Abstract

The use of social media in society continues to increase over time. The ease of access and familiarity of social media makes it easier for an irresponsible user to do unethical things such as spreading hatred, defamation, radicalism, pornography, etc. Although there are regulations that govern all the activities on social media, it is still not working effectively due to the impossibility of classifying the comments manually. Therefore, we conducted this study to classify comment into their toxicity categories using machine learning methods for convenience purposes on social media usage. The method that we used in this study is SVM with TF-IDF as the feature extraction and Chi Square as the feature selection. We also performed several exploration scenarios, including implementing SVM kernels and preprocessing stages to find out the best performance of the model. The best performance obtained using the SVM model with a linear kernel, without implementing Chi Square, and using stemming and stopwords removal with the $F\,1-$ Score equal to 76.57%.

Keywords: text classification, toxic comment, social media, support vector machine