ABSTRAK

Hate speech atau ujaran kebencian di Indonesia yang sering dijumpai di media sosial memiliki dampak buruk bagi semua pengguna. Meskipun banyak kasus yang tercatat oleh Kemkominfo, ujaran kebencian yang tersirat dalam teks kalimat bisa dianggap sebagai kebencian atau bukan kebencian, tergantung orang yang menafsirkannya. Maka dari itu, diperlukan proses klasifikasi suatu teks sehingga dapat diketahui apakah teks tersebut termasuk *hate speech* atau bukan. Tantangan yang dihadapi dalam proses klasifikasi teks *hate speech* adalah bahwa *hate speech* merupakan fenomena yang sulit untuk didefinisikan dan tidak monolitik, sehingga diperlukan sebuah model klasifikasi yang dapat mengenali pola-pola kalimat tersebut.

Solusi untuk masalah tersebut adalah dengan menerapkan *deep learning* untuk melakukan klasifikasi teks *hate speech*. Penelitian ini menggunakan algoritma *Long-Short Term Memory* (LSTM) yang dikombinasikan dengan *word embedding* untuk melakukan klasifikasi teks *hate speech*. LSTM merupakan pengembangan dari algoritma RNN yang termasuk algoritma *deep learning*. Studi literatur menunjukkan bahwa LSTM dengan *word embedding* memiliki performa yang bagus untuk melakukan klasifikasi dengan data berupa teks. Data yang digunakan pada penelitian ini bersumber dari data penelitian terdahulu serta penambahan data yang bersumber dari Twitter.

Hasil pengujian model menggunakan *K-Fold Cross Validation* menunjukkan nilai *Accuracy* sebesar 85,04%, *Precision* sebesar 84,19%, *Recall* sebesar 79,17% dan *F1-Score* sebesar 81,60%. Model yang telah dibangun menggunakan algoritma *Long Short Term Memory* (LSTM) dapat mengenali pola-pola kalimat *hate speech* sehingga dapat melakukan klasifikasi teks *hate speech* dengan hasil yang lebih baik.

Kata kunci: *Hate Speech*, Media Sosial, Twitter, Klasifikasi Teks, *Deep Learning*, *Word Embedding*, *Long Short Term Memory*