NLP Capstone Report:

The dataset used was of a set of reviews for a product sold on Amazon. The dataset was very large and the initial dataset contained a lot of information not pertinent to the task such as URLs to the review and dates of the review posting. This information needed to be taken away by choosing only the “Reviews.text” column of the dataset (in which the actual text of the reviews was kept). It was also important to remove the entries into the dataset that did not contain the review text by using the dropna() attribute from pandas. There were not many entries that needed to be removed, however it was an important step to ensure the dataset could be analysed properly. It was then important to clean the data so that it could be analysed better. This was done by creating a function that would perform data cleaning on the chosen review that was passed through the function. The data cleaning that was done within this function was ensuring that the string was all lowercase so that there was no confusion that certain words were the same and the strip() attribute, which removed any white spaces so each review was seen to start from the first character. These cleaned review strings were then converted to an nlp doc so that natural language processing could be executed on it. The “stop words” were then removed from the reviews by iterating through the nlp tokens and removing the identified stop words using the is_stop() function. This removed any perceived unimportant words so that only the key elements of the sentence remained allowing for easier analysis. The remaining words were then rejoined back into a string and then once again converted to nlp tokens so that the sentiment analysis could be done on them.

The sentiment attribute analyses and finds the polarity and subjectivity of each review. The polarity is assigned on a scale of -1 to 1, where -1 is a negative review and 1 is a positive review. Therefore an if/else statement was set up so that it would assess the polarity rating and then assign a sentiment label to the review. The majority of the sample reviews analysed were labeled to be positive and only a handful were labeled to be negative. This could be as the amazon product has a high star rating and therefore most of the reviews will be positive. Some of the reviews could also be falsely labeled due to incorrect assumptions made by the algorithm.

The sentiment analysis was good at determining the sentiment of the individual words and whether there were more positive or negative words. This was good for the majority of reviews as it would assess whether they were positive or negative with fair accuracy. However, it was less accurate for reviews in which certain words that when put together would alter the meaning to be more negative or positive. For example, one review stated “I bought this in order to read books. I like using my iPad better though”. The removal of the stop words removed the word “though” and as it would assess the words individually it assessed the words “like” and “better” individually and therefore labeled this a positive review despite it being negative. The sentiment analysis also occasionally ignores certain words that would be important for acknowledging the sentiment of the review, such as one review which read “Alexa really amazes me. She learns new things everyday”. It only analysed the words “new” and “everyday” and thus concluded that the review was negative as it ignored the word “amazed”. The analysis also labeled “everyday” a negative word which is not necessarily true.