1. Pendahuluan

Indonesia adalah negara Asia yang memimpin dalam hal penggunaan media sosial, dengan kehadiran yang signifikan di platform-platform yang sangat populer di era digital saat ini [1]. Menurut sebuah studi yang dilakukan oleh Carley dkk. [2], ditemukan bahwa 2,4% tweet di seluruh dunia diposting oleh pengguna yang berada di Jakarta, ibu kota Indonesia. Platform media sosial membantu meningkatkan komunikasi, berbagi informasi, dan mengekspresikan pendapat, serta mendorong praktik kebebasan berpendapat. Namun demikian, penggunaan platform media sosial dapat menimbulkan dampak negatif, salah satunya adalah ujaran kebencian. Oleh karena itu, dibutuhkan sebuah sistem yang dapat mendeteksi ujaran kebencian secara otomatis untuk bahasa Indonesia.

Selama proses pendeteksian, banyaknya variasi kosakata yang digunakan untuk membuat sebuah tweet hanya akan dimengerti jika ada konteksnya [3]. Untuk menangani masalah tersebut dapat diatasi dengan menggunakan *word embedding*. *Word embedding* merupakan teknik yang memetakan kata-kata dari sebuah kosakata ke dalam representasi vektor. Dengan demikian, teknik ini dapat mencari konteks berdasarkan kalimat [4]. Berdasarkan penelitian yang dilakukan oleh Birunda dkk. [4], mereka membandingkan tiga *word embedding*, yaitu *Traditional word embedding*, *Static word embedding*, dan *Contextualized word embedding*. Dari hasil penelitian tersebut, BERT merupakan penyematan kata yang efisien.

Peneliti lain telah melakukan penelitian tentang deteksi ujaran kebencian, seperti *Convolutional Neural Network* (CNN) [5] dan *Bidirectional Gated Recurrent Unit* (BiGRU) [6]. *Convolutional Neural Network* adalah jaringan saraf *feed-forward* yang memperbaiki eror dari *back propagation network*. Pendekatan ini menunjukkan kemahiran ketika berhadapan dengan tantangan pembelajaran mesin yang berkaitan dengan gambar. *Bidirectional Gated Recurrent Unit* merupakan pengembangan dari *Gated Recurrent Unit* (GRU) di mana terdapat dua tumpukan lapisan GRU yang prosesnya berlawanan arah. GRU bertujuan untuk memungkinkan menangkap *depedencies* pada berbagai skala waktu yang berbeda-beda untuk setiap *recurrent unit*.

Penerapan metode *attention mechanism* dalam deteksi ujaran kebencian juga sudah pernah dilakukan sebelumnya. Das dkk. [7] mengusulkan deteksi ujaran kebencian dengan RNN berbasis *attention*. Mereka melakukan klasifikasi dengan CNN, BiLSTM, dan GRU. Pada saat yang sama, *attention mechanism* ditambahkan ke model, yang memungkinkan setiap kata memperoleh setiap bobotnya. Sehingga akan terlihat bobot yang berat dan bobot ringan untuk setiap kata kunci.

Beberapa penelitian telah mencoba menggabungkan berbagai algoritma *deep learning*, yang sering disebut *hybrid deep learning*. Pada tahun 2019, Zhang dan Luo [8] menggabungkan CNN dengan GRU dengan Word2Vec sebagai *word embedding* untuk mengklasifikasikan ujaran kebencian. Huynh dkk. [9] menggabungkan tiga algoritma *deep learning*, yaitu BiGRU, LSTM, dan CNN, dengan menggunakan FastText sebagai *word embedding* untuk mendapatkan f1-score sebesar 70,58% dengan pendekatan ini.

Tujuan utama dari penelitian ini adalah untuk membuat sistem pendeteksi ujaran kebencian yang efektif dalam bahasa Indonesia dengan mengintegrasikan model CNN dan BiGRU. Untuk mencapai tujuan tersebut, kami akan menggunakan BERT sebagai teknik *word embedding*, yang akan membantu sistem untuk lebih memahami konteks dan makna kata. Selain itu, *attention mechanism* akan ditambahkan pada model-model ini untuk meningkatkan kemampuan mereka dalam memprioritaskan informasi penting selama proses deteksi. Kami bertujuan untuk menentukan model hibrida yang paling efektif dengan mempertimbangkan nilai akurasinya.

Makalah ini disusun sebagai berikut. Bagian II membahas isu-isu yang berkaitan dengan deteksi ujaran kebencian. Bagian III menguraikan metode yang kami usulkan. Bagian IV memberikan hasil eksperimen dan analisis kami, sementara bagian V memberikan kesimpulan.