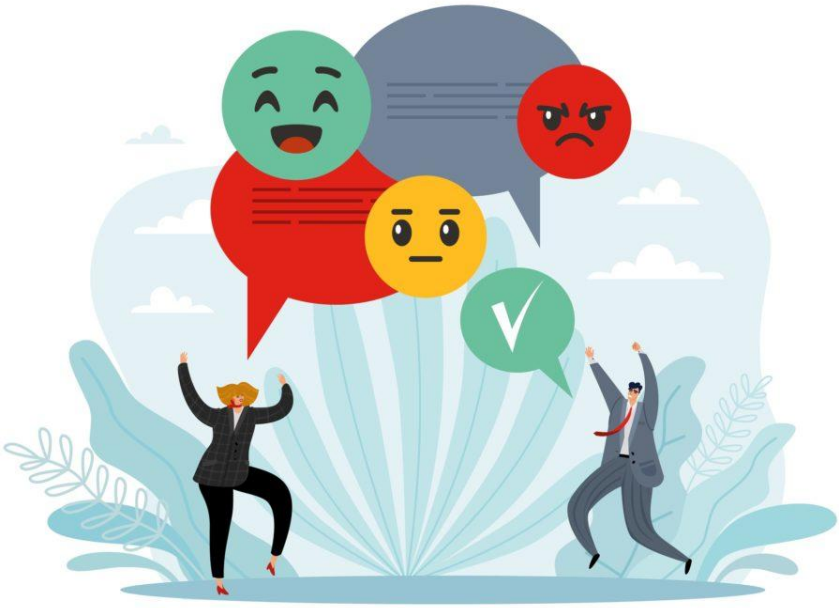
Shrey Patel

Project:

Sentiment Analysis of Amazon Fine Food Reviews



1. Introduction:-

This project involves performing sentiment analysis on a dataset containing reviews of fine foods from Amazon. The dataset spans over ten years, including ~500,000 reviews up to October 2012. For this analysis, I used a sample of 2,025 reviews. The goal is to extract insights into the sentiments expressed in the reviews and understand the patterns of customer feedback using various Natural Language Processing (NLP) techniques.

Sentiment Analysis in Python:

Doing sentiment analysis in Python using two different techniques:

1. VADER (Valence Aware Dictionary and sentiment Reasoner) - Bag of words approach
2. Roberta Pretrained Model from
3. Huggingface Pipeline

2. Exploratory Data Analysis (EDA):-

What is happening:

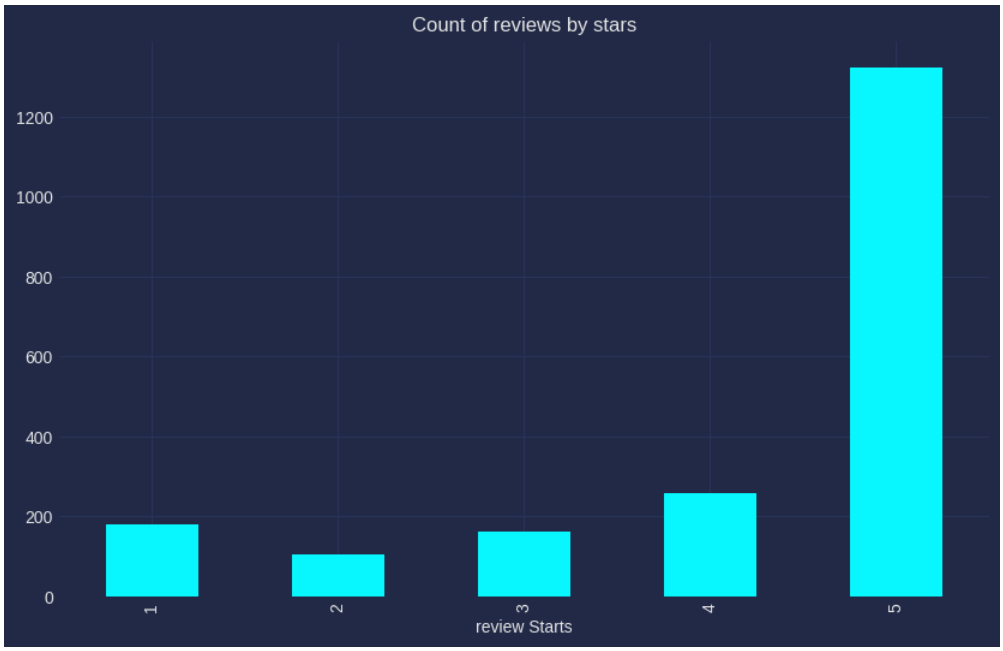
- We loaded the dataset and examined the distribution of review scores (stars) using a bar chart.
- The Score column represents the star ratings given by the reviewers.

How it helps: EDA provides an initial understanding of the dataset, identifying patterns and distributions that guide subsequent analysis. For example, plotting the count of reviews by star ratings reveals potential class imbalances.

```
[33] ## EDA

     ax = df['Score'].value_counts().sort_index().plot(
         kind='bar',
         title='Count of reviews by stars',
         figsize=(10,6))

     ax.set_xlabel('review Starts')
     plt.show()
```



3. Tokenization and Part-of-Speech Tagging:

What is happening:

- Tokenization splits text into individual words or tokens.
- Part-of-Speech (POS) tagging assigns grammatical categories to tokens.
- Named Entity Recognition (NER) identifies entities in the text.

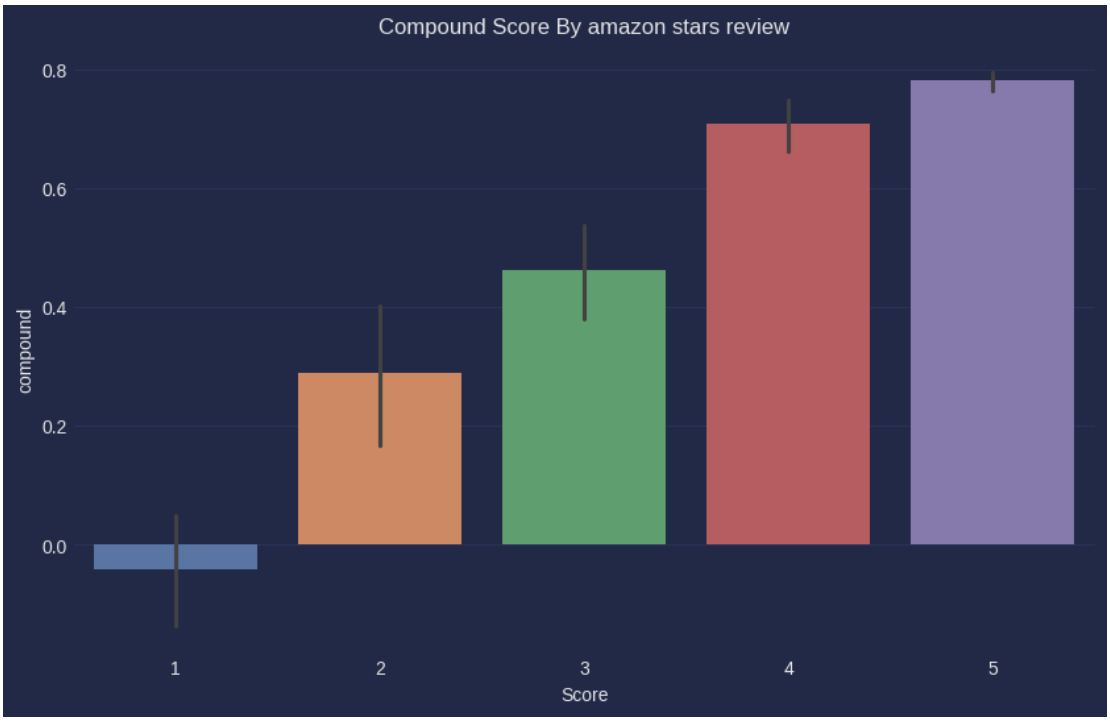
How it helps: These processes prepare the text data for further analysis by extracting meaningful features from raw text.

4. VADER Sentiment Scoring: -

What is happening:

- Using NLTK's SentimentIntensityAnalyzer, we calculated sentiment scores (negative, neutral, positive, and compound) for each review.
- This approach uses a Bag of Words method, where stop words are removed, and each word is scored.

How it helps: VADER provides an efficient and interpretable sentiment scoring system, especially suited for short text data such as customer reviews.

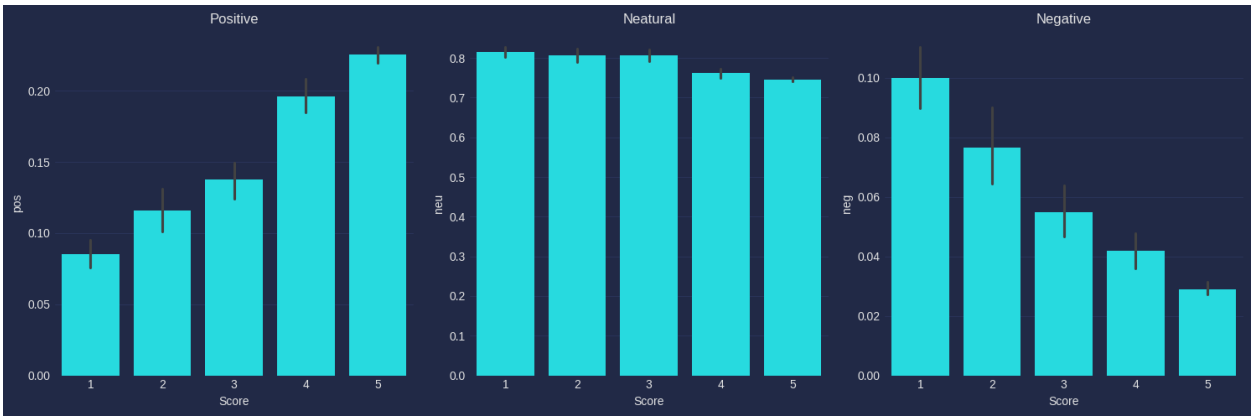


5. Visualizing Sentiment by Review Scores: -

What is happening:

- Bar plots display the average sentiment scores (positive, neutral, negative, compound) across different review star ratings.

How it helps: These visualizations reveal relationships between the star ratings and the sentiment intensity of the text, enabling us to verify if the sentiment aligns with the score.



6. RoBERTa Pretrained Model for Sentiment Analysis: -

What is happening:

- A transformer-based model (RoBERTa) is used for context-aware sentiment scoring.
- The model processes the review text and generates probabilities for negative, neutral, and positive sentiments.

How it helps: RoBERTa provides more sophisticated sentiment predictions by considering the context and relationships between words in the text.

From transformers import AutoTokenizer, AutoModelForSequenceClassification

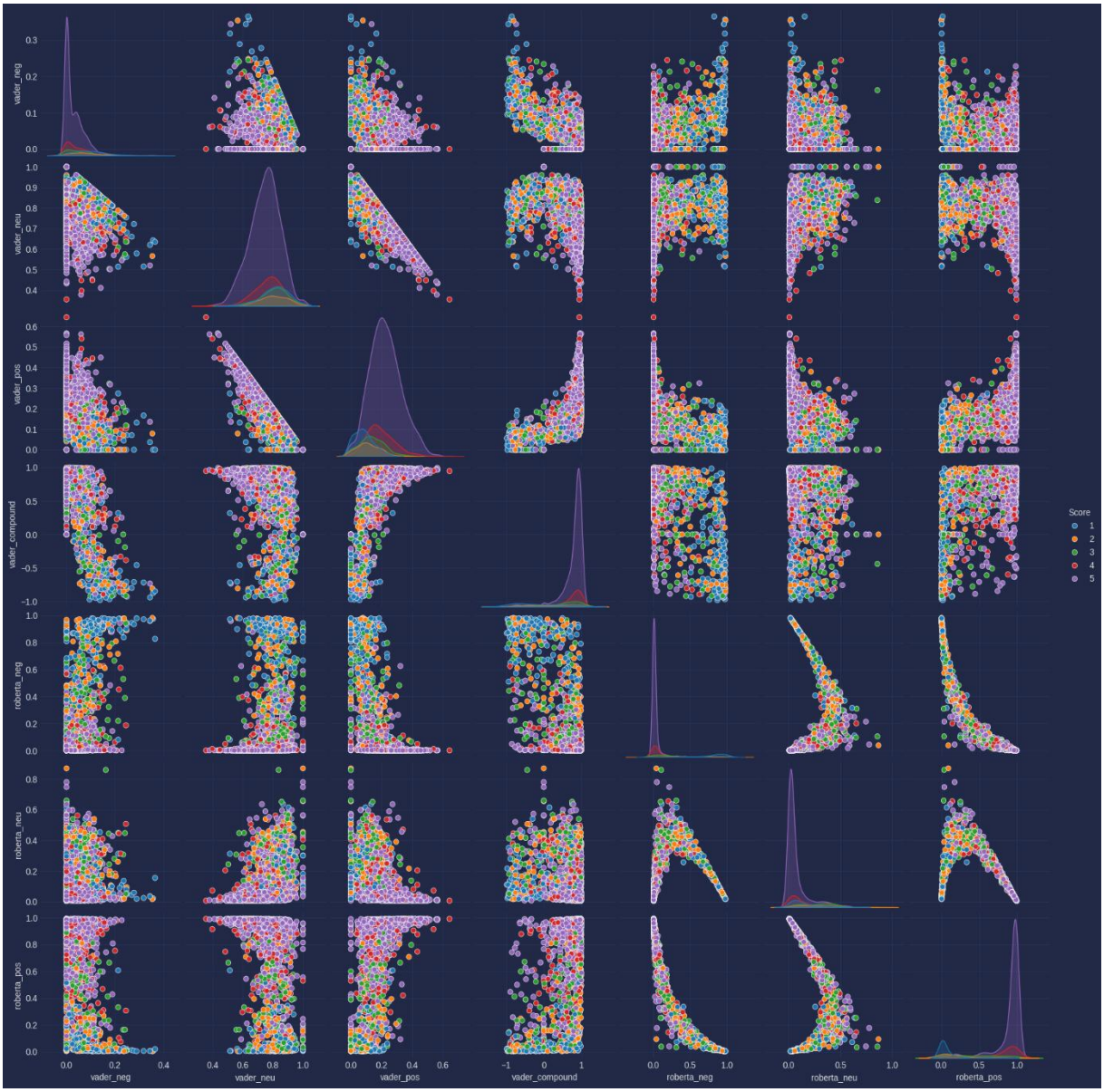
from scipy.special import softmax

7. Comparison of VADER and RoBERTa Models: -

What is happening:

- Both VADER and RoBERTa models were applied to the reviews to compute sentiment scores.
- Results were stored and visualized to compare their predictions.

How it helps: By comparing two models, we identify their strengths and limitations. VADER excels in simplicity and speed, while RoBERTa captures nuanced sentiment with higher accuracy due to its contextual understanding.



8. Future Scope: -

Automation:

- The process can be automated using APIs to fetch real-time reviews and classify them instantly.
- Pipeline integration with cloud services (e.g., AWS or GCP) can handle large-scale data.

Improvements with Advanced Technologies:

- Fine-tuning transformer models on domain-specific data (e.g., Amazon reviews) can enhance prediction accuracy.
- Leveraging unsupervised learning techniques like topic modeling can uncover hidden themes in reviews.
- Incorporating multilingual models to analyze reviews in different languages.

9. Conclusion: -

This project demonstrated how sentiment analysis, using VADER and RoBERTa, provides actionable insights into customer reviews. Visualizations and comparative analysis enhance understanding, enabling data-driven decision-making for product improvements and customer satisfaction.