

ABSTRACT

Latent Semantic Analysis (LSA) is one of the most popular methods used in classification task. LSA is used to extract and represent the contextual-usage meaning of words in the document. Commonly, TF-IDF is used as the method to build a term-document matrix or to generate the feature before applying Singular Value Decomposition (SVD) in LSA. Based on the initial experiment, TF-IDF feature in LSA has not performed well to classify the similar text such as between *aqidah* and *ibadah* articles as well as between national and regional articles. This could happen due to the gap of TF-IDF to capture semantic information in the article. Referring to this issue, this study contributes to the use of semantic vector as word representation in text classification with word vector, word2vec which is then processed to the LSA method. This study reveals that the result obtained is better than the previous method. In national and regional article classification, the use of word2vec feature in LSA successfully increased the *f-score* from 74% (LSA with TF-IDF) to 75% (LSA with word2vec) as well as in the accuracy scores that increases from 74% (LSA with TF-IDF) to 78% (LSA with word2vec), meanwhile in *aqidah* and *ibadah* article classification the *f-score* improved from 63% (LSA with TF-IDF) to 73% (LSA with word2vec) as well as in the accuracy score improved significantly from 49% (LSA with TF-IDF) to 72% (LSA with word2vec).

Keywords: classification, LSA, word2vec, Indonesian Islamic article, Indonesian news article, Support Vector Machine