Evaluating Text Classification Models on Multilingual Documents

Julia Eigenmann

Master Thesis in Computer Science

Machine learning models often require large annotated datasets for training in order to obtain accurate results. However, the scarcity of the labeled data is a bottle neck for many applications, including text classification. The problem becomes eve n more challenging in the case of multilingual textual documents. In such a case, annotators are required to be experts in annotating the data in different languages. Existing methods have limited performance on classifying textual documents by using only a small set of labeled data.
In this thesis, we propose solving this problem by using heuristic rules to label a large set of multilingual documents and apply different classification models to them. We compare language-dependent with language-independent classification approaches and report the results of our comparison. Our results show that:

- Language-independent classifiers perform overall better than the language-dependent ones for underrepresented languages; this is probably due to their too small training dataset. Language-dependent classifiers with large training dataset might outperform the language-independent classifiers with training dataset of comparable size.

- Linear SVC, Random Forest, FastText, Logistic Regression and Distil Bert are well-performing classification models, whereas Multinomial Naive Bayes achieves only satisfying performance results. DistilBert performs best.

Prof. Philippe Cudré-Mauroux