Abstrak

Twitter merupakan salah satu media sosial yang digunakan untuk menyalurkan opini penggunanya. Tak jarang ada pengguna yang menyalahgunakan fitur ini untuk menuliskan ujaran kebencian. Ujaran kebencian baik disengaja maupun tidak disengaja dapat memicu terjadinya perselisihan. Terutama pada media sosial dengan jumlah pengguna yang besar seperti Twitter, ujaran kebencian dapat dengan mudah menyebar dan menjangkau orang yang menjadi sasarannya. Oleh karena itu dibutuhkan sistem yang dapat mendeteksi ujaran kebencian untuk mencegah penyebarannya serta mengatasinya. Pada penelitian ini penulis akan melakukan fine-tuning terhadap model BERT yang sebelumnya sudah di-*pre-train* untuk mengklasifikasi data yang termasuk ujaran kebencian dari dataset yang sudah dikumpulkan. Adam Optimizer dan AdamW Optimizer akan digunakan pada model BERT yang sudah dibangun untuk kemudian dibandingkan mana yang menghasilkan akurasi terbaik. Penelitian ini dilakukan dengan harapan dapat membantu penanganan kasus ujaran kebencian sehingga dapat dicegah atau diatasi dengan segera. Pada penelitian ini ditemukan bahwa melakukan fine-tuning untuk parameter yang tepat untuk setiap algoritma optimasi dapat menghasilkan akurasi yang lebih tinggi secara signifikan dibandingkan dengan setelan parameter bawaan. Untuk task mengidentifikasi hate speech pada data, model BERT yang dioptimasi dengan AdamW optimizer dapat mencapai akurasi sebesar 90.08% dengan merubah learning rate awal menjadi 1e-5 dan weight decay awal menjadi 1e-3, peningkatan sebesar 40.03% dari baseline. Sementara model BERT yang dioptimasi dengan Adam mencapai akurasi sebesar 90.03% dengan merubah learning rate menjadi 1e-5 dan menggunakan weight decay bawaan, peningkatan sebesar 40.38% dari baseline.

Kata Kunci : Twitter, BERT, Optimizer, Adam, AdamW, Ujaran Kebencian