1. Pendahuluan

Ujaran kebencian atau *hate speech* adalah salah satu kejahatan yang sering terjadi di tengah masyarakat dengan atau tanpa menggunakan teknologi informasi melalui jejaring sosial. Penyebaran ujaran kebencian seringkali dilakukan dengan menyebarkan informasi buruk seseorang atau sekelompok orang melalui media sosial [1]. Perkembangan media sosial mempengaruhi tumbuhnya ujaran kebencian, di mana masyarakat bebas berekspresi tanpa batasan. Twitter merupakan salah satu media sosial terbesar yang digunakan untuk berkomunikasi, media sosial dengan 18,45 juta pengguna pada tahun 2022.

Twitter adalah layanan jejaring sosial yang memungkinkan penggunanya untuk saling berkirim pesan yang dikenal dengan tweet [2]. Pada tahun 2021, Kementerian Komunikasi dan Informatika mengumumkan telah mengatasi 3.640 konten terkait dengan isu Ujaran Kebencian berdasarkan Suku, Agama, Ras, dan antar golongan (SARA) [3]. Dengan jumlah insiden ujaran kebencian tentunya berdampak negatif bagi pengguna media sosial. Oleh karena itu, identifikasi ujaran kebencian merupakan alat yang dapat meredam atau mencegah kerusuhan yang dapat memecah belah keutuhan bangsa Indonesia.

Deep learning merupakan algoritma yang banyak digunakan oleh para peneliti untuk mengembangkan klasifikasi ujaran kebencian [4]–[6]. Banyak peneliti yang telah melakukan hal tersebut, namun hasilnya terkadang kurang memuaskan, seperti akurasi yang rendah yang disebabkan oleh terbatasnya data pelatihan [4]. Selain itu, dalam proses pendeteksian, kosa kata yang tidak tepat membuat kalimat yang diunggah dalam bentuk tweet menjadi sulit dipahami. Ekspansi fitur dapat mengatasi masalah ini dengan penyisipan kata dalam tweet [9].

Telah dilakukan penelitian tentang klasifikasi ujaran kebencian yang telah dilakukan sebelumnya [4], [5], [7] dengan menggunakan algoritma *machine learning* dan *deep learning* seperti *Logistic Regression (LR), Random Forest (RF),* dan *Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Long Short-Term Memory* (*LSTM*). Namun, tidak satu pun dari ketiga penelitian tersebut yang menerapkan pembelajaran mendalam *hybrid*. Penelitian [8] telah menggunakan metode *hybrid deep learning* yang telah dilakukan dengan menggunakan metode model deep learning dari *CNN* dan *LSTM* menghasilkan akurasi terbaik sebesar 92,9% namun digunakan dalam analisis sentimen.

Berdasarkan uraian masalah dan penelitian sebelumnya, maka penelitian ini berkontribusi untuk meningkatkan akurasi penelitian sebelumnya dengan menggunakan metode *deep learning hybrid CNN* dan *LSTM* serta metode fitur ekspansi *GloVe* dengan jumlah data yang lebih banyak menggunakan dataset Twitter berbahasa Indonesia. Sepengetahuan kami, *hybrid* dan *GloVe* belum pernah digunakan dalam penelitian sebelumnya tentang ujaran kebencian.