# Capstone Project

## Task Description

- **Your task**: Carry out a small but complete social media analytics project related to topics covered in the course (APIs, web scraping, network analysis, text mining, sentiment analysis, text or token classification, LLM applications, retrieval augmented generation, topic modeling). Describe a real or fictitious scenario for your project, explaining its goal and purpose.
- **Data**: Your project must be based on a real-world data set. You can either use an existing data set or engineer your own data set via APIs, web scraping, and possibly annotating the data. The data context can be social media platforms or other forms of text data such as news articles, product reviews, etc. Hint: If you intend to carry out a text classification task, keep in mind that you need a data set that comes with labels (which is often not the case with text data), or spend some time on annotating data yourself.
- **Focus**: If your project has a significant data engineering part (data acquisition, cleaning, annotation, etc.), then a compact analytical part is sufficient. If your project is based on an existing data set, then I expect a larger scope and/or depth on the analytical side (in-depth evaluation of multiple models and parameter choices).
- **Documentation**: Include all relevant steps of your project into a Jupyter notebook and guide the reader through the project in words. If your project has both a data engineering and an analytical part, split your submission into two different Notebooks. In particular, make sure to carefully evaluate the results and comment on the insights you gained.
- **Language**: Allowed languages are English and German.
- **Resources**: You may use all the code from the lectures. Copying and adapting from other sources is allowed in small quantities. Copying code in large quantities will be treated as intent to deceive and result in a score of zero points. Cite all relevant resources on which your project is based or from which you draw inspiration.
- **Chat GPT usage**: You are encouraged to use AI assistance to learn about concepts and approaches, write better code or similar tasks. However, the main part of the project must be your own work, and you need to understand and be able to explain in person what you are doing in the project. If you use AI assistance, you must clearly state this in your submission.
- **Jupyter Notebook vs Script**: While it can be helpful to shift the definition of functions or classes into separate scripts, the main part of your project should be presented step-by-step in a Jupyter Notebook. It is essential that the reader can follow your thought process and understand your work. So please avoid hiding the entire logic and complex processes in a few functions such as `clean_data()` and `train_model()` and then just calling these functions in your notebook.
- **Submissions**: Submit all that is needed to fully reproduce your work (notebook, scripts, data, etc.) on Moodle, or submit a link to a public GitHub repository. If you are using some API, you should not provide API keys, but rather a script that can be used to reproduce the data acquisition process. The raw data used in the project must be included in the submission.

## Continuous Work and Feedback

- In the beginning of the semester, it can be difficult to make a decision on the project topic, because you may not yet have a complete understanding of the topics covered towards the end of the semester. However, you should start thinking about potential project ideas early on. And you should continue working on the project throughout the semester.
- There will be times in our lectures explicitly reserved for work on your project. In particular, you will be asked early in the semester - as part of your regular problemsets - to investigate data access options and collect project ideas. Towards the end of the semester the focus will shift even more

towards the project. You will be asked to write a short project proposal, present it to lecturer and collect feedback from him.

## Grading

Due to the different nature of the projects, the grading will be based on a holistic evaluation of the project. The following aspects will be considered:

- Creativity, complexity and innovativeness of the project
- Correctness of the approach
- Data engineering efforts
- Thorough evaluation and correct interpretation of the results
- Convincing Storytelling: purpose of the project, insights, reflection on learnings and limitations
- Well-structured, concise and clean submission
- "ChatGPT buzzword bingo" will be considered as a malus. Write in your own words!