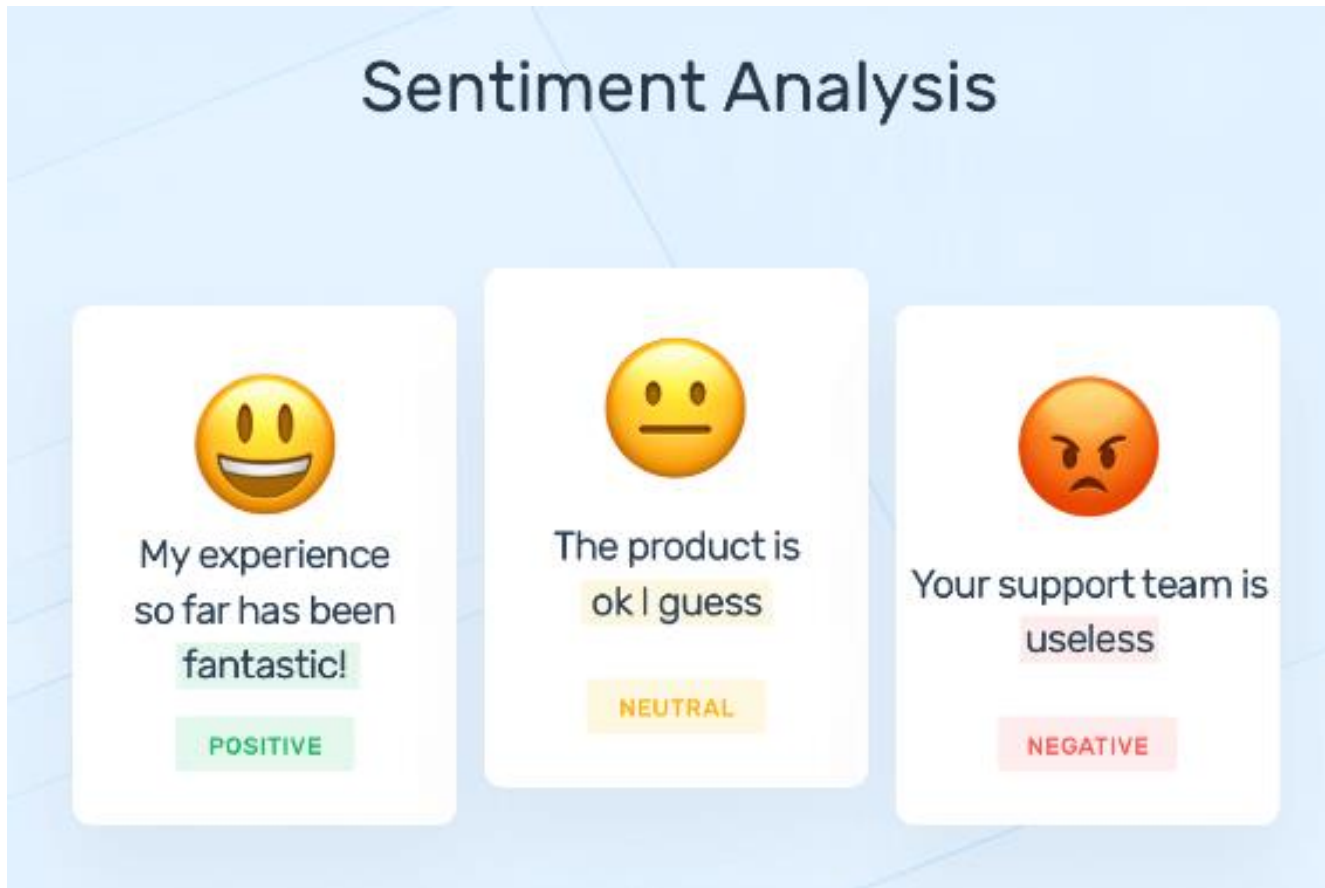
1. Giới thiệu project
2. Triển khai project theo Data Science Process

## □ Sentiment Analysis

- **Sentiment analysis - phân tích tình cảm** (hay còn gọi là **phân tích quan điểm, phân tích cảm xúc, phân tích cảm tính**, là cách sử dụng xử lý ngôn ngữ tự nhiên, phân tích văn bản, ngôn ngữ học tính toán, và sinh trắc học để nhận diện, trích xuất, định lượng và nghiên cứu các trạng thái tình cảm và thông tin chủ quan một cách có hệ thống. Sentiment analysis được áp dụng rộng rãi cho các tài liệu chẳng hạn như các đánh giá và các phản hồi khảo sát, phương tiện truyền thông xã hội, phương tiện truyền thông trực tuyến, và các tài liệu cho các ứng dụng từ marketing đến quản lý quan hệ khách hàng và y học lâm sàng.

[https://vi.wikipedia.org/wiki/Ph%C3%A2n\\_t%C3%ADch\\_t%C3%ACnh\\_c%E1%BA%A3m](https://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_t%C3%ACnh_c%E1%BA%A3m)

# Giới thiệu project



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

# Giới thiệu project



- Sentiment Analysis là quá trình phân tích, đánh giá quan điểm của một người về một đối tượng nào đó (quan điểm mang tính tích cực, tiêu cực, hay trung tính,...). Quá trình này có thể thực hiện bằng việc sử dụng các tập luật (rule-based), sử dụng Machine Learning hoặc phương pháp Hybrid (kết hợp hai phương pháp trên).
- Sentiment Analysis được ứng dụng nhiều trong thực tế, đặc biệt là trong hoạt động quảng bá kinh doanh. Việc phân tích đánh giá của người dùng về một sản phẩm xem họ đánh giá tiêu cực, tích cực hoặc đánh giá các hạn chế của sản phẩm sẽ giúp công ty nâng cao chất lượng sản phẩm và tăng cường hình ảnh của công ty, củng cố sự hài lòng của khách hàng.

## ❑ Sentiment Analysis trong ẩm thực

- Chúng ta luôn thích thú khi được khám phá những điều mới mỗi ngày và ẩm thực là một lĩnh vực thu hút sự quan tâm rất lớn.
- Để lựa chọn một nhà hàng/quán ăn mới chúng ta có xu hướng xem xét những bình luận từ những người đã thưởng thức để đưa ra quyết định có nên thử hay không?

- Điều này trở nên quan trọng hơn trong ngành dịch vụ ẩm thực. Các nhà hàng/quán ăn cần nỗ lực để cải thiện chất lượng của món ăn cũng như thái độ phục vụ nhằm duy trì uy tín của nhà hàng cũng như tìm kiếm thêm khách hàng mới.

=> Xây dựng hệ thống hỗ trợ nhà hàng/quán ăn phân loại các phản hồi của khách hàng thành các nhóm: tích cực, tiêu cực, trung tính dựa trên dữ liệu dạng văn bản.

## ❑ Business Objective/Problem ShopeeFood

- ShopeeFood là một kênh phối hợp với các nhà hàng/quán ăn bán thực phẩm online.
- Chúng ta có thể lên đây để xem các đánh giá, nhận xét cũng như đặt mua thực phẩm.
- Từ những đánh giá của khách hàng, vấn đề được đưa ra là làm sao để các nhà hàng/ quán ăn hiểu được khách hàng rõ hơn, biết họ đánh giá về mình như thế nào để cải thiện hơn trong dịch vụ/ sản phẩm.



## ❑ Các kiến thức/ kỹ năng cần để giải quyết vấn đề này:

- Hiểu vấn đề
- Import các thư viện cần thiết và hiểu cách sử dụng
- Đọc dữ liệu (dữ liệu project này được cung cấp)
- Thực hiện EDA cơ bản
- Tiền xử lý dữ liệu: làm sạch, tạo tính năng mới, lựa chọn tính năng cần thiết...

# Giới thiệu project

---



- Trực quan hóa dữ liệu
- Lựa chọn thuật toán cho bài toán classification
- Xây dựng model
- Đánh giá model
- Báo cáo kết quả

# Nội dung

---



1. Giới thiệu project
2. Triển khai project theo Data Science Process

# Triển khai project theo Data Science Process

---




- Thư viện sử dụng

- numpy, pandas, matplotlib, seaborn
- underthesea
- glob
- wordcloud
- scikit-learn (sklearn)
- Pyspark
- ...

## □ Triển khai dự án

### ● Bước 1: Business Understanding

#### ■ Dựa vào mô tả nói trên => xác định vấn đề:

- Xây dựng hệ thống dựa trên lịch sử những đánh giá của khách hàng đã có trước đó. Dữ liệu được thu thập từ phần bình luận và đánh giá của khách hàng ở  **ShopeeFood**

=> Mục tiêu/ vấn đề: Xây dựng mô hình dự đoán giúp nhà hàng/ quán ăn có thể biết được những phản hồi nhanh chóng của khách hàng về sản phẩm hay dịch vụ của họ (tích cực, tiêu cực hay trung tính), điều này giúp cho nhà hàng hiểu được tình hình kinh doanh, hiểu được ý kiến của khách hàng từ đó giúp nhà hàng cải thiện hơn trong dịch vụ, sản phẩm.

# Triển khai project theo Data Science Process



## ● Bước 2: Data Understanding/ Acquire

- Từ mục tiêu/ vấn đề đã xác định: xem xét các dữ liệu cần thiết:
  - Dữ liệu được cung cấp sẵn trong tập tin 2\_Reviews.csv với gần 30.000 mẫu gồm các thông tin: ID (mã), User (người dùng), Time (thời gian đánh giá), Rating (điểm đánh giá), Comment (nội dung đánh giá), và IDRestaurant (mã nhà hàng)
  - Ngoài ra, còn có tập tin chứa thông tin về nhà hàng: 1\_Restaurants.csv với hơn 1.600 mẫu gồm các thông tin: ID (mã), Restaurant (tên nhà hàng), Address (địa chỉ), Time (giờ mở cửa), Price (khoảng giá), District(quận)

*HV tự thu thập thêm dữ liệu có các thông tin như mô tả trên và đưa vào tập dữ liệu sẽ được +0.25đ*



# Triển khai project theo Data Science Process

---



=> Có thể tập trung giải quyết bài toán

- Sentiment analysis với các thuật toán thuộc nhóm Supervised Learning – Classification như: Naïve Bayes, KNN, Logistic Regression, Tree Models...
- Triển khai với cả các thuật toán phù hợp trong môn Machine Learning with Python (MDS6) và Big Data in Machine Learning (LDS9)

# Triển khai project theo Data Science Process



## ● Bước 3: Data preparation/ Prepare

### ■ Thực hiện các công việc:

- Vì đây là dữ liệu text Việt nên cần thực hiện công việc chuẩn hóa text là chính.
- Cũng như những ngôn ngữ khác, khi làm việc với tiếng Việt, chúng ta phải tiền xử lý dữ liệu. Tuy nhiên, có rất ít thư viện NLP hỗ trợ công việc này ngay từ đầu nên chúng ta phải tự thực hiện các xử lý làm sạch dữ liệu thô ban đầu trước khi dùng thư viện cung cấp để chuẩn hóa sau giúp có dữ liệu chuẩn đưa vào huấn luyện mô hình.

**demo\_VN\_pre\_underthesea\_032024.ipynb**





# Triển khai project theo Data Science Process



- Bước 4&5: Modeling & Evaluation/ Analyze & Report
  - Xây dựng các Classification model dự đoán giá (trong MDS6 & LDS9)
    - Naïve Bayes, KNN, Logistic Regression, Tree Algorithms
    - ...
  - Thực hiện/ đánh giá kết quả các Classification model
    - R-squared, acc, precision, recall, f1, confusion matrix, ROC curve...
    - Kết luận

# Triển khai project theo Data Science Process

---



- Bước 6: Deployment & Feedback/ Act
  - Đưa ra những cải tiến phù hợp để nâng cao sự hài lòng của khách hàng, thu hút sự chú ý của khách hàng mới.

# Triển khai project theo Data Science Process



Các công việc cần thực hiện...

# Triển khai project theo Data Science Process



## □ Với project phía trên

- Yêu cầu 1: Thu thập dữ liệu review trên <https://shopeefood.vn/> (+0.25 đ).
  - Chọn địa điểm: có thể chọn thành phố Hồ Chí Minh..., chọn loại đồ ăn/ đồ uống/... ngoài dữ liệu được cung cấp
  - Kết quả: file dữ liệu thu thập thêm có các cột thông tin như file dữ liệu đã cung cấp ...

# Triển khai project theo Data Science Process



The screenshot displays the ShopeeFood website interface. At the top, the ShopeeFood logo is on the left, and navigation links for 'TP. HCM', 'Đồ ăn', 'Thực phẩm', 'Bia', 'Hoa', 'Siêu thị', 'Thuốc', and 'Thủ cưng' are in the center. A search bar and a 'Đăng nhập' button are on the right. The main content area features a large banner with the text 'Đặt Đồ ăn, giao hàng từ 20'...' and a search bar. Below the banner, there are several category buttons: 'All', 'Đồ ăn', 'Đồ uống', 'Đồ chay', 'Bánh kem', 'Tráng miệng', 'Homemade', 'Vía hè', 'Pizza/Burger', 'Món gà', 'Món lẩu', 'Sushi', 'Mì phở', and 'Cơm hộp'. A section titled 'Ưu đãi' (Offers) displays a grid of food items with their names, addresses, and discount percentages. Below this, there is a 'Bộ sưu tập' (Collection) section with three promotional banners for 'FREESHIP 0Đ', 'COMBO 33.000 ĐỒNG', and 'SIÊU DEAL 50%'.

**Đặt Đồ ăn, giao hàng từ 20'...**  
Có 75426 Địa Điểm Ở TP. HCM Từ 00:00 - 23:59

Tìm địa điểm, món ăn, địa chỉ...

All Đồ ăn Đồ uống Đồ chay  
Bánh kem Tráng miệng Homemade  
Vía hè Pizza/Burger Món gà  
Món lẩu Sushi Mì phở Cơm hộp

Sử dụng App ShopeeFood để có nhiều giảm giá và trải nghiệm tốt hơn

Available on the App Store  
Get it on Google play

**Ưu đãi** Xem tất cả

- Bún Bò Đắt Thánh - Sh...  
221/16 Đắt Thánh, P. 6, Tân...  
Giảm món
- Tiên Tiên - Bún Thái C...  
243/4 Chu Văn An, P. 12, Bìn...  
Giảm hết 10%
- Bún Thịt Nướng Di 7 ...  
1779/21/6 Khu Phố 2A, Quố...  
Giảm hết 10%
- Bánh Mì Bắp Hồ  
19 Huỳnh Khương Ninh, P. ...  
Giảm món
- Cô Liễu - Gỏi Cuốn  
11 Đường Số 8, P. Trương T...  
Giảm hết 10%
- TocoToco Bubble Tea - ...  
892C Tạ Quang Bửu, P. 5, Q...  
Giảm hết 30%
- Chè Ngon Cô Ba - Cô G...  
176 Cô Giang, P. Cô Giang, ...  
Đồng giá 19K - Deal 1K
- Texas Chicken - Quan...  
578 Quang Trung, P.11, Gò ...  
Giảm món
- Bánh Bò Thốt Nốt Co...  
18E Trần Quang Diệu, P. 13, ...  
Giảm món

Xem thêm

**Bộ sưu tập** Xem tất cả

- FREESHIP 0Đ
- COMBO 33.000 ĐỒNG
- SIÊU DEAL 50%

# Triển khai project theo Data Science Process

---



- Yêu cầu 2: Thực hiện việc tiền xử lý dữ liệu text để làm đầu vào cho việc xây dựng model (0.5 điểm).

# Triển khai project theo Data Science Process



- Yêu cầu 3: Đề xuất các thuật toán có thể, sau đó chọn thuật toán phù hợp để giải quyết vấn đề, report kết quả.
  - Lựa chọn thuật toán phù hợp trong Machine Learning with Python (MDS6) (0.5đ)
  - Lựa chọn thuật toán phù hợp trong môn Big Data Machine Learning (LDS9) (0.5đ)
  - Nếu dữ liệu mất cân bằng gây ảnh hưởng đến kết quả thì xem xét thêm việc xử lý mất cân bằng (0.25đ)
  - So sánh các kết quả. (0.25đ)
  - Có thể đề xuất thêm các thuật toán mới (+0.25đ)

