The Impact of Pre-processing and Feature Selection on Text Classification.

Nur Syafiqah Mohd Nafis
Soft Computing and Artificial Intelligence Research Group (SPINT)
Faculty of Computer Systems & Software Engineering.
University Malaysia Pahang, UMP
Lebuhraya Tun Razak 26300, Kuantan, Pahang, Malaysia
nsyafiqahmnafis@gmail.com

Suryanti Awang
Soft Computing and Artificial Intelligence Research Group (SPINT)
Faculty of Computer Systems & Software Engineering.
University Malaysia Pahang, UMP
Lebuhraya Tun Razak 26300, Kuantan, Pahang, Malaysia
suryanti@ump.edu.my

Abstract:
Nowadays text classification dealing with unstructured and high-dimensionality text document. These textual data can be easily retrieved from social media platform. However, those textual data are hard to managed and processed for classification purposes. Pre-processing activities and feature selection are two methods to process the text document. Therefore, this paper is presented to evaluate the effect of pre-processing and feature selection on the text classification performance. A tweet dataset is utilized and pre-processed using several combinations of pre-processing activities (tokenization, removing stopwords and stemming). Later, two feature selection techniques (Bag-of-Words and Term FrequencyInverse Document Frequency) are applied on the pre-processed text. Finally, Support Vector Machine classifier are used to test the classification performances. The experimental results reveal that the combination of pre-processing technique and TF-IDF approach achieved greater classification performances compared to BoW approach. Better classification performances hit when the number of features is decreased. However, it is depending on the number of features obtained from the preprocessing activities and feature selection technique chose.

Keywords: Unstructured; High-dimensional; Pre-processing; Text Classification; Feature Selection

ACKNOWLEDGMENT

We would like to show our gratitude to Universiti Malaysia Pahang grant (RDU vote number RDU190174) for supporting this research.