Beginning:

Collecting the data from twitter:

After some research we found a file containing all twitter account names for members of the German Bundestag with their respective party. This file had some formatting errors which we had to clean up before using it.

For downloading the required data from twitter we decided to use the opensource framework selenium in combination with google chrome as web browser. This setup allowed us to download as many tweets as we wanted with all of the metadata visible on Twitter.

Because the twitter page is not loaded all at once but continuously as you scroll down, we had to imitate a real user scrolling down the page. For implementing this, we used selenium to scroll to the end of the current view and wait a certain amount of time for the page to load new content. To find the ideal time to wait we had to assess the trade-off between increasing the runtime and the risk of losing some tweets because they weren’t loaded in that short amount of time. We still noticed that we didn’t get all of the tweets in the first run so we decided to do another run in a different network and combine the results.

Here is an overview of the two runs which shows doing a second run increased the total number of tweets by 5%. Because this value is so low, we assume that we got nearly all the data.

223457

235939

For every politician we created an own CSV-file containing the username, post date, text of the tweet, the number of replies, retweets and likes. From these files we created several numpy arrays for training and testing, in which we deleted all retweets to guarantee that all texts were written by the politician.

In our first step we wanted to train a model to detect which political party the sender of the tweet belongs to. The model therefore has to process the tweets and output a likelihood value for each of the 7 parties we used.

To tune the hyperparameters for our model we decided to use the keras tuner. The result we got was 16500 neurons for the hidden layer and a learning rate of ?. We then used these hyperparameters to fully train a model, which reached an accuracy of about 55%. To visualize the difficulties the model had with classifying the tweets, we used it to predict the party for every tweet in our test dataset. Out of these results we created the following confusion matrix. It shows that the model has the highest accuracy for identifying tweets belonging to the AfD party. Tweets from the party “Grüne” however seem harder to classify as the model confuses a relatively large amount of tweets with the SPD and LINKE. It also shows that the model is biased against the CSU which we already expected since there are way less tweets from the CSU in our dataset than from the other parties.

We were also curious how our model would respond to tweets by members of the LKR party which it had never seen before. The result was …

In the next part of our project we wanted to use our dataset to analyse the big5 personality traits of the politicians and find out if there are any major differences. We were also interested in whether there is a noticable difference between the parties when taking the average over all of the members.

The first problem we had was that the tweets in our dataset are in german whereas all the datasets we could find for the big5 analysis were in english. We decided to use the english essays dataset for training our model and then translate all of the tweets into english for evaluating them. To find out how much information is lost during the translation process, we decided to translate the essays dataset to german and retranslate it to english and evaluate it with the original test dataset.

We therefore needed some sort of bi-directional translation between german and english. The first thing we thought of was the transformer model we discussed during the meetings. To use it for our purposes, we first had to find a suitable dataset containing translations between german and english sentences. TED\_English\_German.txt

Furthermore we had to create our own tokenizer for the german language since the tensorflow tutorial only provided a Portuguese and English tokenizer.