Investigate temporal shift in hate speech detection

- **Supervisor**: Antonis Maronikolakis
- **Examiner**: Prof. Schütze
- **BSc, MSc, Open**: BSc, MSc
- **Summary**: Hate speech online is a task that has challenged researchers and engineers developing systems to foster healthy online platforms. To tackle this challenge, language models are often used as classifiers to detect and filter hateful content. An issue plaguing language models in general is that of a temporal shift between training and real-life data. With the evolution of language, models trained on old data can become obsolete if the divide between training and contemporary data is too large. This is especially salient in hate speech detection, where the vocabulary and terminology used by cyberbullies and peddlers of hate speech constantly shifts and changes, not only through online language evolution, but also to avoid detection. Models trained on older datasets may lose their efficiency on contemporary data. Thus, developing models and methods to mitigate this issue is of paramount important to the health of online spheres. The goal of this project is to investigate this phenomenon and propose potential solutions to mitigate it. You will analyze temporal data and design a model training pipeline.
- **Prerequisites**: Data processing, Python, Machine learning (basic)