

ABSTRACT

Stemming words to remove suffixes has applications in text search, machine translation, document summarization, and also text classification. In the the Indonesian language, *stemming* is of crucial importance: words have prefixes, suffixes, infixes, and confixes that make matching related work difficult.

In this final paper, the three algorithms of *stemming* Indonesian such as Nazief&Adriani's algorithm, Arifin&Setiono's algorithm and Vega's algorithm are implemented and the performance of these algorithms is compared each other. Parameter of the *stemming*'s performance is Terms that have been stemmed correctly are divided with the whole terms. The results show that, the algorithm of Nazief&Adriani with the improvements correctly stems around 96.5447%. We conclude that *stemming* for Indonesian should be performed using Nazief&Adriani with the improvements approach.

In this final paper, the influence of *stemming* to text categorization will be analyzed. In this case, *stemming* is performed with the three algorithms of *stemming* Indonesian (Nazief&Adriani's algorithm, Arifin&Setiono's algorithm and Vega's algorithm), purely manual *stemming*. Parameter of the categorization performance are accuracy and f-measure. The accuracy of each case above is compared with the accuracy of non-*stemming* dataset. The categorization is performed with Multinomial Naïve Bayes. The results shows that dataset which is stemmed by the purely manual *stemming* has 97.5% accuracy of categorization and be the best accuracy of categorization of it all. Compared with 3 algorithms for stemming Indonesian that are implemented, stemming using Nazief&Adriani's Algorithm has 95% accuracy of categorization and that's the best accuracy compared with Arifin&Setinoo's algorithm and Vega's algorithm. From this experiment we conclude that stemming is influenced text categorization because stemming can increase the accuracy of categorization.

Keywords: *stemming, text categorization, accuracy, F-Measure*