Mohammed Shahzaid - MS24020013136 - NLP Task 26

Task 5

5.1

The data set used is amazons data set recording the following information about their products.

```
Index(['id', 'name', 'asins', 'brand', 'categories', 'keys', 'manufacturer',
       'reviews.date', 'reviews.dateAdded', 'reviews.dateSeen',
       'reviews.didPurchase', 'reviews.doRecommend', 'reviews.id',
       'reviews.numHelpful', 'reviews.rating', 'reviews.sourceURLs',
       'reviews.text', 'reviews.title', 'reviews.userCity',
       'reviews.userProvince', 'reviews.username', 'preprocessed_text',
       'sentiment', 'polarity', 'subjectivity'],
      dtype='object')
```

5.2

Preproccessing steps involved the following:

- Iniitialising spacy which is a library that cleans the data.
- We then remove stopwords and also punctuation so that the data is reduced and cleaned.
- We lemmatise the data to reduce words further
- Finally combine the strings together back into a sentence which is now cleaned

5.3

From using the below code, we can see that the majority of feedback given is positive regarding the amazon products based on the analysis of the sentiment column.

```
sentiment
Positive    2074
Neutral      139
Negative      87
Name: count, dtype: int64
```

```
sentiment_counts = clean_data['sentiment'].value_counts()
sentiment_counts
```

5.4

Spacy is very easy to use and the code is very compact and readable. It is definitely useful when gathering insights into data quickly and whether or not to invest in a better NLP model.

The main issue I incurred was when using it against a large data set. The time to complete was incredibly slow. Other models may be quicker in time to compute.