Abstract

Along with the development of the era of technological development also has an increase. Information dissemination occurs very quickly on social media, especially Twitter. On Twitter, only some news circulating is necessarily accurate information. Lots of information that is spread is hoax news that irresponsible individuals apply. In this final project, the author will build a system to determine the optimal amount of data trained in the hoax news classification process. In this study, the authors will use the support vector machine and word2vec algorithms to classify hoax and non-hoax news on the system to be created. In this study, five experiments were carried out with the number of train data used as many as 5000, 10000, 15000, 20000, and 25000. 5000 data train results in an accuracy of 77.28%, 10000 data train produce an accuracy of 79.68%, data 15,000 trains produce an accuracy of 79.892%, 20,000 data trains produce an accuracy of 80,416%, and 25,000 data trains produce an accuracy of 81,184%, by using a combination of unigram with token full token selection.

Keywords: **Hoax, Classification, Support Vector Machine, Word2Vec, Twitter**