Link to project: https://github.com/SanderGi/SentimentAnalysisAPI

Sentiment Analysis ML Report

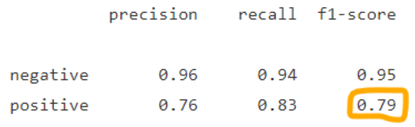
Models

- A Logistic Regression based classifier using the bag-of-words representation of texts as vectors. Simple data pre-processing employed (removing punctuation and contractions). Works well and is computationally and development-time-wise effecient. Useful as a baseline.
- A Naive Bayes based classifier with tokenization and slightly more advanced data preprocessing (regex to remove links, etc.). Still the same principle though. Useful as a sanity check because it can model feature importance so it can be verified that the words that indicate positive/negative sentiments make sense (e.g. “wonderful” and “fantastic” strongly indicate positive sentiment according to the model which makes sense).

EDA and Metrics

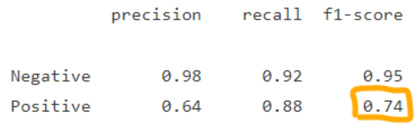
During data exploration, it was clear that the training data distribution was biased (more negative than positive observations/texts). So I opted for evaluating the models using precision and recall as summarized by the **f1-score** instead of possibly misleading accuracy scores.

- Logistic Regression had a score of 79%

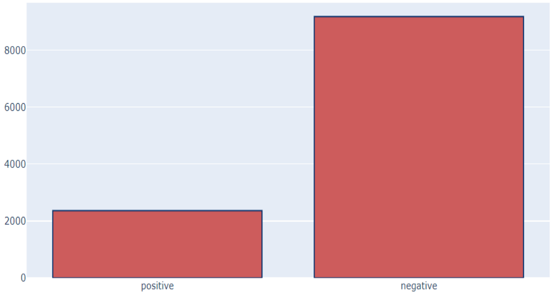


```
              precision    recall  f1-score

    negative       0.96      0.94      0.95
    positive       0.76      0.83      0.79
```

- Naive Bayes scored 74% on test data



```
              precision    recall  f1-score

    Negative       0.98      0.92      0.95
    Positive       0.64      0.88      0.74
```



Both fared comparably on regular accuracy (about 92% correct predictions) so I chose to use the Logistic Regression for the API.

Model Tuning

I varied two hyperparameters:

- Amount of training data (from 50 to 80%) using a fixed amount (20%) of test data
  - To understand the tradeoff between high variance and high bias
  - 80% yielded the best f1-scores on test data
- Type of data pre-processing
  - Did not seem to have a significant effect whether or not I removed punctuation, etc. (or used twitter specific tokenizers)

I just varied these parameters manually in a notebook since the models were so simple. If I was to spend more time on this project, I would want to look at tuning regularization and other parameters using a more automated cross validation setup.