Sentiment Analysis Report

Dataset Description

The dataset utilised for this sentiment analysis project is sourced from Kaggle, specifically the Consumer Reviews of Amazon Products. The original file, named "Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv," contains 24 columns and 5000 rows, with each row representing a customer review. The relevant columns for analysis include 'id' (unique product ID), 'name' (product name), and 'reviews.text' (text of the review).

Details of Preprocessing Steps

To prepare the data for sentiment analysis, the following preprocessing steps were applied:

1. **Removal of missing values**: One row containing missing values was removed from the dataset.
2. **Lowercasing**: All words in the reviews were converted to lowercase to ensure uniformity.
3. **Trailing whitespaces removal**: Any trailing whitespaces in the reviews were removed.
4. **Stop words removal**: Stop words were excluded from the reviews to enhance the efficiency and accuracy of the analysis.

Evaluation of Results

1. **Comparing two reviews:** A similarity score of 0.55 was obtained when comparing a review stating the product is "too small" with another stating it is "light and easy to use." The score accurately reflects the correlation between "too small" and the implication of being lightweight.

2. **Testing sample reviews:** The model produced positive polarities for positive reviews, close to zero values for most neutral reviews, and zero or negative values for most negative reviews. While the positive predictions align well, some polarities may seem lower than expected.

3. **Average polarity:** The model assigned positive average polarities to reviews for each product in the dataset, aligning with the overall positive sentiment expressed in these reviews.

Model’s Strengths and Limitations

Strengths:

- **Effective stop words removal:** Excluding stop words contributed to faster and potentially more accurate sentiment analysis.
- **Accurate reflection of positive sentiment:** The model demonstrated accuracy in assigning positive sentiments to positive reviews.

Limitations:

- **Polarity values:** Some polarity values, especially in positive reviews, might be lower than anticipated, indicating a potential area for improvement.
- **Limited dataset exploration:** While the entire dataset was analysed, a more extensive exploration could reveal additional insights into the model's performance.

In conclusion, the sentiment analysis model exhibits strengths in stop words removal and capturing positive sentiments. However, there is room for enhancement in accurately reflecting the magnitude of positive sentiments, and further exploration of the entire dataset could provide a more comprehensive evaluation.