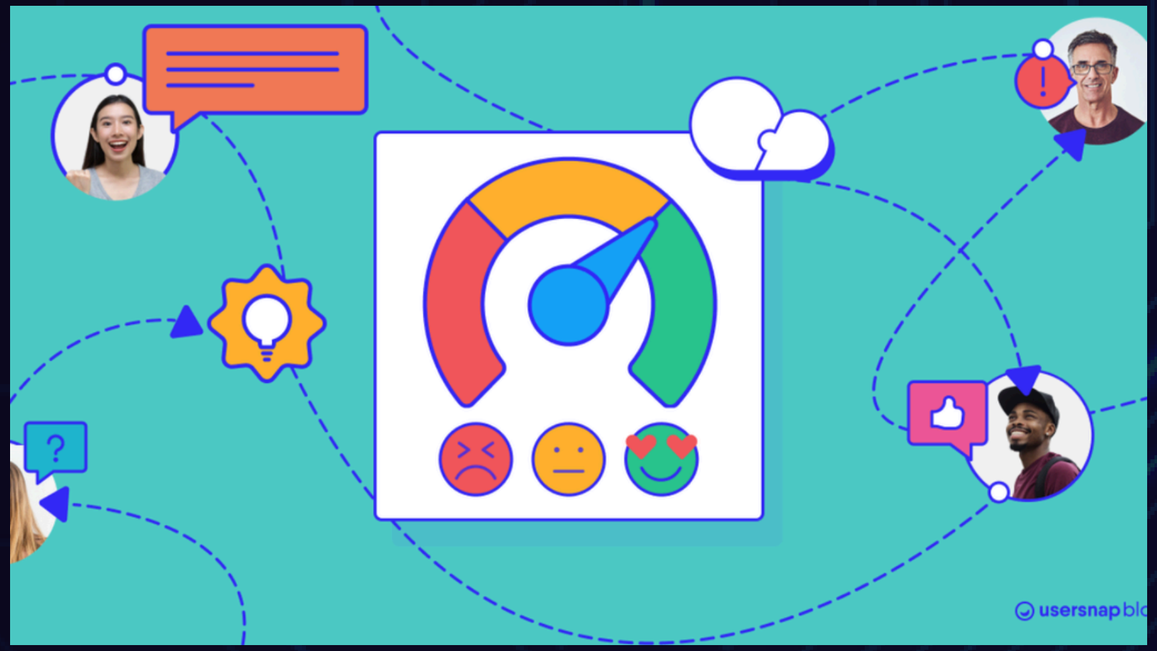
SENTIMENT ANALYSIS



WHAT IS IT

Sentiment analysis or opinion mining is the use of natural language processing and text analysis that allows the user to take a text and feed it into an AI and through machine learning, output what the AI thinks is the most likely sentiment with the most commonly used sentiments being Positive and Negative.

METHODOLOGY

01

Look for open-sourced models that have already been pre-trained with data as they have been trained with more comprehensive datasets

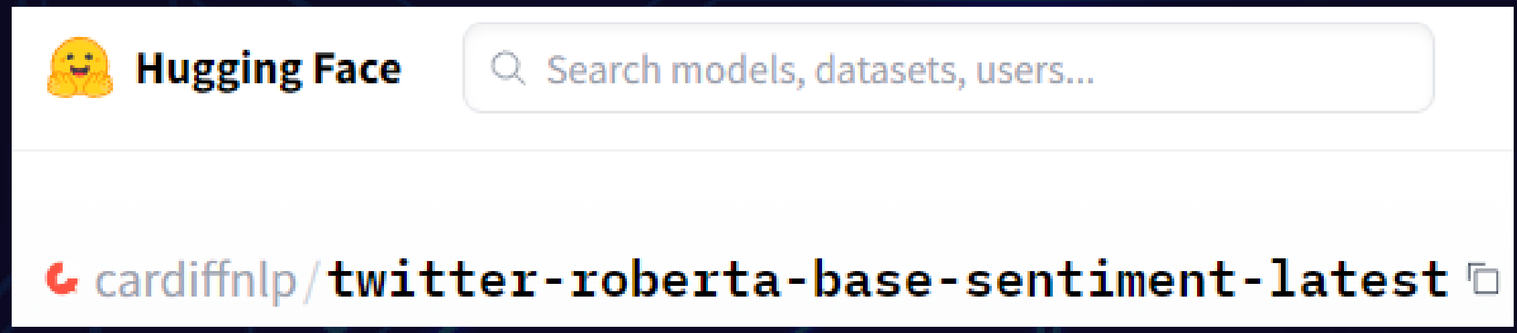
02

Make modifications to their script to be able to calculate for the confidence_score and add it to the dataframe

03

Test run it against dataset Sentiment Analysis Dataset found on Kaggle for a rough gauge of the results(Actual cases were not included as a result the results of the model will be skewed)

MODEL USED



Twitter-roBERTa-base for Sentiment Analysis

- Data trained on tweets
  - People are more free with emotions normally not as binary responses allowing for more neutral entries
- In comparison to TweetNLP, resulted in higher accuracy

OUTPUT

```
Accuracy: 71.00%
```

01 Accuracy only reached 71% potentially due to the model being slightly out of date or due to the dataset including characters that is being processed incorrectly.

02 CSV output was correctly created through the model used accompanied with the confidence score.

text	expected_sentiment	model_output	confidence_score
Last session of the day http://twitpic.com/67ezh	neutral	neutral	0.9
Shanghai is also really exciting (precisely -- skyscrapers galore). Good tweeps in China: (SH) (BJ).	positive	positive	0.99
Recession hit Veronique Branquinho, she has to quit her company, such a shame!	negative	negative	0.93
happy bday!	positive	positive	0.98
http://twitpic.com/4w75p - I like it!!	positive	positive	0.97
that`s great!! weee!! visitors!	positive	positive	0.99

POSSIBLE FUTURE IMPROVEMENTS

01 One possible improvement is by hand-creating a labelled dataset to specifically fit the kind of purpose the sentiment analysis will accomplish ex. Labelled dataset for reviews for online reviews

02 Better text processing to prevent any errors which may have occurred due to the way it was imported

03 A better way that I wanted to attempt was through the use of polarity instead between positive and negative. It currently outputs the likelihood of a tweet being positive, negative and neutral. Removing neutral as a probability and allocating a range in polarity would be better to distinguish the emotion as to which it leans towards and allow for a more detailed analysis as to the level of intensity of the sentiment

04 The creation of a model that could better understand more complex texts such as multiple sentences should also be looked into