Abstrak

Twitter adalah media sosial populer untuk mengirim pesan teks, tetapi tweet yang dapat dikirim dibatasi hingga 280 karakter. Oleh karena itu, pengiriman tweet dilakukan dengan berbagai cara, seperti slang, singkatan, atau bahkan pengurangan huruf dalam kata yang dapat menyebabkan ketidaksesuaian kosakata sehingga sistem menganggap kata dengan arti yang sama berbeda. Dengan demikian, menggunakan perluasan fitur untuk membangun korpus kesamaan dapat mengurangi masalah ini. Dua dataset digunakan untuk membangun similarity corpus: Twitter dataset sebanyak 63.984 dan IndoNews dataset sebanyak 119.488. Kontribusi penelitian adalah menggabungkan deep learning dan ekspansi fitur dengan kinerja yang baik. Penelitian ini menggunakan FastText sebagai perluasan fitur yang berfokus pada struktur kata. Selain itu, penelitian ini menggunakan empat metode deep learning: Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), dan kombinasi dari dua deep learning CNN-GRU, GRU-CNN dengan representasi boolean sebagai ekstraksi fitur. Penelitian ini menggunakan lima skenario untuk mendapatkan hasil terbaik: best data split, n-grams, max feature, feature expansion, dan persentase dropout. Pada model akhir, CNN memiliki performa terbaik dengan akurasi 88,79% dan meningkat 0,97% dari model baseline, diikuti oleh GRU dengan akurasi 88,17% dengan peningkatan 0,93%, CNN-GRU dengan akurasi 87,47% dengan peningkatan 1,86%, dan GRU-CNN dengan akurasi 87,55% dengan peningkatan 1,32%. Berdasarkan hasil dari beberapa skenario, penggunaan perluasan fitur menggunakan FastText berhasil menghindari vocabulary mismatch, dibuktikan dengan peningkatan akurasi model yang paling tinggi dibandingkan skenario lainnya. Namun, penelitian ini memiliki keterbatasan yaitu dataset digunakan dalam bahasa Indonesia.

Kata kunci : ujaran kebencian, fasttext, feature expansion, hybrid deep learning, convolutional neural network, gate recurrent unit.