Brief Report / Summary : T21 - Capstone Project - NLP Applications

1. **A description of the dataset used.**
   - There are 24 columns (i.e. variables/features) and 5,000 rows (i.e. observations/records) in the dataset.
   - Of those 24 columns, one has the 'bool' (i.e. 'reviews.doRecommend'), one has the 'float' (i.e. 'reviews.id'), two have the 'int' (i.e. 'reviews.numHelpful', 'reviews.rating') and 20 have the 'object' data type.
   - Only four columns have null values: 'reviews.dateAdded' has 3,948, 'reviews.id' has 4,971, 'reviews.title' has 13, and 'reviews.username' has one.
   - The most important column for the purpose of this project is called 'reviews.text'. It has no missing values and has the 'object' data type.

2. **Details of the preprocessing steps.**
   - A user-defined function handles the preprocessing of single text reviews.
   - Using the 'doc = nlp(text)' statement, the function tokenises the review.
   - Then, employing list comprehension, it processes each token that is neither a stop word, punctuation, nor whitespace. Tokens that do not fall into one of those categories are lemmatized, converted to lowercase, and cast to string.
   - Lastly, the function joins all the tokens that were added to the list and returns them as a string.

3. **Evaluation of results.**
   - Polarity scores range from -1 to 1, where -1 is negative, 0 is neutral and 1 is positive. Subjectivity scores range from 0 to 1, where 0 is objective (factual) and 1 is subjective (opinionated).
   - The results of the sentiment analysis are mixed. Some reviews were correctly evaluated, while there are others where the polarity and subjectivity scores do not match up with the sentiment. For example, there is a clearly positive review with a negative polarity score as well as one exceptionally opinionated review with a subjectivity score of 1.
   - Similarity scores ranges from 0 to 1, where 0 implies complete dissimilarity and 1 suggests identicalness.

- In my opinion, the two reviews compared would not warrant a similarity score as high as the one calculated of 0.801. This is because ‘Review A’ is positive, while ‘Review B’ mentions suboptimal battery life which is negative.

4. **Insights into the model's strengths and limitations.**

- Strengths
    - Leveraging the spaCy NLP library, which is optimized for speed and accuracy, and using the small spaCy English language model provides efficiency.
    - The application of the TextBlob library, including its polarity (positive vs negative) and subjectivity (opinionated vs factual) attributes, provides nuance to the results.
    - The ‘preprocessing’ function removes stop words, punctuation, extra whitespace, and performs lemmatization, improving data quality for sentiment analysis.
- Limitations
    - While TextBlob is useful, it provides relatively simple sentiment analysis, which might not be able to accurately handle more nuanced linguistic expressions and sarcasm.
    - The model uses pre-existing components and does not involve model training. This limits its ability to learn and adapt to the specific vocabulary and sentiment patterns of the dataset used.
    - The use of the small spaCy model for sentiment analysis limits the quality of the results produced. Using a larger model would take more time and computing power but would probably deliver better results.