Deteksi Ujaran Kebencian pada Komentar Twitter Indonesia Menggunakan *Convolutional Neural Network*(CNN) dan *FastText Word Embedding*

Fadhilah Nadia Puteri[1], Yuliant Sibaroni [2], Fitriyani[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]fadhilahnadiap@students.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id,
[3]fitriyani@telkomuniversity.ac.id

Abstract

Hate speech is a problem that is often present in Indonesia, including on social media platforms such as Twitter. Refers to any form of communication, whether oral, written, or symbolic, that may offend, threaten or insult an individual or group based on attributes such as religion, race, ethnicity, sexual orientation, or other characteristics. The existence of freedom of expression and communication on social media triggers the spread of hate speech quickly and widely. To avoid this, a system is needed that can detect hate speech on social media. Deep learning is potentially better at recognizing and analyzing language patterns that reflect hate speech in text. In the previous study, the accuracy obtained was 73.2% using the Convolutional Neural Network method. This study proposed a hate speech detection system using Convolutional Neural Network model and FastText word embedding. The performance of Convolutional Neural Network classification model and FastText as word embedding provide excellent performance results in detecting hate speech, by involving the K-Fold Cross Validation process to the appropriate dropout value is able to achieve an accuracy value of 80%. The resulting accuracy value can be a benchmark that the model that has been built is able to avoid the spread of hate speech on social media.

Keywords: hate speech, twitter, deep learning, convolutional neural network, fast text