Title: Temporal Shift in hate speech data

Supervisor: Antonis Maronikolakis

Examiner: Professor Hinrich Schütze

BSc, MSc, Open: BSc

General Topic Area: Social NLP, Hate Speech

Prerequisites: Moderate Python skills

Details: To evade detection, users who post extreme speech content have to get crafty: they add misspellings, they appropriate neutral vocabulary and use code to hide hateful rhetoric from machine learning models. This hate speech "language" changes across time, with new evasion methods and particulars developing in online communities. Thus, training a hate speech classifier on data from a particular timeframe does not guarantee the model will work in later timeframes. In this project, we will explore the effect of this phenomenon, analyzing model performance across time, as well as potentially developing methods to mitigate this.