Abstract

Hate speech has become a hot issue as it spreads massively on today's social media with specific targets, categories, and levels. In addition, hate speech can cause social conflict and even genocide. This research proposes a system that classifies hate speech written in Indonesian language on Twitter. It also handles the noisiness of twitter data, such as mixed languages and non-standard text. We not only use Support Vector Machines (SVM) as a classifier, but also compare it with other methods, such as deep learning, CNN and DistilBERT. Apart from standard text preprocessing, we propose to accommodate the effect of translating in handling the multilingual content. The data transformation methods used in the SVM model are Label Power-set (LP) and Classifier Chains (CC). The experiment result shows that the classification using the SVM and CC without stemming, stopword removal, and translation provides the best accuracy of 74.88%. The best SVM hyperparameter on multilabel classification is the sigmoid kernel, the regularization parameter value of 10, and the gamma value of 0.1. Stemming, stopword removal, and translation preprocessing are less effective in this research. Moreover, CNN has a flaw in predicting labels for the training data with a low occurrence rate.

Keywords: classification, hate speech, social media, support vector machine.