NLP Capstone Project

5.1. A description of the dataset used

- The data was taken from a detailed CSV spreadsheet of Amazon product reviews
- The spreadsheet contained 34660 records (rows)
- Each record consisted of 21 pieces of data (columns)
- We were most interested in the 17th column which contained the `Review Text`

5.2. Details of the preprocessing steps

- I loaded the csv spreadsheet of Amazon consumer reviews
- I took the subset of the Reviews Text only, and discarded all the other columns
- I filtered out all the blank/empty reviews
- I striped the while space from the front and the end of the text
- I striped out all punctuation and special characters
- I made the whole text into lower case
- And finally I removed all `stop words` from the text

5.3. Evaluation of results

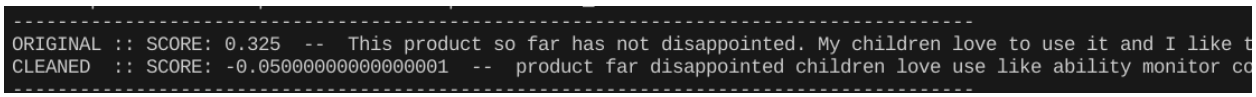
Overall the ability of the program to analyse each review and determine whether it expresses a positive, negative, or neutral sentiment was very accurate.

5.4. Insights into the model's strengths and limitations

The major strength of the model is to quickly analyse a large dataset and score whether the customers have expressed a positive or negative experience for their purchase without the intervention of a person reading the reviews.

However, I have come across one limitation of this model. I have observed that on some occasions the model has scored the sentiment as negative even though from the review text it is evident that the user has had a positive experience.

For Example:



I believe that this may have been caused because of the preprocessing step of removing all the stop words. On this occasion the word `not` was removed from the text, thus giving the opposite meaning.

- This product so far has not disappointed. My children .....
  vs
- product far disappointed children .....