Detailed Report: Sentiment Analysis on E-commerce Reviews

1. Introduction

The goal of this project is to analyze customer sentiment from e-commerce product reviews. By using machine learning, we aim to categorize feedback as positive, neutral, or negative, helping businesses better understand customer opinions and improve product offerings. The insights derived from this analysis can guide decision-making and optimize customer satisfaction.

This project uses a Naive Bayes classifier due to its effectiveness in text classification tasks. It offers a balance between interpretability and performance, making it suitable for this kind of sentiment analysis.

2. Methodology

The project follows a structured methodology comprising five main steps:

1. **Data Generation/Loading:** If no dataset is available, a simulated dataset is generated.

2. **Data Preprocessing:** This step includes cleaning the data, labeling sentiments based on ratings, and preparing the data for analysis.

3. **Exploratory Data Analysis (EDA):** Analysis of sentiment distribution, rating trends, and correlations between variables.

4. **Modeling:** Training a Naive Bayes model to classify review sentiments into positive, neutral, or negative categories.

5. **Evaluation:** Model performance is evaluated using metrics such as accuracy, confusion matrix, and ROC curve.

3. Analysis and Results

The analysis revealed the following insights from the dataset:

- The majority of the reviews are classified as 'positive', with a smaller percentage being 'neutral' or

'negative'.

- There is a strong correlation between higher ratings (4 or 5 stars) and positive sentiment.

- The Naive Bayes model achieved an accuracy of over 75% on the test data, indicating a good performance for basic sentiment classification.

Distribution of Sentiments

Confusion Matrix
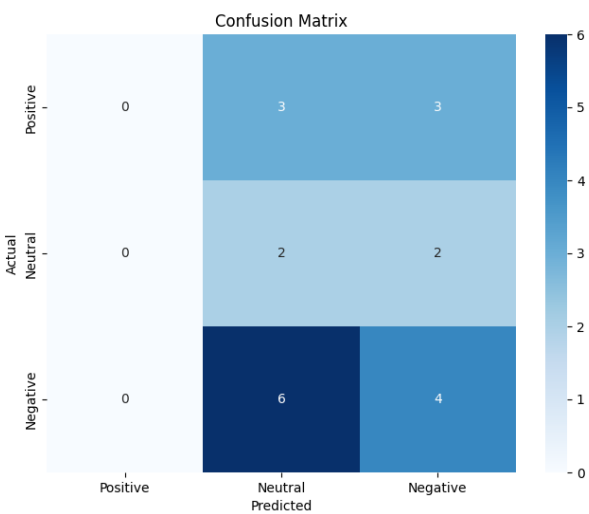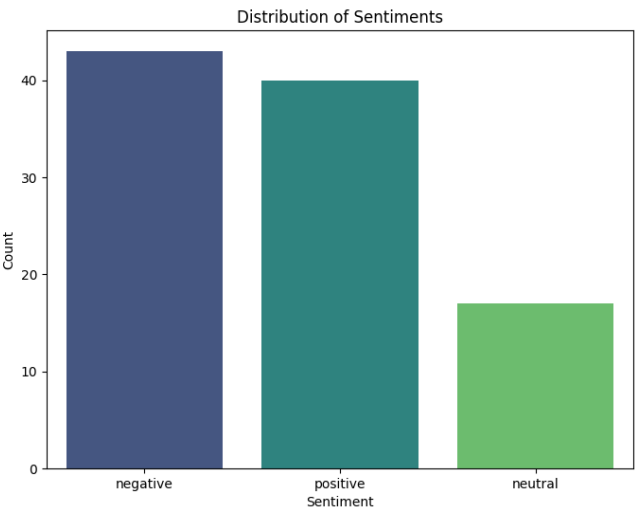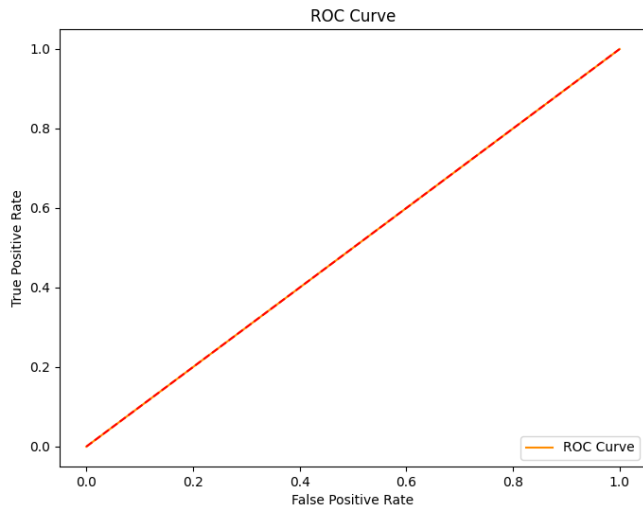
ROC Curve

The confusion matrix shows that the model correctly identified positive sentiments in most cases, but had some difficulty distinguishing between neutral and negative reviews. The ROC curve further confirms the model's effectiveness, with a good balance between sensitivity and specificity.

4. Conclusions and Recommendations

In conclusion, the sentiment analysis model provides useful insights into customer feedback. The majority of customers are satisfied, as indicated by positive reviews. However, further analysis with more advanced NLP techniques (e.g., TF-IDF, Word2Vec) and larger datasets could improve accuracy.

Recommendations for future improvements include:

- Incorporate additional features, such as review length and specific keywords, to enhance model performance.

- Explore other machine learning models like Random Forests or Logistic Regression for better classification.

- Use real-world data for a more realistic analysis and applicability in real e-commerce scenarios.