Report on NPL- Semantic Analysis:

Description of Dataset:
The dataset has 5000 rows and 24 columns. Each row represents a review from an Amazon customer about a product they have purchased (such as Kindle, Amazon Fire Stick etc). Each column provides basic product information, rating, review text, and more, for each product. These reviews were collected by Datafiniti's Product Database and the dataset was downloaded from Kaggle in CSV format.
Semantic analysis of Amazon reviews can be helpful in gauging public opinion of products, and identifying trends (e.g.: if certain product variants receive better reviews). This project performs exploratory analysis on the Amazon review dataset, demonstrating how a more in-depth study might unlock the data and trends embedded in natural language product reviews, in order to make recommendations on why products are popular.

Pre-processing Steps:
The column of interest, *'review.txt'*, was retrieved from the dataset for sentiment analysis. First, all rows with missing data were removed from the column. Next, a new column was formed and the function *'clean_data()'* was applied to each row and cleaned the data. This function converts the review text into lowercase, removes all stop words to make the analysis more meaningful, and removes any white spaces. Finally, it joins all the data together with a space in between each word.

Evaluation of the Results:
To analyze similarity, we compared 2 negative reviews between different products, 2 positive reviews between different products, and a positive and a negative review between the same product. The negative reviews (1 and 3) both got a negative/lower sentiment analysis score and the positive/higher reviews (2 and 4) both got positive scores.

Review	Review Text for Amazon Kindle E-Reader 6" Wifi (8th Generation, 2016)	Cleaned Text	Sentiment Analysis Score
1	*'I thought it would be as big as small paper but turn out to be just like my palm. I think it is too small to read on it... not very comfortable as regular Kindle. Would definitely recommend a paperwhite instead.'*	*'thought big small paper turn like palm think small read comfortable regular kindle definitely recommend paperwhite instead'*	-0.016667
2	*'Love my kindle, makes reading at night so easy and it feels like you are reading a real book.'*	*'love kindle makes reading night easy feels like reading real book'*	0.377778

Review	Review Text for Fire Tablet with Alexa,	Cleaned Text	Sentiment

	7" Display, 16 GB		Analysis Score
3	*‘I like the unit but battery seems to run down more quickly that my previous Fire. Also gets warm while charging.’*	*‘like unit battery run quickly previous fire gets warm charging’*	0.216667
4	*‘This device performs great! For the price it's really a steal. Affordable and easy to use’*	*'device performs great price steal affordable easy use’*	0.616667

Review Comparison	Similarity Score
1 and 3	0.76
2 and 4	0.49
1 and 2	0.65

Reviews 1 and 2 have high similarity, because despite their different sentiments they are still discussing the same product and so using similar language and terms. Reviews 1 and 3 and 2 and 4 have high similarity, because they have similar sentiments even though they are talking about different products. This shows that SpaCy NLP does not have context nuances.

We also calculated the average sentiment score for each product and plotted them against each other (Figure 1). This shows that the ‘All-New Fire HD 8 Tablet’ has the highest average sentiment analysis score out of all the products in the dataset and was the most positively reviewed, whereas the ‘Amazon Fire TV with 4K Ultra HD and Alexa Voice Remote’ has the lowest average sentiment analysis score and was the most negatively reviewed.

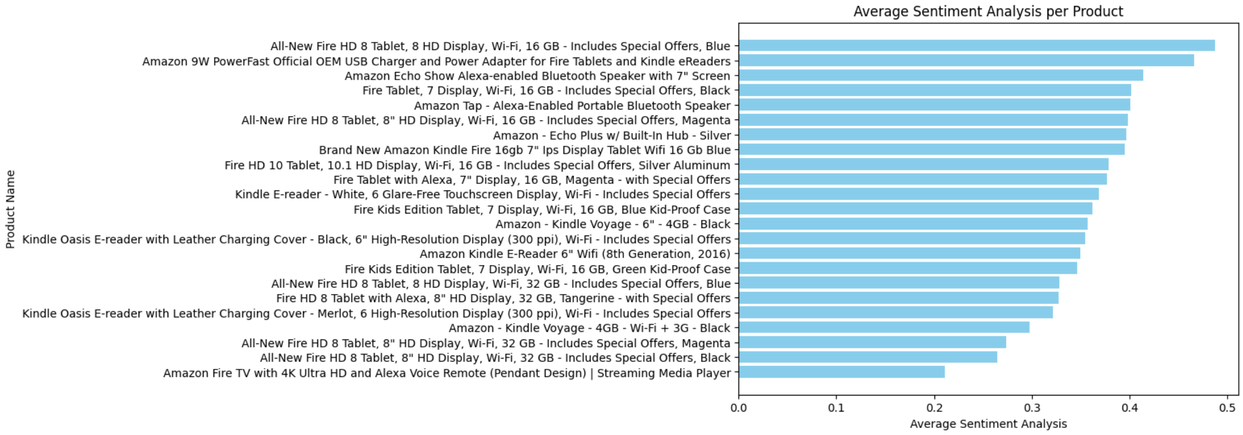

Figure 1: Average Sentiment Analysis per Product

Strengths and Limitation of the Model:

Strengths:

1. The ‘en_core _web_sm’ model is efficient, it requires limited computation resources making its application suitable for all users.
2. The ‘en_core _web_sm’ model is widely used, so it is easy to integrate sentiment analysis with other NLP.

Limitations:

1. The ‘en_core _web_sm’ model may not be capable of identifying more nuanced language and context. Therefore, it may not be as accurate as other models.
2. The ‘en_core _web_sm’ model may have limited vocabulary/lack domain knowledge resulting in inaccurate sentiment predictions.
3. The sentiment analysis we performed gave us the average sentiment score per product, however this did not indicate the total number of reviews, or the distribution of sentiment scores. Additionally, the date of each review was not considered - older positive reviews may be less accurate than more recent, negative reviews, e.g.: if the product has changed supplier and suffered a defect in quality.
4. When SpaCy removes the stop words from the reviews, it can completely change the meaning of the reviews in the cleaned reviews (See example below).

Text Review	Cleaned Text Review
‘Very cheap and was not impressed at all never again’	‘cheap impressed’