1. Pendahuluan

Media sosial memungkinkan orang untuk mencari dan memperluas pertemanan mereka. Saat ini, salah satu *platform* media sosial yang sangat populer adalah Twitter, Twitter merupakan layanan media sosial yang memungkinkan penggunanya untuk membuat, memposting, memperbarui, dan membaca pesan singkat yang disebut *tweet* [1]. Twitter telah berkembang menjadi layanan media sosial yang digunakan secara luas dan forum yang menarik di mana lebih dari 500 juta pesan dipertukarkan setiap hari di antara sekitar 1,3 miliar orang [2]. Berdasarkan laporan We Are Social, jumlah pengguna Twitter di Indonesia diproyeksikan mencapai 18,45 juta pada tahun 2022, menempatkan Indonesia sebagai negara terbesar kelima secara global dalam hal pengguna Twitter [3]. Namun demikian, penggunaan Twitter sebagai *platform* media sosial seringkali disalahgunakan oleh sebagian pengguna dengan memposting *tweet* negatif yang mengandung unsur *cyberbullying*.

Cyberbullying dapat didefinisikan sebagai aktivitas menggunakan internet untuk menyakiti atau mengintimidasi orang lain, terutama dengan mengirimkan pesan yang tidak menyenangkan [4]. Menurut survei yang dilakukan oleh UNICEF U-Report terhadap 2.777 remaja berusia 14-24 tahun di Indonesia ditemukan bahwa 45% diantaranya pernah mengalami *cyberbullying* [5]. *Cyberbullying* dapat berdampak pada mental korban, bahkan tidak sedikit kasus korban *bullying* berakhir dengan bunuh diri karena tidak tahan dengan banyak tekanan [6]. Oleh karena itu, sangat penting untuk menerapkan langkah-langkah pencegahan, seperti mengembangkan sistem deteksi *cyberbullying* pada platform media sosial.

Dalam mengatasi masalah ini, banyak penelitian terkait pendeteksian *cyberbullying* telah dilakukan, beberapa di antaranya telah menggunakan pendekatan *hybrid* yang menggabungkan model klasifikasi *Convolutional Neural Network* (CNN) dengan *Long-Short Term Memory* (LSTM) [7], [8]. Dengan menerapkan konsep ini didapatkan nilai akurasi yaitu sebesar 84%. Di Indonesia juga telah dilakukan penelitian terkait deteksi *cyberbullying* dengan dataset berbahasa Indonesia, beberapa peneliti mengimplementasikan model *Support Vector Machine* (SVM) [6], [9], [10]. Namun nilai akurasi yang diperoleh relatif rendah yaitu 76%. Beberapa peneliti lainnya telah menerapkan pendekatan *deep learning* dengan mengimplementasikan model CNN, LSTM, dan BiLSTM [11]–[13]. Dan nilai akurasi yang diperoleh dari ketiga penelitian tersebut adalah 65%, 76%, dan 81%.

Penelitian ini mengusulkan pendekatan *hybrid deep learning* dan ekspansi fitur menggunakan Word2Vec dalam membangun sistem deteksi *cyberbullying* di Twitter berbahasa Indonesia. Penggunaan ekspansi fitur dengan *word embedding* Word2Vec dinilai dapat membantu mengurangi ketidaksesuaian kosakata pada *tweet* yang menggunakan variasi kata atau tweet yang disingkat oleh pengguna Twitter [14]. Penelitian ini mengusulkan pendekatan *hybrid deep learning* untuk membangun sistem deteksi *cyberbullying* dimotivasi oleh pemahaman kami dari penelitian sebelumnya tentang pembuatan sistem deteksi *cyberbullying* yang menggunakan konsep *hybrid deep learning* pada dataset *tweet* berbahasa lain, secara konsisten memberikan nilai akurasi yang tinggi. Selain itu, penelitian yang ditujukan untuk mengembangkan pendeteksian *cyberbullying* dengan dataset berbahasa Indonesia masih terbatas pada konsep *supervised learning* dan *deep learning*, yang mana dari keterbatasan tersebut menjadi motivasi bagi kami untuk berhasil menerapkan *hybrid deep learning* untuk mendeteksi cyberbullying dalam dataset berbahasa Indonesia.

Penelitian ini disusun sebagai berikut: Bagian pertama berisi pendahuluan. Bagian kedua berisi mengenai penjelasan studi terkait dari penelitian ini. Bagian ketiga berisi penjelasan metode yang digunakan dan penerapannya. Bagian keempat berisi hasil pengujian dan analisisnya. Bagian kelima berisi kesimpulan dari penelitian ini.