QUESTION 5 - REPORT

1. Dataset used 1429_1.csv - I used this dataset as it was the smallest in size compared to the others.

   The Dataset contains details about various types of products sold by amazon e.g Kindle, Fire HD, Fire TV etc.

   The Dataset has 21 columns and 36600 rows, The columns contain various information related to the products being sold and the rows contain the different products being sold

2. Details of preprocessing steps:

   Data Preparation - Download the amazon dataset from Kaggle. I used the file named 1429_1.csv. This file was renamed to amazon_product_reviews.csv for this task. The 'reviews.text' column was selected from the dataset and its data retrieved.

   Loaded the small spaCy model for English language to enable natural language processing.

   Data Cleansing and Formatting- Removing any null values, stop words, punctuations and converting uppercase text to lowercase from the ‘reviews.text’ column in the dataset. The code uses the dropna() to remove any rows that have null values in it in order to improve the quality and accuracy of the data and avoid misleading or biassed results.

   Tokenization - Split the text into individual words/tokens. This process breaks down the sentences in the review into manageable units for easier analysis.

   Sentiment analysis - Applying the textblob component for spaCy to calculate the polarity and sentiment value of the text reviews. The code defines a function that takes a text review and returns its sentiment score and label based on the polarity value. The code also tests the function with other sample reviews.

   Term Frequency-Inverse Document Frequency (TF-IDF) -
   This process assigns weights/numerical vectors to words in the sentences based on their importance in the document and across the entire dataframe.

3. The polarity is a float value between -1 and 1, where -1 indicates a negative sentiment, 0 indicates a neutral sentiment, and 1 indicates a positive sentiment.

   The similarity score is a value between 0 and 1, where 0 indicates an objective statement and 1 indicates a subjective statement.

   These values allow us to see which products or categories have the highest or lowest customer satisfaction and identify any outliers or trends that may indicate customer preferences or issues.

4. Insights into model strengths and limitations:

Strengths of the model are:

The model is easy to implement and use, as you only need to add the textblob component to your spaCy pipeline to access the polarity and sentiment attributes.

The model can handle different types of data, such as reviews, tweets, comments, numerical values etc., and provide a fast effective way to measure their sentiment.

The model can also provide the subjectivity score. This score can be important in applications such as news and social media analysis, where it is important to identify whether a particular article or post is providing factual information or expressing a personal opinion.

The model allows fast automated processing of large volumes of text data making it suitable for handling the large amounts of amazon reviews.

Limitations of the model are:

The model cannot handle sarcasm or irony. A positive review could actually be a negative one

The model does not account for the intensity or degree of the sentiment, such as very positive or very negative, which may affect the accuracy and usefulness of the polarity score.

The model does not consider the structure or syntax of the text, such as negations, modifiers, or conjunctions, which may alter the meaning and sentiment of the text.

The model lacks contextual understanding. For example ‘battery life’ could refer to a genuine tablet or a toy tablet

The model needs robust preprocessing steps in order to remove misspelt words and abbreviations found in many product reviews.