Report for T26 - Natural Language Processing with SpaCy

Student Name: Wai Man, PUN
Student Number: MP24020014526

Description of the dataset used

This is a list of over 34,000 consumer reviews for Amazon products like the Kindle, Fire TV Stick, and more provided by Datafiniti's Product Database. The dataset includes basic product information, rating, review text, and more for each product. For this task, the feature ‘review text’ is used for users’ sentiment analysis with SpaCy NLP model.

Details of the preprocessing steps

1. Remove all the review texts with null value, to avoid error in analysis. There is no analysis that can be done for null text.
2. Remove stopwords and punctuation with SpaCy. SpaCy evaluates if a word is a stop word or not. In theory, stopwords are common words that do not add much meaning to a sentence, which does not help with the analysis.
3. Lemmatize the text with SpaCy.
4. Apply basic text cleaning strip() and str() to remove heading space and change the data type to string, which helps on SpaCy’s Named Entity Recognition and avoid type error.

I tried to change the text to lower / upper case to see if it has any impact on the analysis. From the attribute of SpaCy’s Named Entity Recognition (NER), changing cases can reduce the ability of identity recognition. Therefore, changing cases are not applied.

Evaluation of results

The sentiment is correct in most of the samples taken for analysis. The polarity may not be very accurate compared between the reviews. However, it is usually correct to interpret the positive sentiments. Just some cases like below are not accurate:

Below case should be positive. However, the model should have bias on online games and rates it as -0.4.

index: 58
Review: My daughter likes this tablet to play her online games!
Preprocessed Text: daughter like tablet play online game

Review Rating: 5.0
Sentiment: Negative
Polarity: -0.4

Below case should be positive. However, the model rates it as neutral.

index: 13
Review: Simply does everything I need. Thank youAnd silk works wonders
Preprocessed Text: simply need thank youAnd silk work wonder
Review Rating: 5.0
Sentiment: Neutral
Polarity: 0.0

Below case should be negative. However, the model rates it as positive. Because the stopword ‘not’ is removed in data pre-processing.

index: 10
Review: Not easy for elderly users cease of ads that pop up.
Preprocessed Text: easy elderly user cease ad pop
Review Rating: 4.0
Sentiment: Positive
Polarity: 0.43333333333333335

Below case should be overly rated compared with other positive reviews.

index: 48
Review: Tablet is perfect for beginners who just want basic
Preprocessed Text: Tablet perfect beginner want basic
Review Rating: 5.0
Sentiment: Positive
Polarity: 0.5

Insights into the model's strengths and limitations

1. The model is mostly correct in the analysis, which is helpful for a large amount of review analysis which does not need very accurate results and saves a lot of time to review manually.
2. It is challenging for the model to understand the review’s meaning and compare the polarity between reviews. The polarity may not be very accurate.
3. It takes longer time in analysis compared with other simpler learning models.
4. It may give bias from the data it learnt from, just like negative sentiment on online games.

5. Removing Stopwords may not be helpful in the analysis. From the result, ‘Not’ is removed and the model wrongly rated as positive. Having an experiment without removing Stopwords, the model still works well.