Abstrak

Ujaran kebencian menjadi isu hangat mengingat maraknya penyebaran ujaran kebencian di media sosial saat ini. Tindakan tersebut memiliki target, kategori, dan tingkat tertentu. Selain itu, ujaran kebencian dapat menimbulkan konflik sosial bahkan genosida. Pada penelitian ini dilakukan klasifikasi multilabel pada ujaran kebencian berbahasa Indonesia di Twitter. Selain itu, dilakukan penanganan *noise* pada data Twitter, seperti bahasa campuran dan kata yang tidak standar. Model klasifikasi utama yang digunakan pada penelitian ini adalah SVM. Namun, metode tersebut dibandingkan dengan metode *deep learning*, yaitu CNN dan DistilBERT. Selain praproses teks yang standar, kami juga melaporkan efek *translate* dalam menangani twit multi-bahasa. Metode transformasi data yang digunakan pada model SVM adalah *Label Power-set* (LP) dan *Classifier Chains* (CC). Hasil dari eksperimen menunjukkan bahwa klasifikasi menggunakan model SVM dengan CC yang menggunakan dataset tanpa *stemming*, *stopword removal*, dan *translate* menghasilkan nilai akurasi terbaik mencapai 74.88%. Sementara itu, *hyperparameter* SVM terbaik dalam klasifikasi multilabel adalah kernel sigmoid, nilai parameter regularisasi sebesar 10, dan nilai gamma sebesar 0.1. Proses *stemming*, *stopword removal*, dan *translate* kurang efektif dalam penelitian ini dan CNN memiliki kekurangan dalam memprediksi label yang memiliki tingkat kemunculan rendah pada data latih.

Kata Kunci: klasifikasi, ujaran kebencian, media sosial, *support vector machine*.