## 1. Pendahuluan

### Latar Belakang

Sebanyak 202,6 juta individu di Indonesia memanfaatkan layanan internet, mewakili 73,7% dari keseluruhan populasi Indonesia [1]. Mayoritas dari para pengguna internet ini terlibat dalam berbagai platform media sosial. Data mencatat bahwa jumlah individu yang memanfaatkan media sosial mencapai 170 juta, yang setara dengan 83,9% dari total pengguna internet di Indonesia [1]. Salah satu contoh platform media sosial yang sangat digemari adalah Twitter. Berdasarkan informasi yang diterbitkan oleh Departemen Riset Statistik, jumlah pengguna aktif Twitter tercatat sebanyak 229 juta dari tahun 2017 hingga 2022 [2]. Di Indonesia, pengguna Twitter diperkirakan mencapai 18,45 juta individu [2].

Media sosial digunakan oleh masyarat sebagai media untuk berkomunikasi secara daring. Pada Twitter sendiri, pengguna dapat saling berbagi *tweet* yang dapat dilihat secara publik, selain itu pengguna lain dapat memberikan komentar pada *tweet* tersebut. Tidak jarang platform ini dimaanfaatkan untuk hal yang tidak baik yaitu melakukan ujaran kebencian. Ujaran kebencian adalah tindakan menyebarkan kebencian terhadap individu atau kelompok tertentu atas dasar suku, agama, ras, dan karakteristik lain yang dapat menimbulkan diskriminasi, kekerasan, dan konflik sosial [3]. Ujaran kebencian dapat mencakup penghinaan, pencemaran nama baik, perbuatan yang tidak menyenangkan, tindakan provokatif, ajakan, dan penyebaran informasi palsu. Sanksi terhadap ujaran kebencian diberlakukan karena perilaku tersebut dapat menyebabkan terjadinya diskriminasi, tindakan kekerasan, hilangnya nyawa, atau bahkan konflik sosial. [4]. Ujaran kebencian pada media sosial harus dideteksi untuk menghindari konflik antara warga dan menghindari generasi muda untuk mempelajari ujaran kebencian atau bahasa yang tidak pantas dari media sosial yang mereka gunakan [5].

Salah satu tindakan untuk mengatasi hal tersebut adalah melakukan deteksi kalimat ujaran kebencian pada media sosial yang dapat berjalan secara otomatis. Terdapat penelitian yang telah dilakukan terkait deteksi ujaran kebencian pada media sosial, Pada studi sebelumnya, penelitian telah dilaksanakan menggunakan teknik-teknik seperti SVM (*Support Vector Machine*), NB (*Naive Bayes*), dan RFDT (*Random Forest Decision Tree*) bersama dengan pendekatan BR (*Binary Relevance*), LP (*Label Power-set*), serta CC (*Classifier Chains*) sebagai metode untuk mentransformasi data [6]. Tetapi dalam beberapa tahun ini model bahasa menggunakan pre-trained model telah terbukti efektif untuk meningkatkan banyak tugas dalam bidang *Natural Language Processing* (NLP) [7]. *Pre-trained* model adalah sebuah model bahasa atau model *machine learning* yang sebelumnya telah melewati tahapan pelatihan dengan *dataset* besar dan beragam kemudian dapat digunakan untuk menyelesaikan tugas lain [8]. Beberapa penelitian yang telah menggunakan metode *pre-trained* model seperti deteksi ujaran kebencian dan bahasa kasar pada Twitter menggunakan *Bidirectional Encoder Representations From Transformers* (BERT) dan deteksi penggunaan kalimat abusive pada teks Indonesia menggunakan *Indonesia Bidirectional Encoder Representations from Transformers* (IndoBERT).

Penelitian ini menggunakan metode IndoBERTweet untuk mendeteksi ujaran kebencian di Twitter. Metode ini merupakan model BERT Indonesia yang dilatih secara monolingual (hanya menggunakan bahasa Indonesia) dengan penambahan kosakata khusus untuk mendukung domain tertentu, khususnya dalam adaptasi model yang efisien terhadap data berbasis Twitter [10]. Dalam proses pengerjaan penelitian ini, dilakukan pengumpulan data dengan metode crawling untuk membangun *dataset*. *Dataset* tersebut kemudian diberi label sebagai ujaran kebencian dan bukan ujaran kebencian. Selanjutnya, dilakukan eksperimen dengan melakukan perubahan pengaturan *hyperparameter* yang berbeda-beda guna menganalisis dampaknya terhadap performa model yang digunakan.

### Topik dan Batasannya

Topik penelitian adalah deteksi ujaran kebencian dalam teks bahasa Indonesia yang berasal dari *tweet* pada platform Twitter. Sebagai langkah awal, dibuat *dataset* dari *tweet* dengan bahasa indonesia dan kemudian dilakukan proses pembersihan data agar dapat diproses oleh model. Setelah itu, dengan menggunakan metode IndoBERTweet dilakukan klasifikasikan jenis kalimat yang terdapat dalam data tersebut kedalam dua label yaitu ujaran kebencian (*hate speech)* dan bukan ujaran kebencian (*not hate speech)*. Gambaran sistem dapat dilihat pada gambar 1 berikut:


**Gambar 1**. Gambaran Umum Sistem

### Tujuan

Tujuan dari penelitian ini adalah melakukan deteksi ujaran kebencian pada Twitter Indonesia dan mengevaluasi penerapan metode IndoBERTweet. Pada tahap eksplorasi penerapan metode IndoBERTweet,

dilakukan pencarian hyperparameter yang berbeda untuk menemukan kombinasi nilai hyperparameter yang menghasilkan model terbaik. Model terbaik kemudian dipilih berdasarkan akurasi yang dicapai dari setiap percobaan dengan variasi nilai hyperparameter. Dari model tersebut, akan dilakukan pengujian untuk mengidentifikasi jumlah ujaran kebencian dalam lingkup besar.

**Organisasi Tulisan**

Bagian 2 menjelaskan mengenai dasar teori yang menjadi landasan dalam melakukan penelitian ini. Teori-teori tersebut akan memberikan pemahaman yang mendalam mengenai konsep-konsep yang relevan dan mendukung penelitian ini. Pada bagian 3, akan dijelaskan alur penelitian yang telah dilakukan. Alur penelitian ini meliputi tahapan-tahapan yang telah dilakukan untuk mencapai tujuan penelitian. Bagian 4 akan berfokus pada hasil yang diperoleh dari sistem yang telah dibangun. Hasil tersebut akan dianalisis dan dijelaskan dengan baik, sesuai dengan tujuan penelitian. Bagian 5 adalah kesimpulan yang didapatkan dari penelitian ini. Selain itu juga akan diberikan saran-saran yang dapat dikembangkan untuk penelitian selanjutnya. Saran ini dapat meliputi pengembangan metodologi, perluasan cakupan penelitian, atau penelitian lanjutan yang dapat dilakukan untuk meningkatkan pemahaman dan pengetahuan dalam bidang ini.