1. Pendahuluan

Media sosial semakin menjadi bagian integral dari kehidupan sehari-hari, terjalin erat dalam interaksi pribadi dan profesional [1]. Di antara berbagai platform, Twitter telah muncul sebagai pilihan populer, menawarkan pengguna kemampuan untuk mengirim dan membaca pesan orang lain, yang biasa disebut tweet. Pembaruan singkat ini, dibatasi hingga 280 karakter, berfungsi sebagai gambaran singkat tentang pemikiran, minat, dan aktivitas pengguna. Sebagai salah satu platform media sosial terbesar, Twitter melihat volume aktivitas yang sangat besar, dengan data menunjukkan sekitar 867 juta tweet dikirim setiap hari [2]. Namun, konten yang disebarluaskan melalui platform ini tidak semuanya bersifat positif. Di antara sekian banyak tweet, sebagian besar berisi konten negatif, dengan cyberbullying sebagai contoh utama. Dalam arti luas, cyberbullying mencakup penggunaan alat komunikasi digital, seperti media sosial, untuk mengirimkan pesan yang mengancam, mengintimidasi, atau merendahkan. Sifat platform seperti Twitter yang tersebar luas menghadirkan serangkaian tantangan unik dalam memantau dan memoderasi interaksi negatif tersebut, sehingga menjadikannya area fokus penting untuk penelitian dan strategi intervensi.

Cyberbullying melibatkan tindakan yang disengaja dan agresif yang dilakukan oleh individu atau kelompok menggunakan saluran komunikasi digital, mengirimkan pesan dan komentar yang berbahaya, mengancam, mengintimidasi, atau menghina kepada individu yang ditargetkan [3]. Dalam voting survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), dari total 5.900 responden, ditemukan sekitar 49% masyarakat di Indonesia mengaku pernah mengalami *bullying* di media sosial [4]. Kemudian, penelitian yang dilakukan oleh Nikon [5] menunjukkan bahwa dampak cyberbullying sangat parah, seperti depresi, stres emosional, dan penggunaan narkoba, serta dapat berujung pada bunuh diri. Sehingga perlu dikembangkan sistem pendeteksi cyberbullying khususnya di twitter secara otomatis, karena banyak pesan harian yang dapat mengandung cyberbullying.

Dalam membuat pesan *cyberbullying* sering dijumpai penggunaan kata-kata yang disingkat dan kata-kata slang, sehingga sulit untuk dipahami, dan terjadi kesalahan kosakata. Sehingga, untuk mengurangi perbedaan kosakata, dapat dilakukan ekspansi fitur menggunakan *word embedding* [5]. Namun, berdasarkan pengetahuan saya, penelitian tentang deteksi *cyberbullying* menggunakan ekspansi fitur masih sangat minim. Beberapa *word embedding* yang digunakan sebagai ekspansi fitur antara lain Word2Vec, FastText, dan Glove [6]. FastText dapat mengungguli Word2Vec dan Glove karena FastText memiliki kemampuan untuk menangani kata-kata langka atau kata-kata yang keluar dari kosakata [8]. Dengan demikian, FastText menjadi pilihan yang tepat untuk menjadi perluasan fitur, terutama untuk kata-kata yang sulit dipahami, seperti di tweet.

Banyak penelitian terkait deteksi *cyberbullying* telah dilakukan. Penelitian terbaru telah mengimplementasikan Hybrid Deep Learning, seperti Joshi, dkk. [7] dan Aldhyani, dkk. [8] menggunakan CNN dan BiLSTM, Dewani, dkk. [9] dengan RNN dan BiLSTM, menghasilkan nilai akurasi yang sangat baik. Kemudian, penelitian terkait cyberbullying untuk dataset berbahasa Indonesia memiliki tantangan karena kebiasaan masyarakat Indonesia berbicara dalam berbagai bahasa seperti bahasa daerah, bahasa gaul, dan singkatan kata yang terkadang sulit dipahami. Jadi, kemampuan FastText diperlukan untuk mengatasi keragaman ini. Namun, pengembangan penelitian terkait deteksi *cyberbullying* menggunakan dataset berbahasa Indonesia masih sebatas Supervised Learning atau Deep Learning, seperti yang dilakukan oleh beberapa peneliti dengan menggunakan model Support Vector Machine (SVM) [10]–[13]. Dalam penelitian lain, Shindy, dkk. [14] mencoba pendekatan Deep Learning dengan metode Convolutional Neural Network (CNN). Berdasarkan pengetahuan saya, ekspansi fitur dan Hybrid Deep Learning belum banyak dieksplorasi dalam tweet berbahasa Indonesia. Oleh karena itu, CNN dan BiLSTM akan diimplementasikan sebagai Hybrid Deep Learning dan ekspansi fitur menggunakan FastText untuk mendeteksi *cyberbullying* di Twitter berbahasa Indonesia.

Kontribusi utama dari penelitian ini adalah menyajikan kombinasi antara model Hybrid Deep Learning dan ekspansi fitur untuk deteksi *cyberbullying* di Twitter berbahasa Indonesia karena sepengetahuan kami belum ada yang melakukan penelitian ini, dan berpotensi meningkatkan nilai akurasi dalam mendeteksi cyberbullying karena berdasarkan penelitian terkait yang telah dijelaskan sebelumnya, penggunaan hybrid deep learning untuk dataset Twitter di negara lain bahasa memiliki akurasi yang tinggi, dan penelitian terkait pendeteksian cyberbullying menggunakan dataset berbahasa Indonesia masih sebatas Supervised Learning atau Deep Learning dan belum menggunakan perluasan fitur. Kemudian, pemilihan FastText sebagai perluasan fitur dapat mengatasi masalah terkait kesalahpahaman kosa kata karena dataset Twitter berbahasa Indonesia memiliki tantangan tersendiri seperti yang telah dijelaskan sebelumnya. Karena itu, berbagai skenario akan dicoba, seperti pemilihan rasio split data dan ekstraksi fitur menggunakan n-gram terbaik. Kombinasi model Hybrid Deep Learning, CNN-BiLSTM dan BiLSTM-CNN, juga akan diuji. Skenario lain melibatkan ekspansi fitur dalam Hybrid Deep Learning, memanfaatkan FastText dan memilih peringkat teratas dengan performa terbaik dari berbagai korpus.

Struktur penelitian ini akan mencakup bagian-bagian berikut: Bagian pertama mengenai pendahuluan. Bagian kedua akan memberikan bukti penelitian yang relevan tentang topik yang dibahas. Bagian ketiga akan menyajikan metodologi penelitian untuk deteksi *cyberbullying* menggunakan CNN dan BiLSTM, dan Bagian keempat akan membahas hasilnya. Akhirnya, Bagian kelima akan menyajikan kesimpulan dari penelitian yang dilakukan.