Capstone Project

5. Summary of the Amazon review sentiment analysis project

5.1. The used dataset listed 41421 consumer reviews for Amazon products such as the the Kindle, Fire TV Stick. The dataset includes basic product information, whether the product was purchased by the reviewer, rating, review text, review title, as well as information about the reviewing user, such as the user's city and the user's name.

5.2. For preprocessing and cleaning the dataset, initially the correct column, reviews.text, within the dataset was selected and it's data retrieved.
Then missing values were dropped from the dataset.
The data was further cleaned by converting the text to lowercase, removing any punctuation, asterisks and parentheses and any trailing whitespaces.
Finally, stop words were removed from the data set.

5.3. When using the sentiment analysis function and testing the model, it seems that for the majority of times, the model can correctly predict a review's sentiment. E.g. a review talking about how the product was "fine" was correctly classified as positive and a review talking about the product working "perfectly" was rated as very positive, whilst a review talking about "" was correctly classified as negative.

5.4. Limitations of the model can be found when looking at more elaborate reviews, where a reviewer compares the product with a different product. The model cannot distinguish between which token is referring to which product and therefore looks at the overarching sentiment of the review, even though parts of it should be discarded as they relate to a different product. However, overall, the model seems to have a high reliability when it comes to detecting the sentiment of a review, as the tests show.