Portolio I: Sentiment Analysis

Your task is to carry out a sentiment analysis research project.

1 Datasets

You can choose among the following provided datasets:

- **Yelp:**
  - `yelp_reviews_kiel_de.csv`: more than 3000 Yelp reviews and 5 star rating of locations in Kiel (German reviews only)
  - `yelp_reviews_hamburg_en.csv`: more than 3000 Yelp reviews and 5 star rating of restaurants in Hamburg (English reviews only)
  - Some additional metadata on the locations is provided in `yelp_businesses.json`
- **Tagesschau:**
  - `tagesschau_de.csv`: More than 4000 articles from Tagesschau Meldungsarchiv since January 2022 (German articles only). Note that this dataset is not labeled for sentiment analysis, so it cannot be used for supervised approaches to sentiment analysis.
- **Other datasets:** It is also possible to choose some other data set, but please ask me first.

2 Research focus

You can put the focus of your analysis on whatever aspect is interesting for you. Here are some suggestions:

- **German language**: What are best practices for NLP and sentiment analysis for German texts? Which python packages or other resources are available? Are there linguistic aspects of the German language which need to be considered?
- **Model evaluation**: Carry out several different approaches and evaluate their performance: (1) rule-based vs. machine-learning based, (2) Bag of Words vs. Word Embeddings, (3) different preprocessing strategies, (4) different machine learning algorithms, . . . You should show quantitative evaluations, but you can also show anecdotal evidence of text examples where the algorithms fail badly. What are situations in which rule-based or ML-based algorithms typically fail to predict correct results?
- **Linguistic aspects**: How can we associate sentiments with more specific parts of the texts to get more nuanced results? How can we combine sentiment analysis with Part of Speech tagging, named entity recognition, or other linguistic approaches? Can we extract structured information about sentiments towards the food, service, location, weather (YELP review) or persons, countries, topics, . . . (Tagesschau articles)?
- **Data acquisition**: Although data is already provided, you are free to put your focus on the acquisition of data. How can we use the YELP API or web scraping to access relevant information?

3 Formalities

- **Documentation**: Use Jupyter Notebooks to document your work. Not only show code and output, but guide in detail through your project: What are your goals? Which steps do you carry out and why? What are your findings and insights? What are possible limitations?
- **Submissions**: Submit all that is needed to fully reproduce your work on Moodle, or submit a link to a public GitHub repository.
- **Due date**: 2022-05-31