Topic Selection for Malay Articles

Leow Jia Ren, Rayner Alfred
Faculty of Computing and Informatics
Universiti Malaysia Sabah,
Jalan UMS, 88999, Kota Kinabalu, Sabah, Malaysia
jiaren23@gmail.com, ralfred@ums.edu.my,

Abstract. Malay language is the major language that is in used by citizen of Malaysia, Singapore and Brunei. As the language is widely used, there are abundant of text or articles in Malay language are available on the internet. This result in the increasing of the articles in Malay language and the number of articles has increased greatly over the years. Thus, the studies for topic selection for Malay articles are very important in order to help clustering the articles into their respective class. In this paper, the approach used was the k-Nearest Neighbors (k-NN) classifier and Naïve Bayes classifier. Both classifier was used to classify and assign a topic to the documents according to a predefined topic sets. The approach will be tested by comparing the effects of using different distance method which is the Cosine Similarity and the Euclidean distance on the k-NN classifier. Other than that, the effect of stemming on the classifier and the different values of k used for the k-NN classifier were also tested. In conclusion, the proposed approach had shown that the k-NN classifier performs better than Naïve Bayes classifier in performing topic selection for Malay articles. Other than that, the stemming also improves the overall performances of both the classifier in the proposed approach. The findings also show that the application of Cosine Similarity as the distance measure improve the performance of the k-NN classifier too.

Keywords: Topic Selection, Feature Extraction, Text Preprocessing, k-Nearest Neighbours, Naïve Bayes, Classification.