Sentiment analysis report

Introduction

The dataset used were reviews of various Amazon products. In order to perform semantic analysis on a review inside the spreadsheet, I had to firstly remove all of the empty entries inside the spreadsheet. This was done by using the dropna() function of the pandas module. Next was removing the stop words such as ‘the’, ‘is’, etc as they have no meaning. The final step for preprocessing was converting all characters to lowercase and removing whitespace for the NLP to make the sample easier to analyse. On top of using the sentiment attribute to output the sentiment score, I have also implemented the polarity and assessment attributes into my function to provide more detailed results to the user. For clarity’s sake, the sentiment attribute outputs a float between 0 and 1. The closer it is to 1 the more positive the document is deemed. The polarity attribute outputs the sentiment score and the subjectivity score. The subjectivity score also gives a float between 0 and 1, with the higher the number, the more subjective it is. The sentiment assessment attribute returns a list of assessed tokens in the document and their given polarity and subjectivity scores. The sample method of the pandas module was used to test random reviews in the spreadsheet.

I tested my sentiment function on 3 reviews:

Sample reviews

The first review tested was the 4th row in the reviews column which reads: “This make an excellent ebook reader. Don't expect much from this device except to read basic ebooks. The good thing is it's cheap and good to read in the sun”. My sentiment function gave it a polarity score of 0.599 and a subjectivity score of 0.605. This is a positive score and it appears very logical as this review’s tone is very positive with words such as ‘excellent’. Assessed tokens like ‘excellent’ received the highest polarity score of 1 and ‘good’ tokens were recorded at 0.7. Something interesting is that the token ‘basic’ returned a polarity score of 0 and a subjectivity score of 0.0125 while ‘cheap’ got a score of 0.4 and 0.7 respectively. At first I found this odd as cheap has negative connotations while basic is more neutral language but upon seeing the high subjective score I realised that it can imply something positive or negative. ‘cheap’ in this context is used as a compliment but its score suggests it’s used negatively.

The second review tested was the 19th one. It reads “I was looking for a kindle whitepaper. I saw online for $80. What a deal. I ordered it online and picked it up in the store. I got it home and couldn't adjust the brightness. After a lengthy time with online customer service I called customer service. After 20 minutes with speaking to a female Elmer Fud that doesn't speak english well I figured I would just return it. Although it looks Identical to the $120 model, you can not adjust the brightness. That would have been good information before I bought it.” The overall tone is very negative with very little compliment given for the product. The polarity score returned was 0.2333 and subjectivity score of 0.255 which appears to be very accurate.

The final review tested was the 28th one. It reads “ I have several Kindles so knew what to expect. It is a great reading device and I keep several in different locations.” This review is quite ambiguous. The product may be described as ‘great’ but the language is definitely a lot more general. The polarity score was given a 0.4 and a pretty high subjectivity score was 0.675. I feel that the polarity score is low but the subjectivity score makes a lot of sense.

Conclusion

Overall I found that my sentiment function provided solid results. I feel as if though that results become slightly questionable when the language of the review isn’t obvious and simple. When there is a sarcastic tone or the writer talks about other products, the model may struggle to pick up on it. Another limitation, is the model not being able to decipher the correct meaning when assessing words that can mean different things in different contexts. An example that I highlighted earlier was the word ‘cheap’ that has opposite definitions depending on the context. The use of the implementation of the assessment attribute and, more specifically, the subjectivity score helps to mitigate this and explains score in some cases.