ABSTRACT

Hate speech is a form of communication containing hatred by doing things, such as inciting, insulting, disparaging, or demeaning a person or group. Hate speech issues in Indonesia are often related to politics. In 2018 and 2019, for example, the hate speech related to the local leader and presidential elections. The hate speech actors commonly use social networks, such as Instagram, to spread their hatred words. About 60% of hate speech was found in the comments of the posts and it would be a real threat if not quickly detected. This study aims to detect hate speech in Instagram comments. It proposes the use of a word2vec method with skip-gram models and a modified TextCNN to learn and detect hate speech texts. Furthermore, random oversampling, random undersampling, and class weight methods are used to solve imbalanced dataset problems. The results show that the best accuracy, in terms of F-score, is 99.25% which is better than the method from previous research that was a combination of Fasttext, word Bi-gram, 33 hashtags dataset, and random oversampling method.

Keywords—Hate speech comments, Instagram, Word2vec, TextCNN, Fasttext, Imbalance dataset