Data understanding

Repo: https://github.com/martentyrk/FakeNewsDetection
Group ID: D15

Gathering data

Because we wanted to make a fake news detector, we needed a dataset, that would have examples of both fake news and real news articles. The main things we were looking for in a dataset were the title of the article, the content and a separator whether it was real or false. The data needed to be csv format, so that it would be easier to manage.

Describing data

On our search for data we headed to Kaggle, where we also found what we were looking for. We found a dataset that had everything we needed and had separated the fake and real news into two separate .csv files. The size of the two files together is 110.98MB in which true news articles took 51.1MB. So the distribution of data is to our liking. In order to have a better accuracy, we decided to look for more and found another dataset also from Kaggle with the size of 10.46MB. It also includes everything we were looking for. The first dataset has the following columns: title, content, subject, date. The second dataset has the following columns: author, date, title, content, language, site_url, real_or_fake, bias, title_without_stopwords

Exploring data

When having a closer look at the bigger dataset I can already see that the subject column only has two different values: political news and worldnews. This makes the column pretty much useless because we can't possible know what the worldnews in itself contains. The fake.csv has the subject column with classifications news, politics and other. Because the classifications are so different then that makes it hard to use it. The second dataset with the size of 10.46MB has

more so called fake news articles with the percentage of 62% from the whole dataset and 38% of the whole dataset contains real news articles. In addition, 2% of the content column is with null value. I'm suspecting this dataset also includes twitter posts and the 2% are these. We do consider twitter posts as news but regarding our other data we plan to use, then we can't really use them.

Data quality

As an overall, I'd say we have what we need. The quality is mainly low only with the columns that we don't plan on using. The distribution of fake and real news articles with the second dataset is 62/38, but thanks to the other dataset that is much bigger, we believe that it won't play that big of a role.