Abstract

Keyword extraction is a key step in text classification to reduce the feature space complexity by selecting the most relevant terms in the document to train the classification algorithm. Arabic language is an inflectional and a derivational language. The language has a complex morphology and a rich semantic. The preprocessing of Arabic text can suffer from over-stemming and under-stemming problems. In this research, we investigated 12 models in Arabic text classification with Light stemming, Root-based stemming, Morphological Roots Extractor and a proposed Hybrid of the Light stemming and the Roots Extractor. Term Frequency Inverse Document Frequency (TFIDF) was used as the term weighting schema. The 12 models were compared as two sets of 9 models in the first set (Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD), k-Nearest Neighbors (kNN) and Rocchio) and 4 models in the second (Random Forest (RF), Bagging, Gradient Boosting (GBoosting) and Adaptive Boosting (AdaBoosting)). Models were compared in each set in terms of accuracy and F1-measure. Results showed that Light stemming is better in Arabic text preprocessing, SGD model is better when compared with other models in the first set and GBoosting ensemble model is better when compared with other models in the second set.