Description of the Dataset Used:

The data set has 34660 rows, and 21 columns but for the sake of this task, we will be focussing more on the “Reviews.text” column which is an object data type. Using print (dataframe["reviews.text"].isnull().sum()) I could tell I have just 1 missing data in that column.

Details of the Pre-processing steps:

After running my data set I first needed to clean my data of missing values and by doing that, I first created a new data frame from my existing data frame so I could always refer back to it after creating a new data frame, I used the drop method to remove the rows with missing data.

I also created a function to perform several pre-processing steps: Remove the use of special cases or characters, and keep only words and space, it also did tokenization with the Use of spaCy NLP library to process the cleaned text. Generates a list of lemmatized tokens from the processed text, excluding stop words and punctuation. Concatenates the lemmatized tokens back into a single string. Looped through each letter converted them to lowercase characters and stored them a list where processed or clean characters can be used with the use of append.

Evaluation of results:

After running the program using the below reviews, it was observed that there wasn’t any similarity between both sentences.

Review A: Inexpensive tablet for him to use and learn on, step up from the NABI. He was thrilled with it, learn how to Skype on it already...

Review B: I've had my Fire HD 8 two weeks now and I love it. This tablet is a great value.We are Prime Members and that is where this tablet SHINES. I love being able to easily access all of the Prime content as well as movies you can download and watch laterThis has a 1280/800 screen which has some really nice look to it its nice and crisp and very bright infact it is brighter then the ipad pro costing $900 base model. The build on this fire is INSANELY AWESOME running at only 7.7mm thick and the smooth glossy feel on the back it is really amazing to hold its like the futuristic tab in ur hands.

The similarity score of the two reviews: 0.897

Review: product far disappoint child love use like ability monitor control content ease

Sentiment Score: Positive

Review: great beginner experienced person buy gift love

Sentiment Score: Positive

Review: inexpensive tablet use learn step nabi thrilled learn skype

Sentiment Score: Positive

Insights into the model's strengths and limitations:

Using ‘en_core_web_sm’ couldn’t provide better accuracy and performance for tasks that benefit from richer linguistic representations like the task given.

The model is light weight but on my end it took time to load.