Natural Language Processing with SpaCy

Turning Unstructured Natural Language into Structured Natural Language with Understanding Natural Language

The word キリキリ (kirikiri) is a Japanese onomatopoeia, and it can mean "quickly." An example would be: "Come kirikiri and walk." In the film Audition (1999), Eihi Shiina uses kirikiri, which was translated as "deeper, deeper," but a more apt translation would be "ouchy, ouchy."

Side note: We don't say "ouch" when we stub our toes but instead use another word that can have many meanings.

If this project involved translation, it would also require natural language generation. But with sentiment analysis, the model just needs to understand.

The dataset used is: Consumer Reviews of Amazon Products. The dataset has 21 columns and 34,660 rows. The number of rows drops by one in the pre-processing stage due to null data. The columns include reviews.dateSeen, asins, and reviews.userCity. The only data we are using is reviews.text.

In the pre-processing stage, all rows with a null value in the column reviews.text were dropped. Then, Python string manipulation was used to convert all text to lowercase and strip whitespaces. Finally, spaCy uses its English core medium to remove stop words and parse the review.

An example of the parsing is:

"Excellent product. Easy to use, large screen makes watching movies and reading easier."

Becomes:

"excellent product. easy to use, large screen make watch movie and read easy."

When stop words are removed, this changes to:

"excellent product easy use large screen make watch movie read easy"

We have examples of stemming, but other reviews use lemmatization. See:

https://spacy.io/api/token.

For the Sentiment Analysis, we are going to use TextBlob. Sentiment has polarity and subjectivity.

https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment
Polarity gives a score from -1 to 1, with 1 being the most positive score. Subjectivity ranges from 0 to 1.

Natural Language Processing with SpaCy

Example reviews:

"Light, portable and great features. Highly recommended."
Polarity Score: 0.4533
Subjectivity Score: 0.6633

"I bought this tablet on Black Friday, when it was on special for my little girl. I wasn't expecting much from the tablet... boy was I surprised by its capabilities! Love the tablet!"
Polarity Score: 0.1257
Subjectivity Score: 0.5262

The results of the analysis are disappointing. The second review reads as very positive, but due to a negative word being used, it gets a low score. The use of the word ‘boy’ on this occasion acts as a conjunction, changing the meaning of the whole review.

The model as a whole gives better accuracy with longer reviews. For a single short review, it isn’t something I would want to use. But the more reviews, the better the average. It highlights market sentiments but doesn't provide detailed insights.

Example review:

"This is my 4th Kindle - I use it as a tablet when I travel."
Polarity Score: 0.0
Subjectivity Score: 0.0

I can't say if the Kindle is good or bad. Is it the 4th Kindle because they keep breaking?
Finally, there is sarcasm. Below are two snippet reviews:

"Winnie the Pooh: Blood and Honey is history's greatest cinematic masterpiece!"
"Blood and Honey is by far the most detailed, deep, and moving film I've ever had the pleasure of looking at with my own two eyes."

I don't think the model would mark them as negative.