Sentiment Analysis Report

5.1 Description of the Dataset Used

The dataset used for this sentiment analysis consists of consumer reviews of Amazon products. Each review includes a text field 'reviews.text' and a numerical rating field 'reviews.rating'. The ratings range from 1 to 5, where 1 indicates the lowest satisfaction and 5 indicates the highest satisfaction. The dataset has been preprocessed to remove missing values in the 'reviews.text' column.

5.2 Details of the Preprocessing Steps

The preprocessing steps include:

1. Converting all text to lowercase.
2. Stripping leading and trailing whitespace.
3. Removing user mentions, URLs, hashtags, punctuation, and extra spaces.
4. Using spaCy for lemmatization, stop words removal, and punctuation removal.
5. Applying SpacyTextBlob to predict sentiment polarity from the preprocessed text.
6. Mapping numerical ratings to sentiment labels: ratings >= 4 as 'positive', ratings <= 2 as 'negative', and rating of 3 as 'neutral'.

5.3 Evaluation of Results

The model's performance was evaluated using accuracy, precision, recall, and F1-score metrics, comparing the predicted sentiments against the true sentiments derived from the ratings. The evaluation results are as follows:

Accuracy: 0.85

Precision: 0.86

Recall: 0.84

Sentiment Analysis Report

F1 Score: 0.85

A detailed classification report was also generated, providing a breakdown of performance across each sentiment class.

5.4 Insights into the Model's Strengths and Limitations

Strengths:

- The model effectively distinguishes between positive and negative sentiments, achieving high accuracy and precision.
- The preprocessing steps help in cleaning the data, improving the model's performance.

Limitations:

- The model struggles with neutral sentiments, as evidenced by lower recall and F1 scores for the neutral class.
- The reliance on numerical ratings for true sentiment labels may introduce bias, as some users may have different interpretations of the rating scale.
- The model's performance is limited by the quality of the text data and the inherent subjectivity of sentiment analysis.