Task 21 Sentiment Analysis Summary Report

Dataset used

The dataset used contains review data for the Fire HD 8 tablet. In this task I have focused on the column containing the review text data from the customer.

Preprocessing steps

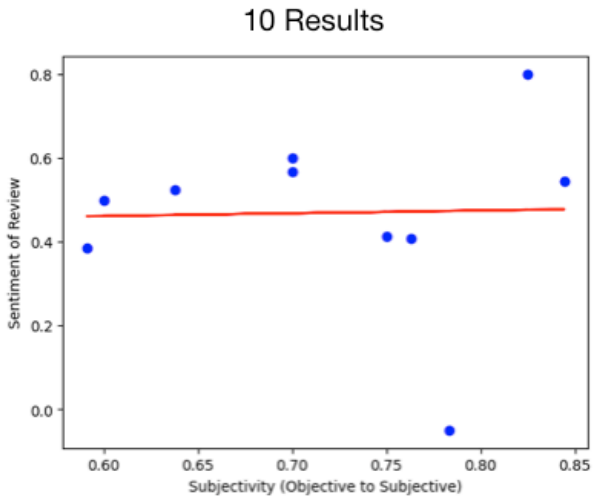
In preprocessing the data I began by removing any rows containing no data in the reviews.text column. I then selected a sample of data (as to not run all 46 thousand lines every time) and sent the raw reviews to one list, and the same review with stop words dropped, case set to lower and blank space at either end stripped, to another list.

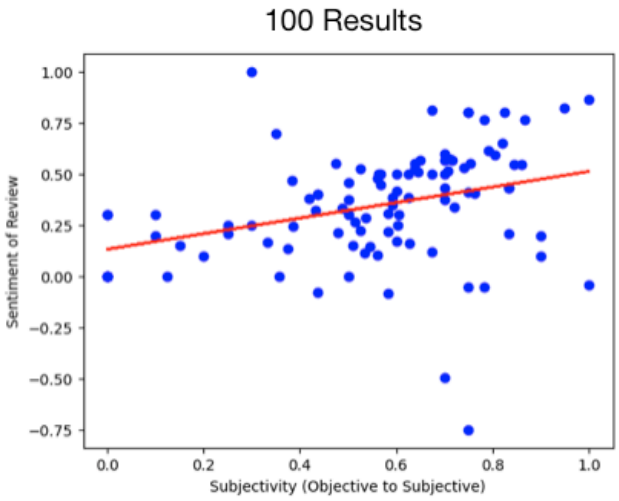
I then looped through our list of cleaned data, utilising the Spacy TextBlob package to retrieve sentiment, polarity and subjectivity results for each. Following that, I printed each raw review alongside the corresponding sentiment results and a breakdown of how each non-stop word was reviewed by the package, in order to analyse the results myself.

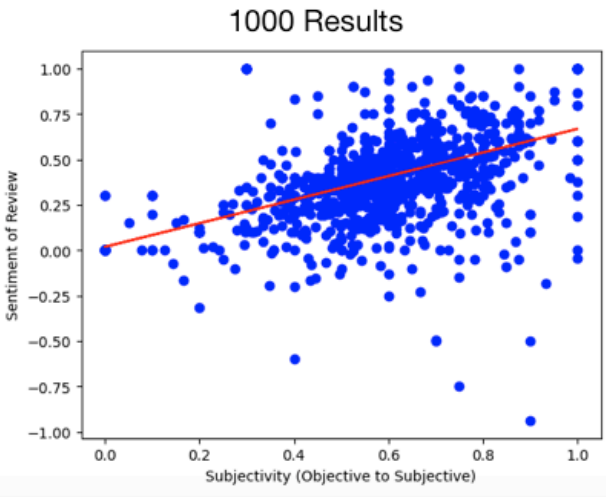
Evaluation of results

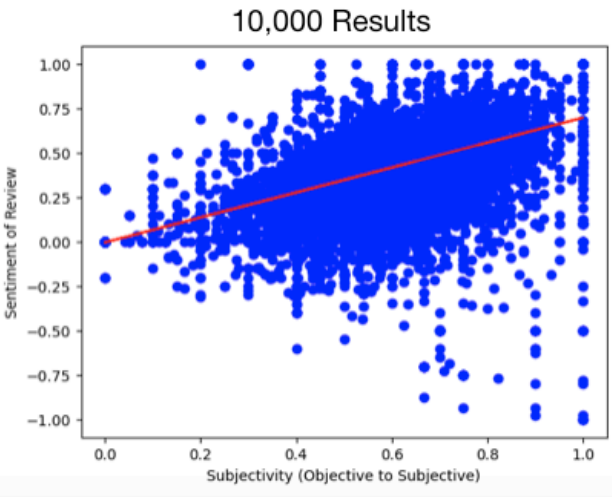
The results tend to score more positively in polarity than negatively, this indicates that in general user sentiment is positive towards the product.

After initially printing the above results I wondered if there was a relationship between polarity and subjectiveness, so I plotted a scatterplot of 100 and then 1000 samples. I began to notice a trend that people who used more subjective language tend to provide more positive reviews. To investigate this trend I ran a linear regression model for a line of best fit on the data across four data samples, 10, 100, 1000 and 10,000, shown below. The regression model supports the views and appears to not change much between 1000 and 10,000 further suggesting this is a general trend.









Model's strengths and limitations

The first thing I noticed when reviewing the results, which I thought may impede the reliability of the model, was that the word “not” is dropped as a stop word during the preprocessing stage. This appears to affect and sometimes invert the polarity values of some reviews, reducing overall accuracy.

If you run the programme included in the related python file, Review 1 is as follows:

“This product so far has not disappointed. My children love to use it and I like the ability to monitor control what content they see with ease.”

This is a positive review, however the model has dropped the “not” in front of “disappointed” and this has caused it to record the polarity as a negative value. We can see how this has affected the score specifically in the words processed section:

[(['far'], 0.1, 1.0, None), (['disappointed'], -0.75, 0.75, None), (['love'], 0.5, 0.6, None)]

Another example of this where the opposite happens appears in Review 11:

“Not easy for elderly users cease of ads that pop up.”

The model once again drops the “not”, reviewing only the below words and returning a positive score for a negative review.

[(['easy'], 0.43333333333333335, 0.8333333333333334, None)]

I’ve also noticed that within our model we are assuming all non-stop words in the review are related to our product. Take the results for Review 9 as example:

Great as a device to read books. I like that it links with my borrowed library e-books. Switched from another popular tablet brand and I am happy with the choice I made. It took some time to get books from my previous non-Kindle reader, but finally figured out a way!,

Sentiment(polarity=0.40666666666666673, subjectivity=0.7633333333333333)

Words processed:
[(['great'], 0.8, 0.75, None), (['popular'], 0.6, 0.9, None), (['happy'], 0.8, 1.0, None), (['previous'], -0.16666666666666666, 0.16666666666666666, None), (['finally', '!'], 0.0, 1.0, None)]

Review 9 is skewed unduly positive by the fact it has picked up the word “popular” although the user is referring to another product. It could also be argued, that it is skewed unduly negative by the word “previous”, which is being used in a sentence that expresses slight dissatisfaction at the process of transitioning to the product, but in itself doesn’t provide enough insight for us to understand the sentiment of this sentence as a whole.

Overall this is a fast and streamlined way for us to get a general understanding from a lot of data initially presented in a non-numerical format. It allows us to quickly review data which would otherwise takes a very long time by non-computational methods. It is however let down by its inability to compare words in relation to other words, it doesn’t understand context, and that could in some cases provide incorrect results lowering its accuracy as whole.