MSE_AnTeDe_Lab3 – Adrian Willi

Unsupervised approach

In the unsupervised approach is first a unique list of words created for each document in the corpus. In a next step, the number of positive respectively negative words is counted. With this numbers a simple score is calculated by subtracting the number of negative words from the number of positive words. If this score is bigger than 0 the document is labeled as positive and if smaller than 0 as negative. This approach reaches an accuracy of 0.709 .

Supervised approach

In the supervised approach different possibilities are evaluated. First, a multinomial naïve bayes classifier was trained and reached an accuracy of 0.802 on the training data.

But there is still space for improvements. For this reason, a binary value is used instead of the token counts. With this change a multinomial naïve bayes classifier is trained again. Now, we receive an accuracy of 0.833 on the training data. This shows that the number of occurrences of a token may not be as important as its presence or absence in sentiment analysis. Next, we evaluate the better approach on the test data and receive an accuracy of 0.825. It can be stated, that is approach is better than the unsupervised one.

In a next step, we train a maximum entropy classifier using logistic regression. The classifier returns an accuracy of 0.83 on the test data.

Now, we inspect which words really have an impact on the results as shown in the figure below. Interestingly, it can be seen that some words affect the decision that are supposed to have neutral semantic orientation. For example, work, plane, people, etc. It seems reasonable to get the classifier to focus on sentiment words. Therefore, a list of positive words and a list of negative words is created. These lists are then used to train the maximum entropy classifier again. Finally, it is evaluated on the test data and gives us an accuracy of 0.835. The improvement is rather small but in the range of what I expected, because the words are still the same as before. If we would delete the words which have an impact, we could see a big decrease in accuracy.

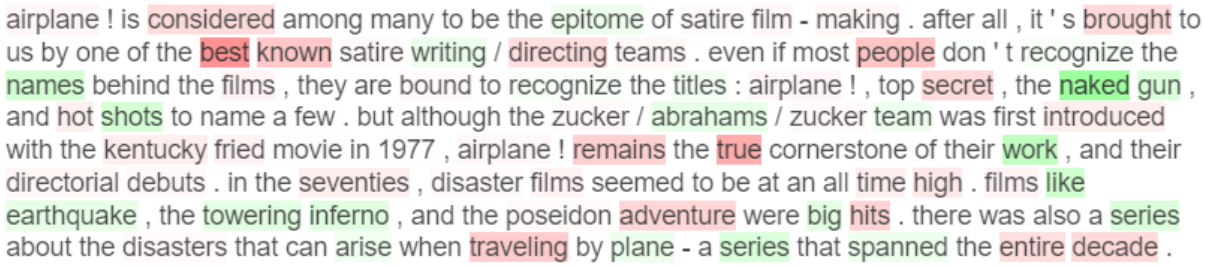

Figure 1 - Words colored by its influence