**BÁO CÁO DỰ ÁN: PHÂN TÍCH CẢM XÚC KHÁCH HÀNG AMAZON**

# **1. Giới thiệu đề tài**

Trong thời đại thương mại điện tử phát triển mạnh mẽ, hàng triệu người tiêu dùng chia sẻ cảm nhận của họ sau khi mua hàng. Các đánh giá (review) không chỉ thể hiện trải nghiệm cá nhân mà còn ảnh hưởng trực tiếp đến uy tín thương hiệu và doanh số bán hàng.  
Dự án **“Amazon Customer Sentiment Analysis”** được thực hiện nhằm **phân tích cảm xúc của người dùng** thông qua hàng trăm nghìn bài đánh giá thực tế trên Amazon, giúp doanh nghiệp hiểu sâu hơn về cảm nhận của người dùng và xu hướng thị trường.

# **2. Mục tiêu dự án**

* Xây dựng mô hình phân tích cảm xúc từ dữ liệu đánh giá sản phẩm Amazon.
* Xác định tỷ lệ phản hồi tích cực, tiêu cực và trung lập.
* Tìm ra **những sản phẩm được đánh giá tích cực nhất**.
* Tạo **dashboard Power BI** trực quan, thể hiện các chỉ số và xu hướng cảm xúc theo thời gian.

# **3. Quy trình thực hiện**

## ****Bước 1. Tiền xử lý dữ liệu (Data Cleaning)****

* Xóa các dòng dữ liệu trống hoặc trùng lặp.
* Chuẩn hóa văn bản: chuyển về chữ thường, loại bỏ ký tự đặc biệt và stopwords.
* Tách riêng các cột: Review Text, Score, ProductId, UserId, Time.
* Xuất file CSV sạch (**sentiment\_predictions\_sample.csv**) để làm dữ liệu đầu vào cho bước phân tích mô hình.

## ****Bước 2. Xây dựng mô hình phân tích cảm xúc (Sentiment Analysis)****

* Sử dụng thư viện **scikit-learn** với mô hình **TF-IDF + Logistic Regression**.
* Gán nhãn kết quả dự đoán vào cột **Sentiment\_predicted** gồm 3 giá trị:
  + **Positive**
  + **Negative**
  + **Neutral**
* Tính thêm các chỉ số thống kê theo sản phẩm:
  + Positive\_Rate, Negative\_Count, Total\_Reviews, Score trung bình
* Xuất dữ liệu sau mô hình sang file sentiment\_predictions\_sample.csv.

## ****Bước 3. Trực quan hóa dữ liệu trên Power BI****

File dữ liệu sau khi xử lý được nạp vào Power BI để xây dựng **dashboard tổng quan** gồm:

|  |  |  |
| --- | --- | --- |
| **Thành phần** | **Visual** | **Ý nghĩa** |
| **Net Sentiment Score of Positive** | Gauge chart | Tỷ lệ đánh giá tích cực (79.68%) – đạt gần mục tiêu 80% |
| **Overall Sentiment** | Donut chart | Phân bố tỉ lệ các loại cảm xúc |
| **Feedback Cloud** | Word Cloud | Từ khóa nổi bật trong các đánh giá |
| **Trend Over Time** | Line chart | Biến động tỷ lệ cảm xúc theo thời gian |
| **Top 10 Products** | Bar chart | Sản phẩm có số lượng đánh giá tích cực cao nhất |
| **Customer Feedback Highlights** | Card/Text box | Trích dẫn một số phản hồi thực tế từ khách hàng |
| **Summary Table** | Table | Thống kê tổng hợp theo mã sản phẩm |

# **4. Kết quả và Nhận xét**

## ****Kết quả tổng quan****

* **79.68%** các đánh giá mang cảm xúc **tích cực**, phản ánh mức độ hài lòng cao.
* **20.32%** còn lại thuộc nhóm **trung lập hoặc tiêu cực**, cho thấy vẫn tồn tại một nhóm nhỏ khách hàng chưa thực sự hài lòng.
* Sản phẩm mã **B007FMH8M8** có **153 đánh giá tích cực**, dẫn đầu danh sách, trong khi các sản phẩm khác dao động quanh 100–120 đánh giá tích cực.
* Các từ khóa nổi bật trong Word Cloud gồm: “good”, “love”, “taste”, “fast”, “price”, thể hiện **khách hàng đánh giá cao chất lượng, hương vị, và tốc độ giao hàng**.
* Ngược lại, các phản hồi tiêu cực (dù ít) thường nhắc đến **“late delivery”, “broken package”, “stale product”**.

### **Insight mở rộng**

* Trung bình **1 khách hàng chỉ để lại khoảng 1.2 review**, cho thấy tính đa dạng người dùng cao (ít trùng lặp).
* Nhóm sản phẩm “đồ ăn nhanh” hoặc “đồ uống” có **tỷ lệ review tích cực cao hơn nhóm hàng hóa gia dụng**.
* Có mối tương quan giữa điểm số Score và xác suất dự đoán Positive\_Probability (R² ≈ 0.82).  
  → Mô hình dự đoán hoạt động khá ổn định.

Giả sử bạn là **quản lý danh mục hàng tiêu dùng trên Amazon**, dashboard này có thể giúp bạn:

* Phát hiện rằng khách hàng **rất hài lòng với sản phẩm cà phê và sô-cô-la**, nhưng lại **phàn nàn nhiều về bao bì và thời gian giao hàng**.
* Dựa vào đó, bạn có thể **làm việc với bộ phận logistics** để cải thiện khâu giao hàng, hoặc **điều chỉnh chiến dịch quảng cáo** nhấn mạnh vào “fast & fresh delivery”.
* Với tỷ lệ sentiment tích cực gần **80%**, bạn có thể sử dụng con số này làm **chỉ số KPI về trải nghiệm khách hàng (Customer Satisfaction Index)** trong báo cáo tháng.

## ****6. Khó khăn và Hạn chế của Dataset****

|  |  |
| --- | --- |
| Khó khăn | Mô tả chi tiết |
| **Không đồng nhất ngôn ngữ & độ dài review** | Một số đánh giá quá ngắn (“good”) hoặc quá dài, gây nhiễu khi huấn luyện mô hình. |
| **Thiếu metadata bổ sung** | Dataset không chứa thông tin về giá, quốc gia, hoặc danh mục sản phẩm cụ thể, nên khó phân tích sâu theo nhóm hàng. |
| **Cảm xúc gián tiếp** | Một số người dùng viết theo phong cách mỉa mai (“I just love waiting 3 weeks for my order…”), mô hình dễ dự đoán sai. |
| **Phân bố dữ liệu lệch (imbalanced)** | Dữ liệu có tỷ lệ đánh giá tích cực chiếm đa số → mô hình dễ thiên lệch về “Positive”. |
| **Thiếu nhãn xác thực** | Các nhãn cảm xúc gốc được gán tự động theo điểm Score, không hoàn toàn phản ánh cảm xúc thật của người viết. |

## ****7. Kết luận****

Dự án đã hoàn thiện toàn bộ quy trình **từ tiền xử lý dữ liệu → huấn luyện mô hình → xây dựng báo cáo Power BI.**Kết quả cho thấy **Sentiment Analysis** có thể trở thành **công cụ hỗ trợ ra quyết định hữu hiệu** cho doanh nghiệp thương mại điện tử.

Tuy vẫn còn hạn chế về dữ liệu, nhưng project đã thể hiện rõ **quy trình xử lý dữ liệu thực tế**, khả năng **ứng dụng AI trong phân tích kinh doanh**, và **tư duy trực quan hóa dữ liệu chuyên nghiệp**.

## ****8. Tệp đính kèm****

🔗 **Power BI Dashboard:** <https://app.powerbi.com/links/uaOcFFzAbV?ctid=b1a9fdc0-1d56-4c3d-a481-809fff8a26db&pbi_source=linkShare>

📁 **Dữ liệu đã qua xử lý:** <https://drive.google.com/file/d/145jderUkfzKmo97v3EHydbzGxHH9gdX3/view?usp=sharing>

🐍 **Source code (Python):** <https://colab.research.google.com/drive/1dFzLdC8nZOnyIpNfaNoahgUU9STa9BT0?usp=sharing>