Abstrak

Penggunaan media sosial yang popular seperti Twitter, memunculkan konten yang positif maupun negatif. Salah satu konten negatif yang sering muncul yaitu *Hate Speech*. Tujuan penelitian ini adalah membangun sistem deteksi *Hate Speech* dengan menerapkan *feature expansion* untuk meminimalisir ketidakcocokan kosakata. Penelitian ini juga menggunakan beberapa metode seperti *TF-IDF* untuk mengekstraksi fitur dari sebuah kalimat. Sedangkan *feature expansion* menggunakan metode *GloVe* dan menggunakan metode klasifikasi *logistic regression, random forest* dan *naïve bayes.* Dihasilkan akurasi yang cukup tinggi sebesar 86.44% pada algortima *random forest* dengan fitur Top 5 dalam mendeteksi *Hate Speech* sehingga dapat lebih akurat dalam mendeteksi mana *tweet* yang mengandung *Hate Speech* dan mana yang tidak mengandung *Hate Speech*.

Kata kunci : *TF-IDF, GloVe, feature expansion*, klasifikasi, Twitter, *Hate Speech*