Sentiment Analysis Report

1. Description of the Dataset Used

The dataset utilised for this sentiment analysis comprises Amazon customer reviews for their Kindle product. This dataset includes various pieces of customer feedback, detailing their experiences and opinions about the product. Each review contains textual content, which forms the basis for our sentiment analysis.

2. Details of the Preprocessing Steps

The preprocessing of the dataset involved several steps to ensure the text data was clean and ready for analysis. These steps included:

1. Tokenisation: Splitting the review texts into individual tokens (words).
2. Stopword Removal: Eliminating common stopwords, such as articles (a/an, the), prepositions (to, from, etc), conjunctions (FANBOYS), pronouns & relative pronouns, auxiliary verbs (to be, to have, to do), that do not contribute meaningful information to the sentiment analysis.
3. Lowercasing: Converting all tokens to lowercase to maintain consistency.
4. Punctuation Removal: Removing punctuation marks from the tokens to focus on the words.
5. POS Tagging: Identifying and extracting adjectives to specifically analyse sentiment-laden words.
6. Sentiment Analysis: Using SpaCy with the `spacytextblob` extension and NLTK to analyse the sentiment of each review, focusing on polarity and subjectivity scores.

3. Evaluation of Results

The sentiment analysis revealed the following key findings:

Key findings	
Mean polarity	0.39
Mean Subjectivity	0.57

These scores indicate that the reviews generally exhibit a moderately positive sentiment, with a fair degree of subjectivity. The most frequently used adjectives in the reviews were:

Adjective	Frequency (occurrences)
Great	1599
Good	670
Old	422
Best	328
New	269

The word "great" was notably the most common adjective, indicating a strong positive sentiment among reviewers. In contrast, the term "good" appeared significantly less frequently, highlighting a considerable gap between the two in terms of usage.

Additionally, the analysis identified the most recurrent negative adjectives used in the reviews:

Negative Adjective	Frequency (occurrences)
Little	219
Small	126
Black	56
Slow	45
Previous	44
Expensive	35
Bad	33
Dark	30
Difficult	29
Wrong	28

However, these adjectives appear to reflect the buyers' surprise and dissatisfaction with the actual size and colour of the product, which are clearly outlined in the product description on the page. Only a small segment of customers complained about technical issues related to the product.

4. Insights into the Model's Strengths and Limitations

Strengths	Limitations
Efficiency: The model rapidly and effectively processes and analyses large volumes of text data (max runtime ~ 2m 45s)	Contextual understanding: The model may struggle to analyse the context in which words are used, which can lead to misinterpretation of sentiment (for example, funny reviews). https://www.wordstream.com/blog/ws/2014/04/15/funny-amazon-reviews https://www.bigleap.com/blog/the-funniest-amazon-reviews-online-part-2/
Ease of interpretation: Polarity and subjectivity provide quantitative measures of text sentiment and bias that are easy to compare.	Subjectivity: The subjective nature of reviews means that different customers might express the same sentiment with varying degrees of intensity, which can affect the accuracy of the analysis.
Focus on adjectives: with the help of POS tagging, the model captures the essence of sentiment-conveying words, which are fundamental indicators of customer viewpoint.	Punctuation and Emojis: While punctuation is removed in preprocessing, certain punctuation marks and emojis can carry sentiment information that is lost in the current analysis approach.

Overall, the sentiment analysis model provides valuable insights into customer reviews, highlighting general sentiment trends and frequently used adjectives. However, there is room for improvement in understanding context and handling subjective expressions more accurately.