Assignment 1

Hanna Kienast, 12020688

I want to contribute to the field of **NLP** and **Fake News Detection**. I found that there are already many well working models out there. I also found a lot of data, however, most of it is from English speaking news outlets. My idea is to **create a dataset** for Fake News Detection for German speaking news networks and train and test well established models with this dataset.

My papers include:

- A paper on web scraping: https://dl.acm.org/doi/pdf/10.1145/3453892.3461333
- Papers on fake news detection using deep learning: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9108841, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9620068, https://paperswithcode.com/paper/exploring-summarization-to-enhance-headline

Time plan	Task
1-2h coding or debugging every day until around 15th November (dataset should be done then)	To create a well sized dataset, I propose the idea to use a web crawler that extracts data from chosen news sites. I have created a list of 30 German speaking news outlets from every political direction from well-known platforms to alternative media from professional journalism to boulevard media. This data will later be stored in a .json and .csv file. I aim for the dataset to contain 10 000 data entries.
1-2h/day for 1 week	Then, for the task of labelling the data, I will manually label around 100-500 data entries. The labels will include: fake news, satire, extreme bias, conspiracy, junk science, hate, clickbait, credible.
1-2h/day for 1 week	After that, I will use a well-established model to label the rest of the data. Then I will use a well-established CNN and BERT model to train and test it on my data (possibly from https://github.com/AIRLegend/fakenews).
17.12.2024	**Deadline Assignment 2**
10h	Build application & finalize
5h	Create video
5h	Write report
21.01.2025	**Deadline Assignment 3**

I will later make the web crawl approach and the dataset public. If I fail to collect a big enough dataset, I will at least try to enlarge existing datasets and use these on the models.

The table includes my aims but, in the end, if I realize some things make more sense if I change them (labels, data size, manual data labelling) I will adjust the plan. Also, the time plan is an estimation, as I do not yet know how much debugging I will have to do.