ABSTRAK

Hate speech atau ujaran kebencian pada salah satu *platform* sosial media yaitu Twitter sudah tidak jarang ditemukan. Pada *platform* Twitter, pengguna bebas mendapatkan, bertukar informasi, serta mengungkapkan opini. Hal ini merupakan salah satu faktor utama seseorang dapat terkena ujaran kebencian pada Twitter. Korban yang terkena ujaran kebencian memiliki kemungkinan menderita gangguan kesehatan mental, dikarenakan sebagian besar korban ujaran kebencian diserang secara verbal atapun emosional. Minimnya penanggulangan deteksi ujaran kebencian pada *platform* sosial media Twitter masih jarang ditemukan.

Pada penelitian ini, dilakukan proses simulasi menggunakan *website* beserta dengan pengujian dan analisis terhadap pendeteksian ujaran kebencian. Pengujian dilakukan dengan cara pengguna akan melakukan *input* kalimat pada *website hate speech*, lalu *website* akan melakukan *preprocessing* dan menganalisa kalimat tersebut menggunakan Algoritma BERT untuk mengklasifikasikan apakah kalimat tersebut termasuk *hate speech* atau tidak.

Dari hasil pengujian diperoleh bahwa pendeteksian *hate speech* pada akun pengguna Twitter menggunakan Algoritma BERT mendapatkan akurasi sebesar 78.69%, presisi sebesar 78.90%, *recall* sebesar 78.69%, dan *F1 score* sebesar 78.77% terhadap pengklasifikasian golongan *hate speech*. Dengan demikian pengguna akan lebih mudah mendeteksi *hate speech* pada Twitter dengan menggunakan *website hate speech*.

Kata Kunci: Algoritma BERT, Aplikasi Web, *Hate* Speech, Twitter