Jordan University of Science and Technology

Cross-lingual Short-Text Document Classification for Facebook Comments

Authors: Mosab Faqeeh, Nawaf Abdulla, Mahmoud Al-Ayyoub, Yaser Jararweh and Muhannad Quwaider

Abstract: Document Classification (DC) is one of the fundamental problems in text mining. Plenty of works exist on DC with interesting approaches and excellent results, however, most of them focus on a long-text documents written in a single language with English being the most studied language. This work is concerned with the natural step beyond such works which is cross-lingual DC for short-text documents. Specifically, we consider two languages, Arabic and English, and compare the performance of some of the most popular document classifiers on two datasets of short Facebook comments. Apart from limited attempts, the addressed problem has not been studied well enough. The results are encouraging and new insights are obtained.