Jordan University of Science and Technology

The Impact of Indexing Approaches on Arabic Text Classification

Authors: Amer Al-Badarneh, Emad Al-Shawakfa, Basel Bani-Ismail, Khaleel Al-Rabab'ah, & Safwan Shatnawi

Abstract: This paper investigates the impact of using different indexing approaches (full-word, stem, and root) when classifying Arabic text. In this study, the na?ve Bayes classifier is used to construct the multinomial classification models and is evaluated using stratified k-fold cross-validation (k ranges from 2 to 10). It is also uses a corpus that consists of 1000 normalized Arabic documents. The results of one experiment in this study show that significant accuracy improvements have occurred when the full-word form is used in most k-folds. Further experiments show that the classifier has achieved the highest accuracy in the 8-fold by using 7/8?1/8 train/test ratio, despite the indexing approach being used. The overall results of this study show that the classifier has achieved the maximum micro-average accuracy 99.36%, either by using the full-word form or the stem form. This proves that the stem is a better choice to use when classifying Arabic text, because it makes the corpus dataset smaller and this will enhance both the processing time and storage utilization, and achieve the highest level of accuracy.