Word Embedding as Word Representations for Clustering Sinhala News Documents

R.I. Weerasiri*, S.A.S. Lorensuhewa and M.A.L. Kalyani

Department of Computer Science, University of Ruhuna
**Corresponding Author E-mail: rochanaweerasiri@gmail.com, TP: +94714746543*

News articles are increasing by the day and the manual clustering or classification has become an impossible task. So there has been a need for new methods for clustering these articles. There is a huge number of text documents created and added to many sources including the internet daily. Manually clustering or classifying these documents into related fields has become an impossible task. Therefore, finding similarities in these documents has turned out to be a very inclusive topic. It helps save time by specifically searching articles. We evaluated the applicability of word embedding mechanisms like fastText to find its applicability to increase the accuracies in the classification process. We explored the feasibility of word embedding models like fastText, doc2vec as a word representation methodology compared to frequent methods like Term Frequency–Inverse Document Frequency in these documents and evaluate its accuracies. The research is based on evaluating the performance of different word representations for clustering and classification of Sinhala news documents. Initially about 10,000 Sinhala news documents were collected by a scraping algorithm from different news websites. They were cleaned, preprocessed to remove irrelevant characters and words. The models were checked for accuracy with changing the number of documents with each model. This model is used for representing words in the model and checked for higher accuracies with various representation mechanisms for both clustering and classification where models like kmeans used for clustering and k nearest neighbours and support vector machines for classification. We have tested the accuracies of various word representations like Term Frequency–Inverse Document Frequency, doc2vec and fastText and upon research and experimenting we have found that fastText models as word representations give best results for both clustering and classification. Therefore, using fastText word embedding models to represent documents for classification and clustering purposes will increase the accuracy.

Keywords: Clustering; Classification; Word embedding; FastText; Sinhala documents