

## **Abstract**

**Semantic similarity is similarity metric between words, sentences or documents that shares element of meaning. Semantic similarity measurement has important role in data mining, information retrieval and even natural language processing. In Indonesian language, semantic similarity measurement has important role because it is widely used for other application, such as text classification. Semantic similarity can be done by corpus based approach and dictionary based approach. In this thesis, the development of corpus based semantic similarity model is represented by distributional semantic vector. The model is then tested on several pairs of words with varying degrees of semantic similarity. The semantic similarity model was build based on Indonesian Wikipedia corpus, with word2vec method. The test result on test dataset which used in previous studies based on SimLex999 dan Rubenstein-goodenough references show the correlation value obtained is 0.2753. Although the correlation value is smaller than value in previous study with the corpus approach, there are numbers of cases where the corpus based semantic model is able to capture the semantic correlation better.**

**Keywords: semantic similarity, Indonesian language, cosinus similarity.**