

# Understanding and Training Language Models: Project

Erik-Jan Senn

Faculty of Mathematics and Statistics, University of St. Gallen

---

CSH Autumn School at University of Hohenheim  
September/October 2024

# Project in Sentiment Analysis

Implement a small project;

- ▶ **Data:** text (and optionally other numerical features), and a target variable to predict (pref. a discrete variable such as an emotion, up/down, positive/negative).  
Need sufficiently many observations and ( $>1000$ , depending on application).
- ▶ **Build the entire modeling pipeline:** Text processing, transformer model / LM, application specific prediction model and model evaluation.
- ▶ **Train the well-performing model:** Estimate and improve the classifier, fine-tune the language model, test different language models, add different features, ... Compare your performance to a naive benchmark.

Hints:

- ▶ Discuss your idea and data with the lecturer and other participants before you implement it.
- ▶ Start simple: For example, you can start with a simple model and pipeline implemented using the transformers library.
- ▶ You can use any code from the class or online resources. AI assistants help!
- ▶ Example data: Data folder on course gitlab, or e.g. [Papers with Code](#), [Kaggle](#).

**Questions ?**