BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Ujaran kebencian (*hate speech*) merupakan salah satu bentuk kekerasan verbal yang semakin sering muncul dalam interaksi digital, khususnya di media sosial. Di Indonesia, fenomena ini menjadi perhatian serius karena tidak hanya mengancam kerukunan sosial tetapi juga berimplikasi pada keamanan dan stabilitas nasional. Menurut laporan *Digital 2025: Indonesia* oleh DataReportal, pada Januari 2025 terdapat sekitar 143 juta identitas pengguna media sosial di Indonesia, yang setara dengan 50,2% dari total populasi negara [2]. Tingginya penetrasi media sosial ini menciptakan ruang yang luas bagi penyebaran ujaran kebencian, sehingga menuntut perhatian dan penanganan yang serius dari berbagai pihak.

Berdasarkan studi Microsoft tahun 2021, Indonesia menempati peringkat 29 dari 32 negara yang dihitung dari angka keberadaban *digital*. Tercatat angka keberadaban *digital* Indonesia menyentuh 76. Semakin tinggi angka keberadaban *digital* semakin tidak sopan komunikasi di dunia *digital* [3, 4]. Pada tahun 2024, berdasarkan laporan Aliansi Jurnalis Independen (AJI), media sosial khususnya X mengandung ujaran kebencian sebesar 36,35% yang menyerang kelompok rentan selama periode kampanye Pemilu 2024, terutama yang menyasar kelompok penyandang disabilitas, Israel, Yahudi dan kelompok minoritas seperti Kristen dan *Tiong Hua* [5].

Upaya untuk mendeteksi ujaran kebencian telah dilakukan dengan berbagai pendekatan. Metode tradisional seperti Support Vector Machine (SVM), Naive Bayes, dan Logistic Regression telah digunakan dalam sejumlah penelitian, namun masih menunjukkan keterbatasan dalam memahami konteks linguistik yang kompleks. Metode SVM hanya mampu mencapai akurasi sekitar 82% dalam mendeteksi ujaran kebencian pada data berbahasa Indonesia [6]. Penelitian lain juga mendukung temuan ini, di mana model tradisional seperti Naive Bayes dan SVM yang digunakan untuk klasifikasi multi-label ujaran kebencian pada Twitter berbahasa Indonesia menghasilkan performa yang terbatas, dengan nilai F1-score rata-rata makro sebesar 0,77 dan mikro sebesar 0,84, bahkan ketika menggunakan kombinasi metode terbaik seperti Random Forest dan Label Powerset [7]. Keterbatasan ini disebabkan oleh ketidakmampuan model tradisional dalam

menangkap konteks kata, terutama dalam kalimat yang mengandung makna ganda atau sarkasme.

Untuk mengatasi keterbatasan tersebut, pendekatan berbasis pembelajaran mendalam (*deep learning*) terutama dengan arsitektur *Transformer* seperti BERT mulai digunakan secara luas. BERT (*Bidirectional Encoder Representations from Transformers*) memiliki keunggulan dalam memahami konteks kata secara menyeluruh karena kemampuannya membaca kalimat secara dua arah (kiri ke kanan dan kanan ke kiri), berbeda dengan pendekatan tradisional atau *word embedding* statis seperti Word2Vec dan GloVe yang hanya melihat konteks secara parsial. Selain itu, BERT memungkinkan fine-tuning pada berbagai tugas pemrosesan bahasa alami (NLP) dengan hasil yang sangat baik meskipun dengan jumlah data pelatihan terbatas [1]. Secara global, studi oleh Jahan dan Oussalah (2023) menunjukkan bahwa model berbasis Transformer seperti BERT dan RoBERTa rata-rata memiliki peningkatan akurasi sebesar 5–12% dibanding model LSTM atau SVM pada tugas deteksi ujaran kebencian, terutama dalam bahasa yang kompleks[8].

IndoBERT, sebagai model bahasa berbasis BERT yang dilatih khusus pada korpus berbahasa Indonesia, menunjukkan kinerja yang jauh lebih baik. Fadhli et al. (2020) memperkenalkan IndoLEM, kumpulan data benchmark untuk bahasa Indonesia, dan menunjukkan bahwa IndoBERT secara konsisten mengungguli mBERT dan pendekatan tradisional lainnya dalam berbagai tugas seperti part-of-speech tagging, named entity recognition, dan dependency parsing, karena kemampuannya memahami konteks dan struktur sintaksis bahasa Indonesia secara lebih dalam [9]. Penelitian oleh Wibowo [10] menunjukkan bahwa IndoBERT mampu mencapai akurasi hingga 91% dan F1-score 0.88 dalam tugas klasifikasi ujaran kebencian dengan *epoch* 10, *learing rate* 1e-1. Lalu penelitian oleh [11] menunjukan IndoBERT mencapai akurasi tertinggi 89% dengan *epoch* 3, *learning rate* 2e-5 dan *batch size* 16 dengan *dataset* komentar Youtube.

Selain itu, model IndoBERTweet yang dikembangkan oleh Koto et al. (2021) merupakan model pralatih yang dilatih menggunakan lebih dari dua juta tweet berbahasa Indonesia. Model ini menunjukkan peningkatan performa dengan peningkatan F1-score dibanding FastText dan CNN. [12]. Pada penelitian [13] IndoBERTweet mampu mencapai akurasi 93% dengan *epoch* 5, *learning rate* 1e-5 dan *batch size* 10 dengan jumlah *dataset* 13.700.

Oleh karena itu, dibutuhkan eksplorasi terhadap IndoBERT dan IndoBERTweet. Penelitian ini bertujuan untuk mengeksplorasi dan

membandingkan efektivitas dari IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian, guna memberikan kontribusi bagi pengembangan sistem klasifikasi yang lebih akurat dan kontekstual.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana implementasi model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian pada data Twitter berbahasa Indonesia?

2. Berapa akurasi, *F1-score*, *Precision* dan *recall* dari model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian pada data Twitter berbahasa Indonesia?

1.3 Batasan Permasalahan

Dalam penelitian ini, terdapat beberapa batasan yang ditetapkan guna mempertajam fokus penelitian agar sesuai dengan rumusan masalah yang telah dipaparkan sebelumnya. Batasan-batasan tersebut mencakup aspek data, metode, dan cakupan penelitian sebagai berikut:

1. Penelitian ini hanya akan fokus pada deteksi ujaran kebencian dalam bahasa Indonesia, khususnya yang ditemukan di platform media sosial Twitter. Bahasa informal dan penggunaan slang dalam teks Twitter akan menjadi fokus utama dalam penelitian ini.

2. Penelitian ini hanya akan membandingkan dua model bahasa berbasis Transformer, yaitu IndoBERT dan IndoBERTweet. Tidak akan dilakukan perbandingan dengan model lain seperti BERT atau model non-Transformer lainnya yang tidak relevan dengan penelitian ini.

3. Penelitian ini hanya akan berfokus pada klasifikasi ujaran kebencian dan tidak akan mencakup klasifikasi kategori lain seperti bahasa kasar, ujaran

kebencian berbasis ras atau agama, dan serangan terhadap individu tertentu.

4. Penelitian ini tidak akan secara mendalam menganalisis konteks sosial dan budaya dari ujaran kebencian yang terdeteksi, melainkan hanya akan mengukur seberapa baik model dapat mengidentifikasi ujaran kebencian dalam data yang diberikan.

1.4 Tujuan Penelitian

1. Implementasi model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian pada data Twitter berbahasa Indonesia

2. Mengukur akurasi, *F1-score*, *precision* dan *recall* dari model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian pada data Twitter berbahasa Indonesia

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan beberapa manfaat, baik secara teoritis maupun praktis, antara lain:

1. Manfaat Teoritis
   Penelitian ini dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan di bidang pemrosesan bahasa alami (Natural Language Processing/NLP), khususnya dalam konteks bahasa Indonesia. Hasil analisis komparatif antara model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian diharapkan dapat menjadi referensi untuk penelitian selanjutnya yang berkaitan dengan klasifikasi teks dan deteksi ujaran kebencian.
2. Manfaat Praktis
   Penelitian ini dapat memberikan informasi bagi praktisi dan pengembang sistem dalam memilih model yang lebih optimal dalam membangun sistem deteksi ujaran kebencian pada platform media sosial. Dengan mengetahui

kinerja masing-masing model, instansi pemerintah, perusahaan teknologi, atau organisasi nirlaba dapat menggunakan pendekatan yang lebih tepat guna dalam menangani konten negatif secara otomatis di media sosial, khususnya Twitter.

1.6 Sistematika Penulisan

Penulisan skripsi ini disusun secara sistematis ke dalam beberapa bab, dengan rincian sebagai berikut:

1. Bab 1 PENDAHULUAN
   Bab ini menjelaskan latar belakang, rumusan masalah, tujuan penelitian, batasan masalah, serta sistematika penulisan.

2. Bab 2 LANDASAN TEORI
   Bab ini membahas kajian pustaka dan teori-teori yang mendukung penelitian, termasuk penelitian-penelitian terdahulu yang relevan.

3. Bab 3 METODOLOGI PENELITIAN
   Bab ini menjelaskan metodologi penelitian yang digunakan, mulai dari pengumpulan data hingga tahapan evaluasi model.

4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil eksperimen serta analisis perbandingan performa antara model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian.

5. Bab 5 KESIMPULAN DAN SARAN
   Bab ini berisi kesimpulan dari hasil penelitian dan saran untuk pengembangan atau penelitian lebih lanjut.