Paper No: SC-08 **Smart Computing**

Impact of Feature Selection Towards Short Text Classification

J.R.K.C. Jayakody[1*], V.G.T.N. Vidanagama[2], Indika Perera[3], H.M.L.K. Herath[4]

[1] *Department of Computing and Information System, University of Wayamba, Sri Lanka, kithsirij@wyb.ac.lk*

[2] *Department of Computing and Information System, University of Wayamba, Sri Lanka, tharinda@wyb.ac.lk*

[3] *Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka, indika@cse.mrt.ac.lk*

[4] *Department of Agribusiness Management, University of Wayamba, Sri Lanka, keminda@wyb.ac.lk*

Feature selection technique is used in text classification pipeline to reduce the number of redundant or irrelevant features. Moreover, feature selection algorithms help to decrease the overfitting, reduce training time, and improve the accuracy of the build models. Similarly, feature reduction techniques based on frequencies support eliminating unwanted features. Most of the existing work related to feature selection was based on general text and the behavior of feature selection was not evaluated properly with short text type dataset. Therefore, this research was conducted to investigate how performance varied with selected features from feature selection algorithms with short text type datasets. Three publicly available datasets were selected for the experiment. Chi square, info gain and f measure were examined as those algorithms were identified as the best algorithms to select features for text classification. Moreover, we examined the impact of those algorithms when selecting different types of features such as 1-gram and 2-gram. Finally, we look at the impact of frequency-based feature reduction techniques with the selected dataset. Our results showed that info gain algorithm outperform other two algorithms. Moreover, selection of best 20% feature set with info gain algorithm provide the same performance level as with the entire feature set. Further we observed the higher number of dimensions was due to bigrams and the impact of n grams towards feature selection algorithms. Moreover, it is worth noting that removing the features which occur twice in a document would be ideal before moving to apply feature selection techniques with different algorithms.

Keywords: *classification, feature selection, n grams, document frequency*