GÖ/HI-Workshop Feb. 2019

Johannes Schäfer

Neural networks at hate speech and offensive language detection with a focus on linguistic features

The detection of hate speech and offensive language in social media microposts is a complex problem where even humans struggle to find definitions of objective and clear boundaries. This talk focuses on methodological aspects and evaluates different models on German *Twitter* data which frequently contains political messages. The methods follow an approach aiming for an empirical solution using neural networks as these can learn highly inter-dependent features for complex tasks on given data. Word embeddings are used to model text segments where linguistic information is integrated to improve meaning representations. I present the current state of my work, discussing the performance of different neural network architectures (LSTM and CNN) with extensions using various types of linguistic features (part-of-speech tags, morphological: compounding). Future plans are outlined where the focus will lie on more fine-grained linguistic features such as the morphological analysis of words considering affixes (flexion and derivation) as well as possibilities for integrating metadata features on top of word embeddings directly.