Abstract

Text categorization is a method to categorize a document based on its contents. Latest research said that a document can be represented as a "bag of words" where a document is represented as a set of *terms* appearing in this document. When we categorize a text, we can not just considerate the frequency of a word, but we must also considerate about the structure of the words. For that reason, we should use *Distributional Feature* for text categorization based on the *compactness of appearance* and *position of the first appearance of a word*.

In this paperwork, the writer tries to combine the *Distributional Feature* with *Term Frequency Inverse Document Frequency* (TFIDF) and *Inverse Term Frequency* (ITF). The writer compares the result of each method by watching the value of its precision, recall, and F1 measure.

The result shows that the use of *Distributional Feature* gives a good result. It is proved by its recall, precision and F1 measure value which is near 1. The performance of *Distributional Feature* (TFIDF·DF, ITF·DF) when use 25%, 50% and 75% of dataset is better than TFIDF. At the other hand, TFIDF shows better performance when use 100% of dataset.

Keywords: text categorization, *Distributional Feature*, TFIDF, TFIDF·DF, ITF·DF, performance