Abstract

Gender identification is a part of author profiling task in which we try to detect the gender of authors from their written texts automatically. Using classification algorithms, exploiting lexical features such as word and character n-grams yields good results on in-genre experiments (e.g. Twitter). In this thesis, by replicating a state of the art lexical approach, we applied the same model on different genres and datasets. Our results show that the lexical model is not a robust approach in cross-genre settings. For this reason, we tried classification with pre-trained word and sentence embeddings. Our experiments show that a simple word vector averaging technique to represent documents outperforms the lexical model in cross-genre settings. Also, we tried multi-task learning models to check if learning two tasks at the same time could help gender profiling results. Depending on the size of datasets, word embeddings coverage and the results of separate single-task models, in some cases, we got better scores. Finally, we talk about our approach to dividing one of the datasets to smaller pieces to have more instances for classification. We show that this strategy works better on some cross-genre models, but in some cases, it can hurt the results.