Task 21 – Capstone Project – Sentiment Analysis Report
Ben Pollins

Brief
To create a program capable of using sentiment analysis to categorise customer reviews of a product.

Data
The data used was a collection of over 34,000 customer reviews for Amazon products like the Kindle and Fire TV Stick downloaded from a public library on Kaggle.com. The review body was the only data used in the sentiment analysis.

Data Pre-processing
The review data was cleaned first by dropping all NaN values from the dataset, then by removing the stop words from the review and finally by converting all characters to lower case.

Results
I sampled 5 random reviews and categorised them using the defined algorithm – the results are shown in the table below (note: the un-processed review wording is shown for clarity)

No.	Review	Predicted Sentiment
1	What a excellent strean=ming player and with the 4k resolution goes great with my Samsung 4k tv	Very Positive
2	Have had it awhile and still learning and adding to it's skills but so far it's pretty good. I'll be adding the additional home automation features soon so we'll see how that goes.	Positive
3	Alexa works a lot easier than Siri, Alexa is more like an personal assistant. Works great!!	Positive
4	This is an older tablet. I probobly expected too much. If I'm going to deal with lagginess I prefer my ipad2.	Neutral
5	Great little streaming box. Better than the fire stick.	Positive

Evaluation
Generally, the model categorises the reviews correctly – 4 of the 5 predictions are correct and appropriately categorised. Review 4 is predicted as a neutral review; however, this is clearly a negative inditement on the product in question, suggesting the equipment is old and laggy and inferior to its competitors.

Insights
While it is clear from the results that the model can recognise positive and negative vocabulary to classify the reviews (e.g. ‘excellent’, ‘good’, ‘great’), it seems as though the model struggles with categorising more nuanced reviews that don’t use such clear language – such as Review 4. This error may also be a result of the jargon used in the review itself, words such as ‘lagginess’ and ‘ipad2’ are unlikely to be as common as other words in the model’s training data and therefore may not be fully understood by the model. ‘Lagginess’ refers to the poor performance of the device, while ‘ipad2’ is a reference to a competing product – if these words were fully understood, the model would likely recognise this review as negative.