Sentiment Analysis Report

Dataset used

The Dataset used was a csv file containing Consumer reviews of Amazon product. The file stores the reviews themselves alongside considerable additional data, such as the name of the product being reviewed, name of the reviewer, date of the review and more. However much of this extra information is not needed for this particular task.

Preprocessing steps

The following make up the preprocessing steps used in the code:

1. **Importing Libraries**: The necessary libraries are imported including `numpy`, `pandas`, `spacy`, `SpacyTextBlob` for sentiment analysis, and relevant functions from `sklearn` for metrics and data splitting.
2. **Loading spaCy Model**: The English language spaCy model (`en_core_web_sm`) is loaded along with the `SpacyTextBlob` extension for sentiment analysis.
3. **Loading Dataset**: The dataset containing Amazon product reviews is loaded into a Pandas DataFrame. The dataset is assumed to be stored as a CSV file named "Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products_May19.csv".
4. **Text Preprocessing Function** (`preprocess_text`): A function is defined to preprocess the text data. This function tokenizes the text using spaCy, converts tokens to lowercase, removes stopwords, and punctuation from the text. The resulting clean text is then returned.
5. **Applying Text Preprocessing**: The `preprocess_text` function is applied to the 'reviews.text' column of the dataset, and the preprocessed text is stored in a new column named 'clean_text'.
6. **Removing Rows with Missing Values**: Rows with missing values in the 'clean_text' column are dropped from the dataset to ensure data quality.

Evaluation of results

The result show a positive sentiment across the reviews with 81.98% of said reviews being positive. There is still potential remove for improvement with 10.70% of reviews being neutral and a further 7.32% being negative.

Model strengths and limitations

The model offers strong feedback for general feeling amongst reviewers and therefore is useful for gaining a snapshot of how positively this collection of products are being received by amazon customers.

However the model is limited in the specific in can draw from the data, for example it dos not highlight common elements within negative reviews which may need addressing. It also crucially does not distinguish between the different products within the dataset, failing to highlight which individual products are reviewing especially well or poorly.