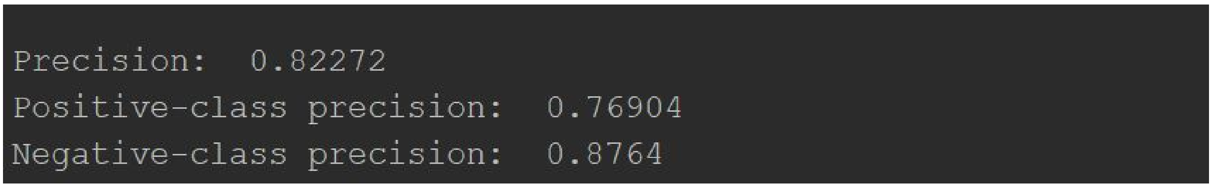
# Report Oblig 1

## Performance:

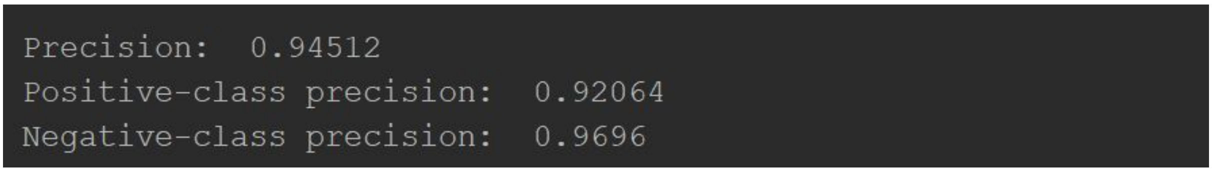
When we measure the performance for our classifier, we have to look at two different things. We have given an option to clean the data before it is classified. The cleaning add time to the classification, but the cleaned data gives a higher precision rate.

### Non-cleaned:



```
Precision:  0.82272
Positive-class precision:  0.76904
Negative-class precision:  0.8764
```

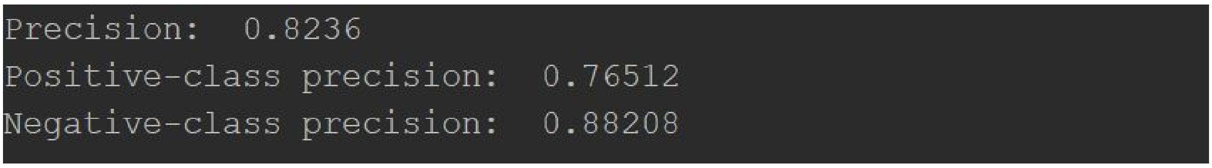
On the training set of non-cleaned data we get a precision of 0.82272 ~ 82.2%.



```
Precision:  0.94512
Positive-class precision:  0.92064
Negative-class precision:  0.9696
```

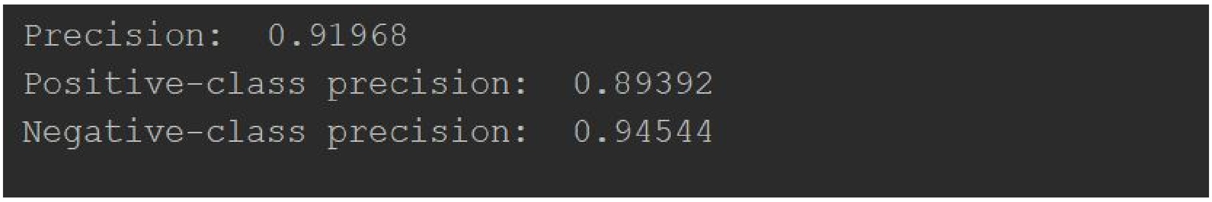
On the test set of the non-cleaned data we get a precision of 0.94512 ~ 94%

### Cleaned:



```
Precision:  0.8236
Positive-class precision:  0.76512
Negative-class precision:  0.88208
```

On the cleaned training set we get a precision of 0.8236 ~ 82.3%, which is 0.001 higher than the non-cleaned code.



```
Precision:  0.91968
Positive-class precision:  0.89392
Negative-class precision:  0.94544
```

On the cleaned test set we get a precision of 0.91968 ~ 91%, which is 0.03 points lower than the non-cleaned dataset.

## Error rates:

### Non-cleaned:

On the non-cleaned training set we got an error rate of 0.17728 ~ 17.7%. But on the non-cleaned test set we got an error rate of 0.05488000000000004 ~ 5%, which is a significant improvement of the training set error rate.

### Cleaned:

On the cleaned training set we got an error rate of  0.1764 ~ 17.6, which is only a 0.001 improvement from the non-cleaned training set. On the cleaned test set we got an error rate of 0.08031999999999995 ~ 8% which is an improvement compared to the training set, but is higher than the non-cleaned set.

## Explanation and how to improve:

In the code the cleaning variant is slower than the non-cleaned code. Because we search through the stopwords list in a sequential way. We could improve the run-time on the cleaned by changing the searching method of the stopwords.