Amazon Reviews Sentiment Analysis Capstone

The Dataset:

This dataset is a collection of 28332 amazon product reviews for various electronic products. There are 24 columns, but the most relevant of these for our purposes is the 'Reviews.text' column, which contains the review text as written by the users of the products.

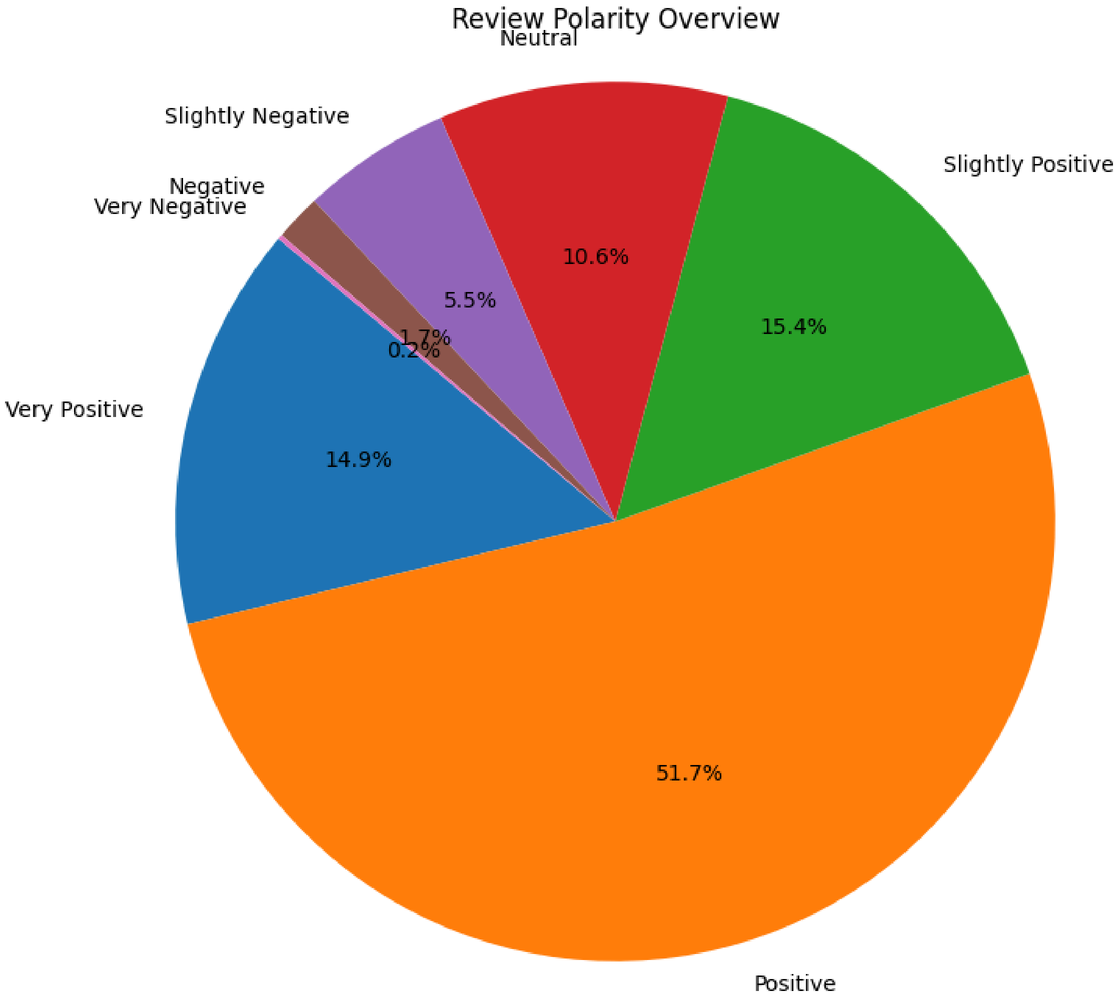
Preprocessing Steps:

Before utilising our data, we first loaded the entire dataset, then selected the 'reviews.text' column. We then dropped NaN values, as these are not possible to analyse and could skew our results. We next created a random integer between 0 and the length of the dataset (minus the dropped NaN values) which we used to select a random sample, using indexing. We stripped trailing whitespaces from the selected sample, tokenised the sample and then stripped all stop words to increase efficiency and accuracy. We finally stripped trailing whitespaces one more time, to ensure our analysis would work without issue.

Findings:

The program was run several times, and evaluation of its output suggested that it was broadly able to analyse sentiment, but that it struggled to properly interpret positivity. For example, the review 'Fantastic for reading in dark conditions. Very light and easy to hold.' was only given a value of 0.27, which seems low for a review that reads very positively. Similarly, 'Works well for the price point and is completely comparable to name brand batteries. I will be purchasing all of my future battery needs from amazon basics.' sounds incredibly positive, but the model gave this a polarity score of only 0.05. However, other reviews were accurately analysed, such as 'No better or worse than Panasonic or the like. Cheaper though.' which the function deemed 'slightly positive', returning a value of 0.05.

Given the vast number of reviews in this dataset, simply analysing a handful of examples did not provide a useful overview of the model's effectiveness, or even the general trend of the dataset as a whole. I therefore decided to analyse the polarity of all of the reviews, to divide these values into groups based on general sentiment (a scale of 7 values from very negative to very positive) and to display these groups as a pie chart. This gave percentage values for each of the groups, and can be seen below:



Model Insights:

The model is a useful way of analysing the sentiment of a piece of text, and is essential for processing large amounts of data. However, the model's accuracy leaves much to be desired; generally it can identify positivity or negativity in a text without problem, but it seriously struggles to accurately quantify this further. There are also several occasions where the model assigns no value to important words, such as 'recommend', which results in the review 'I would recommend this to other grandparents for their grandkids' being viewed as entirely neutral and given a score of 0. With rigorous testing of the model with data throughout the dataset it may be possible to gain insight into why the model struggles in certain cases, but due to limitations of the hardware being used, this is not feasible.