World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:14, No:11, 2020

Use of Interpretable Evolved Search Query Classifiers for Sinhala Documents

Authors : Prasanna Haddela

Abstract : Document analysis is a well matured yet still active research field, partly as a result of the intricate nature of building computational tools but also due to the inherent problems arising from the variety and complexity of human languages. Breaking down language barriers is vital in enabling access to a number of recent technologies. This paper investigates the application of document classification methods to new Sinhalese datasets. This language is geographically isolated and rich with many of its own unique features. We will examine the interpretability of the classification models with a particular focus on the use of evolved Lucene search queries generated using a Genetic Algorithm (GA) as a method of document classification. We will compare the accuracy and interpretability of these search queries with other popular classifiers. The results are promising and are roughly in line with previous work on English language datasets.

Keywords : evolved search queries, Sinhala document classification, Lucene Sinhala analyzer, interpretable text classification, genetic algorithm
Conference Title : ICDAR 2020 : International Conference on Document Analysis and Recognition
Conference Location : Jerusalem, Israel
Conference Dates : November 26-27, 2020