PROJECT REPORT

COMPARISON OF GLOVE AND FASTTEXT ALGORITHMS ON CNN FOR CLASSIFICATION OF INDONESIAN NEWS CATEGORIES

TJONG, GENESIUS HARTOKO
19.K1.0001

Faculty of Computer Science
Soegijapranata Catholic University
2023

ABSTRACT

Computers cannot understand natural language as humans do, so natural language needs to be converted into something that computers can understand. Word embedding is a term that refers to a method for representing words in natural language into vectors so that computers can understand and perform mathematical operations. In a previous study, the classification of Indonesian news using CNN was carried out but only using the GloVe word embedding algorithm, while in another study it was found that fastText outperformed GloVe in terms of accuracy when classifying English news using CNN. However, because each language has different characteristics, grammar, and structure, this research was conducted to find out whether fastText would also outperform GloVe when using Indonesian news data. The dataset used in this study is a Wikipedia article to train the fastText and GloVe models which will produce a text representation in vector form and be used in the CNN model as a weight on the Embedding layer. The next dataset is Indonesian news with 8 categories for CNN model training, validation, and testing. This study will use 3 different numbers of Wikipedia articles to see the performance of each algorithm when given 10000, 50000, and 100000 Wikipedia articles. The results obtained from this study indicate that fastText outperforms GloVe in accuracy with an average difference of 2.51%, macro precision with an average difference of 4.32%, weighted precision with an average difference of 2.86%, and weighted recall with an average difference of 2.51 %, but for fastText macro recall it only excels when there are 10000 articles with a difference of 11.95% while when there are 50000 and 100000 articles GloVe excels with an average difference of 1.96%.

Keyword: glove, fasttext, indonesian news, cnn