Sentiment analysis exercise

Tanja Crijns s4204999

Brief description of problem approach

I chose to do the second assignment on the #trump hashtag tweet dataset. We were given a dataset of 110 twitter messages containing said hashtag. The goal was to automatically determine the attitude of the author towards Donald Trump. The twitter messages were not annotated and I annoted them by hand, there were three categories: positive, negative and neutral. During annotation, I made the following choices:

- Negative tweets about Hillary Clinton are not assumed to be negative for Donald Trump but neutral. Although it seems logical to us that this is negative for Donald Trump, such an assumption should not be made.
- Non-English tweets are considered neutral.

This resulted in 43 neutral, 30 positive and 37 negative tweets.
I decided to integrate the given positive and negeative word lists into a standard text mining procedure.

- **Pre-processing:**

  I removed any puncuation including # and @. Every word written in CamelCase was separated into two words, for example DonaldTrump would become Donald Trump. I lowered all the capital letters. I checked whether the words were present in either the positive word or negative word file. If this was the case, I added "_POSITIVE_" or "_NEGATIVE_" to the tweet. Eventually I did not use this addition of terms, because it seemed to lower the performance.

- **Classification:**

  I applied 10-fold cross validation and two different classifiers on this pre-processed dataset. I chose to use the cross validation technique because the dataset is small. This way, all data is used to train and test which can improve performance in small datasets. I tried two classifiers, Naïve Bayes as a baseline and Logistic Regression because Naïve Bayes has an independence assumption and this might not be the useful for these short tweets.

Result evaluation

Meaned results across 10 folds:

	Accuracy	Recall	Precision	F1
Naïve Bayes	0,4545	0,4545	0,5564	0,4553
Logistic Regression	0,5182	0,5182	0,5861	0,5129

Overall, Logistic Regression performed best. Without the previously mentioned pre-processing steps, the classifiers performed worse. With an exception of the addition of positive and negative terms. Chance level with three classes is 0,33 and we reach a higher performance than that, not very good however.

- What problems did you encounter? What problems could you solve with 10 hours extra time? What problems do you think are real challenges in sentiment

The biggest problem was the dataset. For a classifier, it is hard to determine the subject to what the tweet is adressed. Sentiment recognition is easier if only the postitive/negative/neutral classes are relevant. However, in this task the subject of sentiment was also important. Also, the multiple languages were a problem. Given more time, I would try to detect non-English tweets and remove them from the dataset. Also, even though it is not really relevant. In determining the positive and negative words, I noticed that 'trump' was considered a positive word. In a context outside of Donald Trump, a trump could for instance be an decicive and advantageous move. However, this could cause a bias in this particular problem. I removed 'trump' from the positive word list. I believe that the biggest problem in sentiment analysis is to determine the subject of the sentiment.