Abstract No: 110 ***ICT, Mathematics, and Statistics***

COMBINING FASTTEXT EMBEDDINGS WITH NEURAL NETWORKS FOR SHORT TEXT CLASSIFICATION

J.R.K.C. Jayakody* and V.G.T.N. Vidanagama

Department of Computing and Information Systems, Faculty of Applied Sciences, Wayamba University of Sri Lanka
**kithsirij@wyb.ac.lk*

Using embedding representation is a critical step to improve the classification accuracy of a text dataset. Even though Bag of Word (BOW) models are used with past research work, usage of word2vec, Glove and FastText as embedding techniques helps to represent the features of text documents in a distributed manner, hence improving the accuracy of such models. The latest research work used a combination of embedding techniques and enhanced neural network models to improve the classification accuracy of text documents. FastText as an embedding unsupervised model and CNN, LSTM, and RNN as neural models were used extensively in the latest research work. However, comprehensive analysis with FastText and neural models with text documents has not been undertaken thus far. As a result, it is hard to compare the existing research work, and it is unclear which combination of neural model with FastText performs well over the other techniques. Therefore, it is necessary to investigate the impact of neural networks when the features were represented with the FastText embedding model. A famous movie review dataset was used for the experiment. CNN, LSTM, RNN, NN, and variations of those neural networks were used as neural networks. Hold out stratified Training and testing set was taken with 70 % to 30% split. Seventy per cent of training data was split as 80% of training and 20% of validation set. We compare classification accuracy across a range of neural network models, and our results show that the RNN model outperforms other neural network models with FastText embeddings with 86% accuracy. Moreover, out of various neural networks, the combination CNN-LSTM outperforms all other neural network models with 88% accuracy. The outcomes of this study can be a baseline for future research.

Keywords: Classification, CNN, FastText, LSTM, RNN