ISSN: 1391-8796
Proceedings of 9th Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
January 19, 2022

Word embedding-based sinhala news documents classification

Weerasiri R.I.*, Lorensuhewa S.A.S., Kalyani M.A.L.
Department of Computer Science, Faculty of Science, University of Ruhuna.

News articles are increasing daily, and a huge number of text documents are added to the Internet. Manual classification of these documents has become an impossible task. In Sinhala news document classification, TF-IDF has been used more often as a word representation, but word embedding has rarely been used. We compared the performance of Word2Vec, FastText and Doc2vec with frequently used Term Frequency Inverse Document Frequency (TF-IDF) as word representations for Sinhala news documents classification and applied machine learning approaches for the best word embedding model identified. We also experimented with each representation by removing stop words and investigated the feasibility of using Convolutional Neural Networks (CNN) as well.

Sinhala being a low resource language, it is challenging to prepare a corpus that is comparable with other resource-rich languages. We collected Sinhala news documents from different news websites and used them to create word embedding models, training and testing data. Models were created using *skip* gram method and assigned a 300-dimension vector for each word in the document. For CNN, embedding of each word in the first 100 words was taken for each document. In Word2Vec and FastText embedding models, we used a 300-dimension vector for each document by considering all the words in the document. Our experimental results show that FastText performs better than other representation methods and SVM with the *rbf* kernel always gives the highest accuracy. According to our experiments, the CNN model also showed prominent results but could not surpass the SVM model.

Keywords: Classification, Word embedding, FastText; Sinhala documents

*Corresponding author: rochanaweerasiri@gmail.com