ISSN: 1391-8796
Proceedings of
5th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
February 15, 2018

Sinhala news analysis using text mining and machine learning

Ekanayaka R.K.S.K.[1], Lorensuhewa S.A.S. [2] and Kalyani M.A.L.[3]

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka

Due to the rapid development of information technology, vast amounts of information are generated daily. Unstructured data such as news reports are a significant part of these growing information repositories. This study focuses on analyzing Sinhala news reports published online to extract important features using text mining and machine learning techniques. Then, represent this extracted information in a way that news readers find it easy to read news or do research on past news reports. For a morphologically rich complex language like Sinhala, it makes text mining a difficult task.

In our approach, we first pre-processed dataset with filtering, stop word removal, stemming and then experimented with feature selection methods such as n-gram combinations, count vectorizer and TF-IDF vectorizer. Text classification methods such as Naive Bayes, Support Vector Machines, Decision Trees, K-means and hierarchical clustering methods were evaluated. Later, we represented the mined knowledge using information visualization methods such as charts, tag clouds and tree structures.

Unigram features with TF-IDF vectorizer for feature selection, Naïve Bayes for document classification and K-means for clustering were the most accurate techniques for Sinhala news. The accuracy of the information visualization methods was measured with human experts. Our results reveal that language specific text pre-processing and feature selection increases the efficiency of information retrieval tasks when compared to generally used existing methods and the new representation model saves users' time and effort to find news reports based on their preferences rather than going through existing news websites.

Keywords: Sinhala language, feature selection, text classification, text clustering, information visualization

**Corresponding author: rkskekanayaka@gmail.com*