Exploratory Data Analysis (EDA)

Amazon Fine Food Reviews Dataset

Diệp Thụy An
Đỗ Mai Anh
Cao Tiến Cường
Võ Hoàng Phúc hy

1. Ageda (mục tiêu phân tích dữ liệu)
2. Data summary (tóm tắt dữ liệu)
3. Tiền xử lý dữ liệu
4. Univariate analysis (phân tích đơn biến)
5. Sentiment distribution analysis (phân tích phân phối cảm xúc)
6. Review length analysis (độ dài các đánh giá)
7. Correlation analysis (phân tích tương quan)
8. Conclusion (kết luận)

1. Ageda (mục tiêu phân tích dữ liệu)

Khảo sát dataset của Amazon Fine Food Reviews

- Hiểu rõ phân phối đánh giá sản phẩm trên Amazon.
- Kiểm tra phân phối điểm số đánh giá sản phẩm
- Kiểm tra mối quan hệ giữa độ dài đánh giá và điểm số cảm xúc.
- Kiểm tra dữ liệu có bị mất cân bằng không

2. Data summary (tóm tắt dữ liệu)

Tập dữ liệu AFFR (568 454 đánh giá sản phẩm thực phẩm) có 10 cột dữ liệu, nội dung các cột dữ liệu là:

Id: Mã Id định danh cho mỗi đánh giá

ProductId: Id sản phẩm cho mỗi đánh giá

UserId: Id người dùng đã đánh giá

ProfileName: Tên người đánh giá

HelpfulnessNumerator: Số lượng người thấy đánh giá hữu ích

HelpfulnessDenominator: Tổng số lượt người dùng đã bỏ phiếu mức độ hữu ích của đánh giá

Score: Điểm đánh giá từ 1 đến 5

Time: Mốc thời gian đánh giá theo dạng Unix time

Summary: Phần tóm tắt ngắn hoặc tiêu đề của đánh giá

Text: Nội dung đầy đủ của đánh giá

2. Data summary (tóm tắt dữ liệu)

568,454 dòng bình luận đánh giá sản phẩm

- 256.059 người dùng
- 74.258 sản phẩm

```
Số lượng sản phẩm được đánh giá: 74258
Số lượng người dùng: 256059
Kiểm tra tính toàn vẹn dữ liệu

Dữ liệu trống:
Id                          0
ProductId                   0
UserId                      0
ProfileName                26
HelpfulnessNumerator        0
HelpfulnessDenominator      0
Score                       0
Time                        0
Summary                    27
Text                        0
dtype: int64
```

Dữ liệu sau khi kiểm tra thì phát hiện

- Tồn tại 26 dòng có dữ liệu trống ở ProfileName
- Tồn tại 27 dòng có dữ liệu trống ở Summary

Có 174875 dòng dữ liệu trùng nhau ở Text

```
Số đánh giá trùng lặp:  174875
            Id   ProductId          UserId  ...        Time                                Summary                                               Text
29          30  B0001PB9FY  A3HDKO7OW0QNK4  ...  1107820800            The Best Hot Sauce in the World  I don't know if it's the cactus or the tequila...
574        575  B000G6RYNE  A3PJZ8TU8FDQ1K  ...  1231718400  One bite and you'll become a "chippoisseur"  I'm addicted to salty and tangy flavors, so wh...
603        604  B000G6RYNE  A3PJZ8TU8FDQ1K  ...  1229385600  One bite and you'll become a "chippoisseur"  I'm addicted to salty and tangy flavors, so wh...
1973      1974  B00171650G  A2EPNS38TTLZYN  ...  1312675200                            Pok Chops  The pork chops from Omaha Steaks were very tas...
2309      2310  B0001VWE0M   AQM74O8Z4FMS0  ...  1127606400                        Below standard  Too much of the white pith on this orange peel...
...        ...         ...             ...  ...         ...                                  ...                                                ...
568409  568410  B0018CLWM4  A2PE0AGWV6OPL7  ...  1309651200          Quality & affordable food  I was very pleased with the ingredient quality...
568410  568411  B0018CLWM4   A88HLWDCU57WG  ...  1332979200                           litter box  My main reason for the five star review has to...
568411  568412  B0018CLWM4   AUX1HSY8FX55S  ...  1319500800                        Happy Camper  I bought this to try on two registered Maine C...
568412  568413  B0018CLWM4   AVZ2OZ479Q9E8  ...  1336435200          Two Siberians like it!  When we brought home two 3-month-old purebred ...
568413  568414  B0018CLWM4   AI3Y26HLPYW4L  ...  1330041600          premium edge cat food  My cats don't like it. what else can I say to ...

[174875 rows x 10 columns]
```

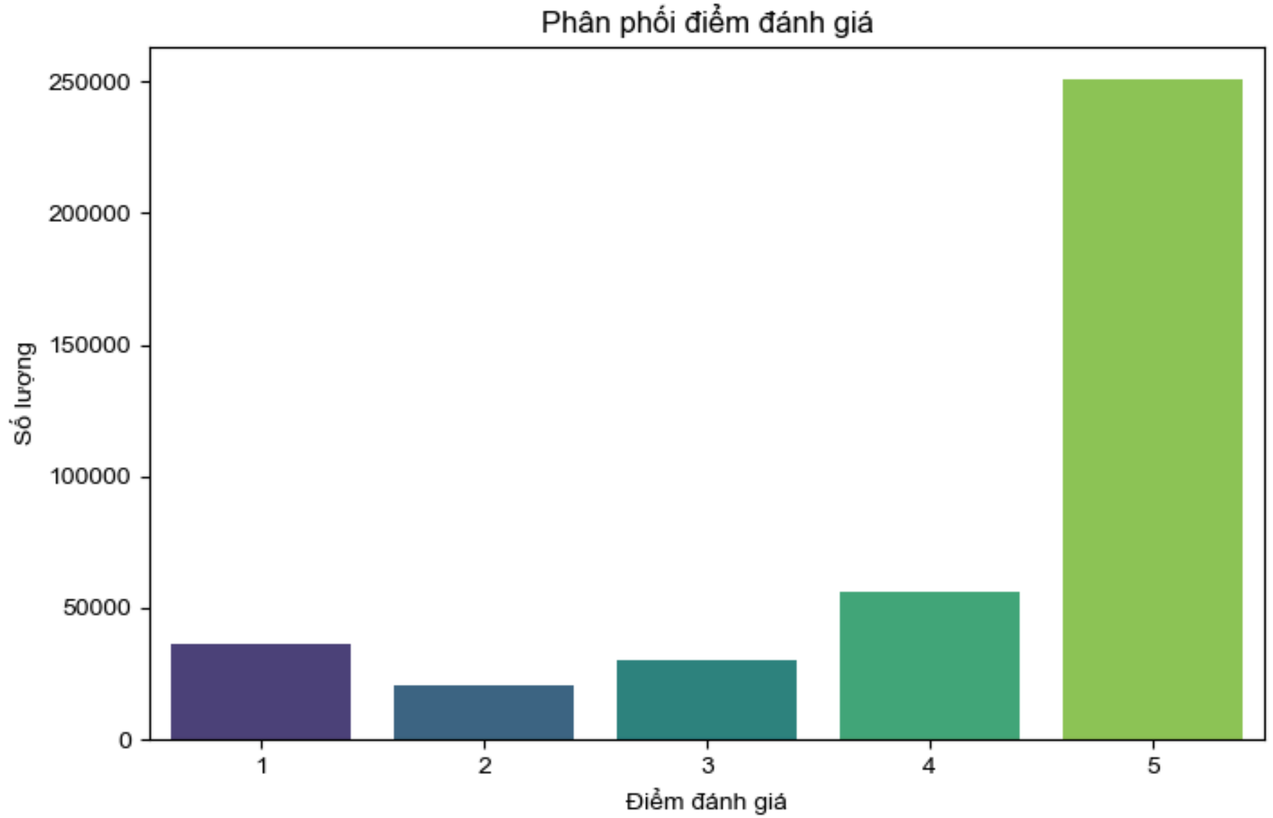
3. Tiền xử lý dữ liệu

Thực hiện tiền xử lý dữ liệu, làm sạch văn bản:
- Loại bỏ dòng trùng lặp
- Loại bỏ dữ liệu trống
- Xóa dấu câu, kí tự đặc biệt
- Loại bỏ stopwords
- Tạo cột mới Text_Cleaned chứa văn bản Text đã qua xử lý

Kết quả còn khoảng 393560 dòng đánh giá sau khi đã tiền xử lý.

4. Univariate analysis (phân tích đơn biến)

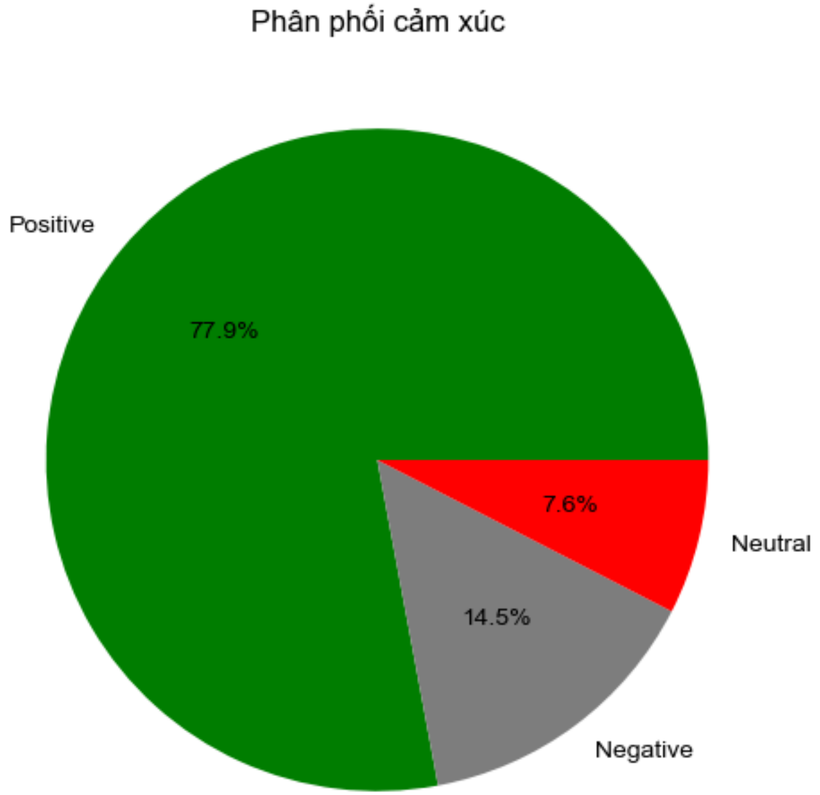
Mục tiêu: Kiểm tra phân phối của điểm đánh giá (Score) theo mức độ từ 1 đến 5 sao



- Điểm số đánh giá thiên về đánh giá 5 sao
- Các đánh giá 1 sao, 2 sao chiếm tỉ lệ thấp
- Khoảng hơn 250 000 đánh giá tốt, cho thấy khách hàng có trải nghiệm tốt với sản phẩm
- Tập dữ liệu có thể bị mất cân bằng khi thực hiện các mô hình học máy, cần xem xét việc cân bằng dữ liệu (resampling)

5.Sentiment distribution analysis (phân tích phân phối cảm xúc)

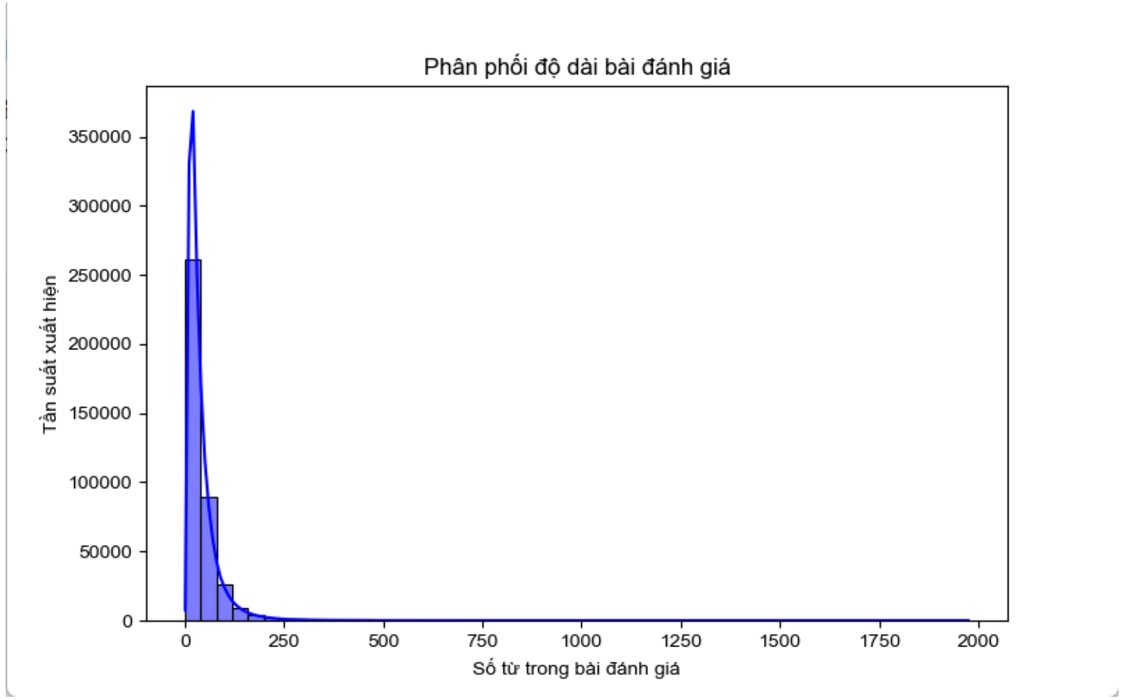
Mục tiêu: chuyển đổi điểm số thành các nhãn tích cực(positive), tiêu cực(negative) và trung lập(neutral)
Positive: Điểm đánh giá >3
Neutral: Điểm đánh giá =3
Negative: Điểm đánh giá <3

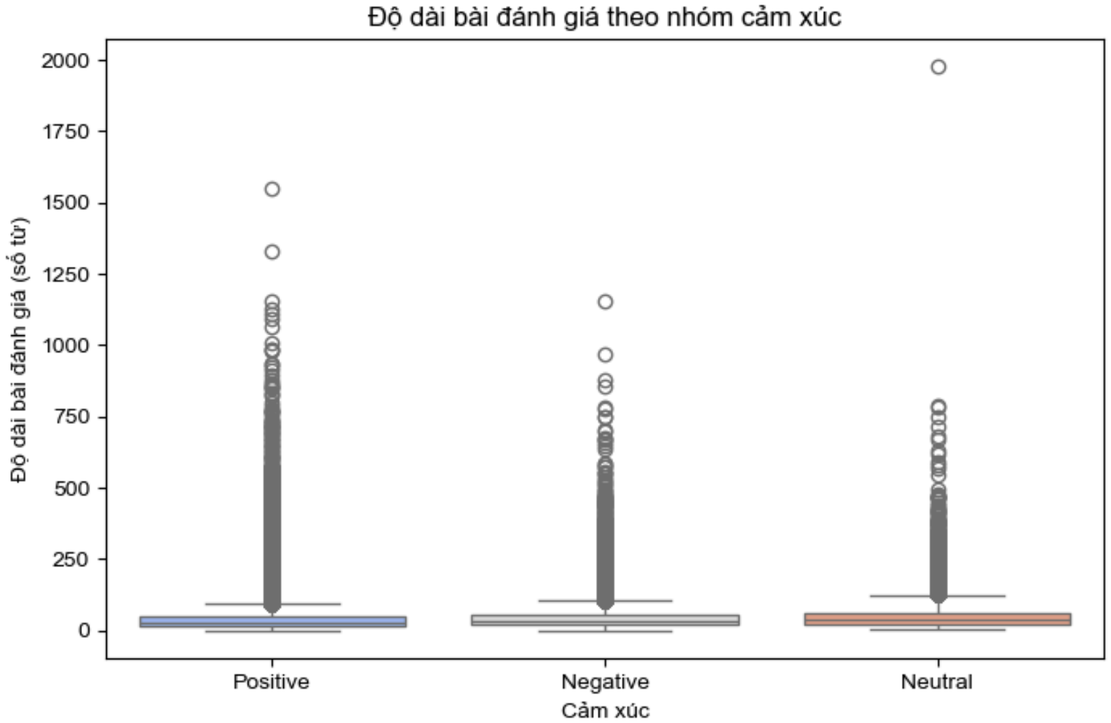


- Đánh giá tích cực (4,5) chiếm đa số với khoảng 77.9%
- Đánh giá tiêu cực (1,2) chỉ chiếm khoảng 14.5%
- Tỷ lệ đánh giá trung lập (3 sao) chỉ chiếm 7.6%, cho thấy phần lớn khách hàng có ý kiến rõ ràng thay vì đánh giá trung tính.
- Tập dữ liệu lệch về hướng tích cực, có thể gây ảnh hưởng đến mô hình dễ dự đoán tích cực hơn tiêu cực

6. Review length analysis (độ dài các đánh giá)

Mục tiêu: Kiểm tra mối quan hệ giữa độ dài đánh giá (Text) và điểm số (Score) của người dùng

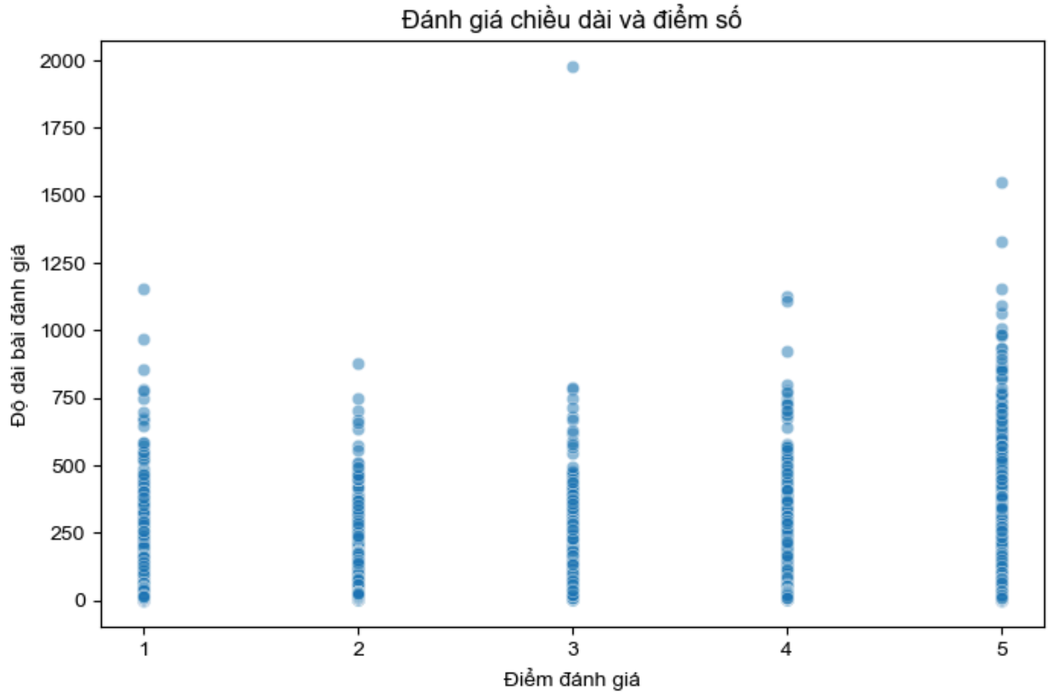


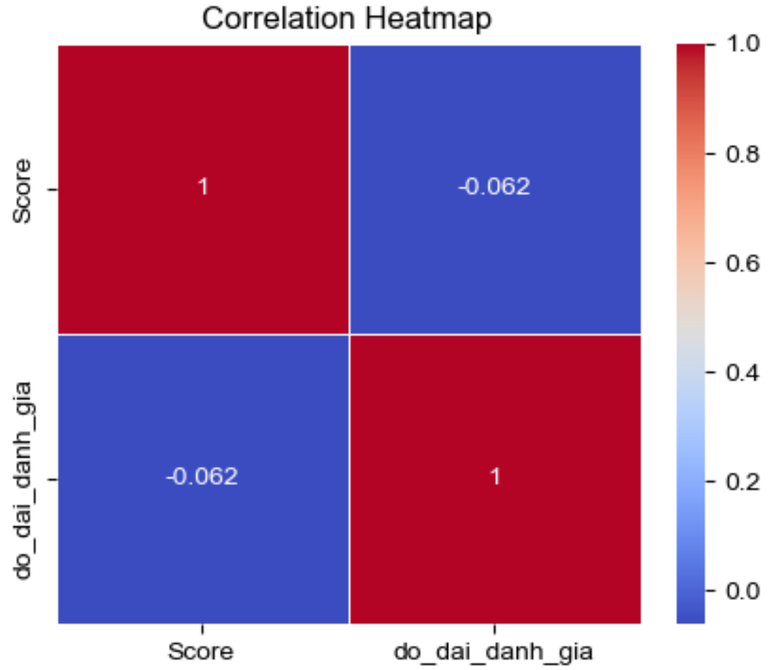


- Độ dài đánh giá tập trung vào 0-100 số từ.
- Một số bài đánh giá có độ dài rất lớn, có thể do spam
- Khi so sánh với nhóm đánh giá dựa trên các nhãn, các đánh giá tích cực có độ dài trung bình cao hơn tiêu cực
- Các đánh giá tiêu cực ngắn hơn (khoảng 30 ~ 50 từ)
- Một số đánh giá có dài rất cao, có thể là người dùng spam
- Người dùng có xu hướng viết đánh giá dài khi họ cảm thấy hài lòng với thực phẩm.

7. Correlation analysis (phân tích tương quan)

Mục tiêu: Kiểm tra các yếu tố có ảnh hưởng đến điểm số đánh giá của người dùng
Biểu đồ Scatter plot so sánh độ dài đánh giá(Text) và điểm số(Score)





- Biểu đồ phân tán các điểm dữ liệu phân bố rời rạc theo các mức điểm
- Không có xu hướng thể hiện mối quan hệ giữa độ dài bài đánh giá và điểm số.
- Biểu đồ Heatmap với hệ số tượng quan âm (-0.062) cho thấy không có mối quan hệ quá mạnh giữa độ dài đánh giá và điểm số.
- Dữ liệu không thể hiện các bài đánh giá dài hơn thì điểm đánh giá sẽ cao hơn.

8. Conclusion (kết luận)

Sau khi thực hiện các bước EDA ta rút ra được các ý sau:

- Dữ liệu chứa các đánh giá sản phẩm với các thông tin quan trọng như Nội dung đánh giá (**Text**), Điểm đánh giá (**Score**), và cảm xúc trong nội dung đánh giá (Sentiment).
- Dữ liệu đánh giá có **sự mất cân bằng:** phần lớn điểm số ở mức 5 sao, dẫn đến nhiều đánh giá có cảm xúc tích cực hơn tiêu cực.
- Điểm đánh giá 3 (**Neutral**) không đóng góp trong việc phân tích, có thể gây cản trở huấn luyện mô hình
- Số lượng từ trong đánh giá ngắn chiếm đa số (**dưới 100 từ**), tuy nhiên các bài đánh giá tích cực và tiêu cực có xu hướng dài hơn đánh giá trung lập.
- Không có mối tương quan rõ ràng giữa điểm số và độ dài bài đánh giá (hệ số tương quan **-0.062**), nghĩa là độ dài không quyết định mức điểm mà người dùng đánh giá.
- Có thể cần cân bằng lại dữ liệu hoặc áp dụng các phương pháp xử lý đặc biệt để đảm bảo các mô hình phân tích dự đoán chính xác hơn.

THANK YOU FOR WATCHING

THANK YOU FOR LISTENING

THANK YOU FOR WATCHING