Abstract

Cyberbullying is the act of threatening or endangering others by posting text or images that humiliate or harass people through the internet or other communication devices. According to a survey from Polling Indonesia and *Asosiasi Penyelenggara Jasa Internet Indonesia* (APJII) about cyberbullying, 49% of 5900 participants claimed they have been bullied. Therefore, this research conducted with the intention to prevent cyberbullying acts, especially in Indonesia. We collected the data from Twitter based on Twitter's Trending keywords which correlated to cyberbully events. Then we combined it with the data from previous research. We obtained a total of 1425 tweets, consists of 393 data labeled as cyberbully and 1032 data labeled as non-cyberbully. Thereupon, we build the Doc2Vec model for features extraction, and a classifier model using the baseline classification method (SVM and RF) and CNN to detect the cyberbully texts. The results shows that the classifier using CNN and Doc2vec has the highest F1-score, 65.08%.

Keywords: cyberbullying, doc2vec, text classification, convolutional neural network, twitter