- Topic: Task Label Aware Classification for Hate Speech and Harassment
- Supervisor: Amir Hossein Kargaran
- Examiner: Prof. Hinrich Schütze
- Open for: **MSc/BSc**

1 Description

Goal: This project aims to train a universal transformer architecture to recognize all forms of hate speech, and harassment.
Why hate speech and harassment: Many studies in this area separate every form of harassment and hate speech into a different category and attempt to detect them. For example sexism and misogyny are a form of harassment but they differ from just using profanity words or racism. In order to restrain the harm caused on digital social platforms, we need to understand each category's potential better and detect them.
Solution: Any text classification problem aims to find a function f:

$$f: \text{ text } \to \{0,1\}^M \tag{1}$$

that maps text to an M-dimensional vector where each dimension corresponds to a certain label. Most of classification models learn the function f for each category of hate speech and harassment separately or in a multi-label classification setup, making it hard to **reuse the existing model for a new task** or to **add more labels to it**. Factoring the text classification problem into a generic binary classification task lets us train a model that takes hate speech and harassment category as inputs besides the text.

$$f :< \text{task label, text} >\to \{0,1\} \tag{2}$$

For example:

- f('sexism', 'Men and women's brains are wired different bro, that's just how it is.') → 1
- f('racism', 'Men and women's brains are wired different bro, that's just how it is.') → 0

This allows us to use our existing model while extending it for the new labels.

2 Prerequisites

- Be interested in the topic
- Proficiency in speaking and writing English
- Python Coding, knowing how to use numpy and pandas libraries.
- Knowledge of classifiers.
- Knowledge of transformer architecture is recommended.
- Knowledge of using PyTorch and Transformers libraries is recommended.

3 Supervisor

Hello, I am Amir, and if you pick this project, I will be your supervisor. You will get 45 minutes per week of guidance and help, which can be accumulated up to 3 weeks. To begin our work together, I describe the project completely, resources and and other ideas that i have. Your works begin with a literature review of the task and the resources available. I will make sure you get the time I mentioned from my side. If I see potential in your work, I can definitely help you even more. If you have any questions contact me directly: amir@cis.lmu.de