Sentiment Analysis Report

Description of the Dataset Used

The dataset used in this analysis is the 'Amazon Product Reviews' dataset. It contains various reviews (5001) from customers about different Amazon products. Each review provides insights into the customer's product experience, including positive, neutral and negative feedback. The column 'reviews.text' was selected for this analysis as it contains the actual text of the reviews.

Details of the Preprocessing Steps

The following preprocessing steps were applied to the text data:

1. Lowercasing: All text was converted to lowercase to ensure uniformity.
2. Whitespace Stripping: Leading and trailing whitespace was removed.
3. Tokenisation and Lemmatisation: The text was tokenised and lemmatised using spaCy.
4. Stopword Removal: Common stopwords were removed to focus on the meaningful words in the reviews.
5. Non-alphabetic Tokens Removal: Tokens that were not alphabetic were removed.

Evaluation of Results

The sentiment analysis was performed on five selected reviews. The results are as follows:

- Review 1: 'item work easy read day light'

This review was determined as positive which is fairly accurate.

- Review 2: 'simple great buy money avid reader'

This review was determined as positive which is fairly accurate.

- Review 3: 'amazon kindle products reliable consistent quality different happy'

This review was determined as positive which is fairly accurate.

- Review 4: 'struggled surprise christmas gift year store sales rep suggested hit'

This review was determined as neutral which may be incorrect. This may have been due to the review discussing the product being a hit but also using the word ‘struggled’. With better removal of stopwords from the review, a more accurate result may have been found for this review.

- Review 5: 'bought mother law buying' – Neutral

This review was determined as neutral which is fairly accurate as there is not much suggestion on how the customer feels about the product.

Additionally, the similarity between Review 1 and Review 2 was calculated to be 63.33%. This is a fairly high score which is understandable because they were both considered positive reviews.

Insights into the Model's Strengths and Limitations

Strengths:

1. The spaCy model, combined with SpacyTextBlob, effectively identifies the sentiment polarity of the reviews.
2. The preprocessing steps help in focusing on the core content of the reviews, therefore improving the sentiment analysis accuracy.

Limitations:

1. The 'en_core_web_md' spaCy model includes word vectors, but the similarity results are still based on context-sensitive tensors, which may not give the most accurate similarity judgements
2. Sentiment analysis using this model may not be accurate for reviews with subtle or complex sentences.
3. The model's performance can be further improved by using larger spaCy models or adding custom word vectors.