

Abstrak

Bahasa kasar adalah ungkapan yang digunakan oleh seseorang dengan penghinaan yang menghina aspek seseorang. Di era modern, penggunaan kata-kata kasar sering ditemukan di internet, salah satunya di bagian komentar artikel berita online yang berisi pelecehan, penghinaan, atau kutukan. Sistem deteksi bahasa yang kasar penting untuk mencegah efek negatif dari komentar tersebut. Mendeteksi bahasa kasar di bagian komentar online adalah tantangan karena bahasa kasar dapat diungkapkan dengan berbagai kata. Selain itu, hanya beberapa penelitian yang dilakukan dalam bahasa Indonesia. Dalam makalah ini, kami menyajikan sistem deteksi bahasa Indonesia yang kasar dengan menangani masalah ini sebagai tugas klasifikasi dan menyelesaikannya menggunakan pengklasifikasi berikut: Naive Bayes, SVM, dan KNN. Kami juga melakukan prosedur pemilihan fitur berdasarkan nilai Mutual Information antara kata-kata. Hasil percobaan menunjukkan bahwa SVM adalah pengklasifikasi terbaik untuk mendeteksi bahasa yang kasar dalam komentar berita dengan skor akurasi 90,19 % dan penggunaan Mutual Information mampu meningkatkan akurasi klasifikasi sebesar 1,63%. Saling Informasi dapat meningkatkan akurasi kinerja penggolong.

Kata kunci : Bahasa Kasar, Bahasa Indonesia, Deteksi, SVM, Mutual Information