Sentiment Analysis Task

5.1. A description of the dataset used.

The dataset contains data on customer reviews. It includes information such as rating, date purchased, date reviewed, username of the reviewer, review text and headers, name and type of product being reviewed, and much more. The dataset contains reviews only for Amazon products, and all of them are purchased from the BestBuy website. Most dataset features contain categorical values, including the date columns, which need to be converted to date types to be recognized. Another thing to keep in mind is that some features require selecting from predefined values, while others require manual input. For example, the product name is a set value, the rating requires selecting from options ranging from 1 to 5, and the review text requires manually inputting your opinion about the product.

5.2. Details of the preprocessing steps.

First of all, I dropped all the missing values. My text preprocessing approach is to remove stop words, punctuation marks, numbers, and named entities. Before doing that, I first need to exclude some important stop words. Words like 'not', 'good', 'bad', 'excellent', and many others could completely change the sentiment of the review, which is why we need to ensure that these words are not removed. Next, we create a function that will go through each review and remove any stop words, punctuation marks, numbers, and named entities. These words or characters are not useful in determining sentiment, and removing them could improve accuracy and time efficiency.

5.3. Evaluation of results.

Before discussing the results, I will explain the metrics used. We have 'polarity' and 'subjectivity'. Polarity ranges from -1 to 1, with -1 representing an extremely negative review and 1 representing an extremely positive review. On the other hand, subjectivity ranges from 0 to 1, with 0 representing an extremely objective review and 1 representing an extremely subjective review. I tested my project on three reviews: the 2nd, 11th, and 101st.

1.	Review 2:	Polarity: 70% quite positive review	Subjectivity: Quite subjective
2.	Review 11:	Polarity: 43% moderate positivity	Subjectivity: Quite subjective
3.	Review 101:	Polarity: 31% moderate positivity	Subjectivity: moderate subjectivity

5.4. Insights into the model's strengths and limitations.

Strengths: One of Spacy's strengths is that it supports multiple languages and provides various functions to easily preprocess text in those languages. It excels in identifying sentiments and finding similarities. Using it can increase both accuracy and time efficiency.

Limitations: First of all, we are using the small English language model. While this can increase time efficiency, it has a limited vocabulary, which could significantly decrease accuracy. The larger language models in Spacy can be more accurate, but they are less time-efficient. Additionally, even with Spacy's larger language models, there can be challenges with out-of-vocabulary (OOV) words. To address this, you need to manually add those words to the language model.