NLP
Problem Formulation

Lennart Hallenberger & Florian Siepe

May 30, 2022

1   Motivation

Over the past years social media has become more relevant in politics. Not only are politicians them self's using social media to gain a growing following, it's also where citizens discuss and share their opinion on political topics.

From this comes an interest to classify social media posts by their political point of view. One use case can be to gain insights on political topics and their distribution amongst voters. Furthermore analyzing posts in a large scale can play a roll in more accurate projections of future elections.

2   Idea and Goals

The goal of this project is to classify users based on their social media posting in regard of their political viewpoint to match them with a political party.

Because of a lack of pre-classified data and the potential of them getting outdated, our approach is to set up a reusable pipeline. The basic idea is to use social media posts of known politicians as training data. Each of these postings will be labeled with the according political party it's author belongs to. Based on the labels we try to extract relevant features which are closely correlated to the political parties.

To test the precision and recall of our approach we use tweets made by members of the republican and democratic party from the United States. Next we apply our model to tweets made by politicians who were not part of the training data.

The next step of this project could be to build a model based on tweets of US politicians before the 2020 election. This model could then be applied on a collection of tweets before the election, in order to predict the users vote for either party. The result can then be matched against the election results to compare the distribution of voters on social media.

3 Datasets

- US Politicians Twitter Dataset
  https://www.kaggle.com/datasets/mrmorj/us-politicians-twitter-dataset

- US Election 2020 Tweets
  https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

- Democrat Vs. Republican Tweets
  https://www.kaggle.com/datasets/kapastor/democratvsrepublicantweets

- Further research during the project

4 Expected Results

We expect to find at least some features of the used language that will be different between the members of opposing political parties. Whether these are significant enough or not to match them with posts from non politicians remains to be seen. To increase the chance of differing features, we can limit the used data to a single political topic with opposing opinions.