Description of the dataset

This report is based on dataset : Consumer Reviews of Amazon Products (Link: https://www.kaggle.com/datasets/datafiniti/consumer-reviews-of-amazon-products)

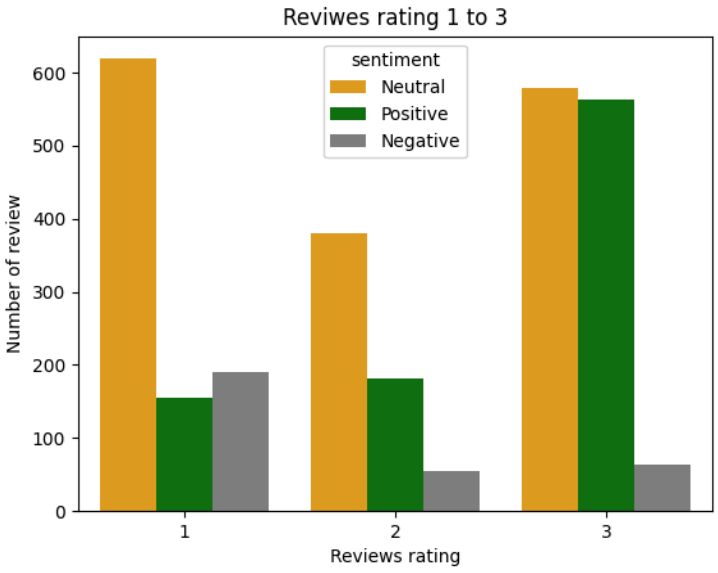
The dataset is a list of over 34,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database. The dataset includes basic product information, rating, review text, and more for each product.

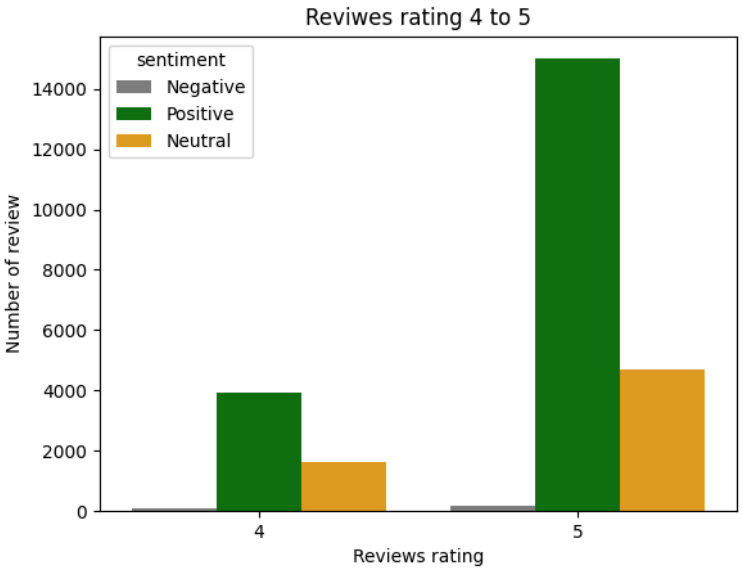
Details of the preprocessing steps

First, I have removed unused column including : dateAdded, dateUpdated, asins, categories, imageURLs, keys, reviews.didPurchase, reviews.dateSeen, reviews.id, reviews.sourceURLs, reviews.username, reviews.doRecommend, reviews.numHelpful, sourceURLs.
Since the current task is to focus on sentiment analysis on review, columns like dateAdded or imageURLS are not useful for now. Also, columns like ‘review.doRecommend’ and ‘reviews.didPurchase’ have too many empty values, I would not consider analysing these data without further information.

Then, the review is cleaned with the function preprocess_text and stored into the column “reviews.text.cleaned”. The function will remove whitespace, stopwords and non-alphabetic words which can enhance the accuracy when analysing sentiment.

Evaluation of results





From the above graph, the model did perform quite well as it reflected the review rating quite correctly. For more negative reviews, the number of Neutral and Negative reviews are much more than Positive reviews and vice versa.

Insights into the model's strengths and limitations

Strengths: The model did provide relatively correct information when analysing sentiment without any needs of extra training data. Also, it provide ways to identify stopwords which increase accuracy when analysing reviews.

Limitations: After analysing some outlier reviews, the review rating is 5 but sentiment analysis is Negative, I realise double negative sentences are one of the limitations.
Here is an example of an outlier review, *'You know, I really like the Amazon line of batteries. At first I has hesitant on using anything other than Duracell, I tried these and especially for the price you simply cannot go wrong....'*.
In the review, we can understand the customer like the product as the product works as intended. However, all the negative words like 'hesitant' and 'go wrong' could lead to more negative analysing results. In my own testing, by putting 'cannot go wrong' would result in a -0.5 in sentiment analysis although we, as a human, can understand it should be Positive sentiment.