ABSTRACT

In recent years the field of Natural Language Processing has seen dramatic changes thanks to the advances of modern transfer learning. This master's thesis aims to fill a research gap by addressing the question of whether introducing a lemmatization pre-processing step affects the accuracy of these state-of-the-art classification models. More specifically the question will be answered in the context of news articles in both English and German. The experiments conducted in this thesis aim to compare the accuracy of models using lemmatization versus models, which do not include a lemmatization step. In a first step, observations were derived from conventional machine learning models. The core question was then answered using a particular transfer learning method, namely: Universal Language Model Fine-tuning also known as ULMFiT. The findings of this thesis have been consistent: in none of the experiments conducted did lemmatization have a significant effect on classification accuracy. As a result the recommendation is not to include lemmatization as a pre-processing step to solve classification problems in the context of English or German news articles.