BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Akselerasi teknologi informasi telah mengubah media sosial dari sekadar *platform* komunikasi menjadi ekosistem informasi global yang membentuk wacana publik dan dinamika sosial masyarakat modern. Indonesia menjadi salah satu negara dengan pertumbuhan pengguna media sosial yang paling signifikan, dengan lebih dari 217,53 juta pengguna aktif pada tahun 2022, menjadikannya negara pengguna media sosial terbesar keempat di dunia setelah Tiongkok, India, dan Amerika Serikat [1]. Secara global, laporan We Are Social dan Kepios pada tahun 2024 mencatat bahwa pengguna media sosial telah mencapai 5,04 miliar atau setara dengan 62,3% dari total populasi dunia, dengan pertumbuhan 266 juta pengguna baru dalam satu tahun [2]. Di Indonesia, dari total populasi sekitar 272 juta jiwa, lebih dari 160 juta orang menjadikan media sosial sebagai sarana utama berinteraksi [3]. Namun, pertumbuhan masif ini diiringi oleh peningkatan penyebaran konten berbahaya di ruang digital, di mana laporan UNESCO menyebutkan bahwa 67% pengguna internet pernah menemukan ujaran kebencian di ruang daring [4], dan Kementerian Komunikasi dan Informatika Indonesia mencatat hampir 1,4 juta konten negatif di media sosial, dengan ujaran kebencian sebagai salah satu kategori dominan [1].

Ujaran kebencian atau *hate speech* didefinisikan sebagai segala bentuk ekspresi yang menyebarkan, menghasut, mempromosikan, atau membenarkan kebencian berbasis ras, xenofobia, antisemitisme, maupun intoleransi lainnya [5]. Dampaknya bersifat multidimensi, mulai dari tingkat individu yang meliputi penurunan kesehatan mental, kecemasan, depresi, dan gangguan stres pasca-trauma, hingga tingkat masyarakat yang dapat memicu polarisasi kelompok, memperdalam divisi sosial, menormalisasi diskriminasi, dan meningkatkan risiko kekerasan fisik terhadap kelompok yang menjadi sasaran [6]. Studi oleh Walther menunjukkan bahwa 42–67% pemuda terpapar ujaran kebencian secara daring dan 21% pernah menjadi korban tulisan atau ucapan yang merendahkan di internet [7]. Bukti empiris menunjukkan bahwa penyebaran ujaran kebencian di media sosial berkorelasi dengan eskalasi konflik sosial di dunia nyata, mengindikasikan bahwa dampak ujaran kebencian tidak terbatas pada ruang digital semata [8]. Dengan demikian,

ujaran kebencian di media sosial tidak hanya menimbulkan konflik sosial, tetapi juga membentuk opini publik dan memperkuat sentimen negatif terhadap individu maupun kelompok tertentu.

Salah satu tantangan utama dalam pendeteksian ujaran kebencian pada media sosial Indonesia adalah keberadaan teks multibahasa dan *code-mixed*, yaitu penggunaan lebih dari satu bahasa dalam satu konteks atau kalimat [9]. Ujaran kebencian di media sosial Indonesia sering kali dikemas dalam bentuk bahasa informal yang mencakup penggunaan kata *slang* atau bahasa gaul, singkatan tidak baku, modifikasi ejaan, serta pencampuran bahasa Indonesia dengan bahasa Inggris yang dikenal dengan *code-mixing* [10, 11]. Fenomena *code-mixing* bahasa Indonesia dan bahasa Inggris sangat umum ditemukan di media sosial, di mana pengguna secara fleksibel mengombinasikan kosakata, frasa, atau klausa dari kedua bahasa dalam satu teks untuk menekankan pesan dan mengekspresikan identitas [6]. Kondisi ini menyebabkan sistem deteksi ujaran kebencian yang umumnya dikembangkan dengan pendekatan *monolingual* masih menunjukkan keterbatasan signifikan dalam menangani teks multibahasa yang bervariasi dan bersifat informal [10].

Pendekatan konvensional dalam deteksi ujaran kebencian, seperti metode *supervised machine learning* berbasis Support Vector Machine (SVM), Convolutional Neural Network (CNN), hingga Long Short-Term Memory (LSTM), umumnya sangat bergantung pada ketersediaan sumber daya linguistik yang spesifik terhadap satu bahasa [12, 13]. Seiring berkembangnya model bahasa besar atau Large Language Model (LLM) yang bersifat multibahasa, muncul peluang untuk mengatasi keterbatasan tersebut. Model berbasis BERT, termasuk *multilingual* BERT dan XLM-RoBERTa, menunjukkan performa kompetitif pada skenario lintas bahasa dan teks *code-mixing* [14, 15]. Namun, sebagian besar penelitian terdahulu masih mengembangkan model secara independen untuk setiap bahasa atau pasangan bahasa, sehingga tidak mampu menangani tiga kelompok bahasa secara simultan dalam satu model terpadu. Selain itu, penelitian yang ada umumnya hanya mengevaluasi model tunggal tanpa mengeksplorasi strategi *ensemble* yang terbukti meningkatkan robustisitas prediksi, serta tidak menyertakan analisis interpretabilitas model yang penting untuk memahami dasar keputusan klasifikasi [5, 16].

Berdasarkan permasalahan dan kesenjangan penelitian tersebut, penelitian ini mengkaji penggunaan model transformer multibahasa untuk mendeteksi ujaran kebencian pada teks media sosial yang bersifat *multilingual*, mencakup Bahasa

Indonesia, Bahasa Inggris, dan teks *code-mixed*. Penelitian ini membandingkan performa dua model transformer multibahasa, yaitu XLM-RoBERTa dan BLOOM-560M, serta mengevaluasi berbagai strategi *fine-tuning* dan metode *ensemble* untuk meningkatkan kemampuan deteksi ujaran kebencian. Untuk meningkatkan robustisitas estimasi performa, penelitian ini menerapkan strategi *5-fold stacking* dengan *out-of-fold predictions* yang secara eksplisit mencegah *data leakage* pada proses pelatihan *meta-learner*. Selain itu, analisis interpretabilitas berbasis Local Interpretable Model-agnostic Explanations (LIME) diterapkan untuk memberikan pemahaman yang lebih mendalam mengenai fitur-fitur linguistik yang mempengaruhi keputusan model, yang merupakan aspek yang jarang dieksplorasi dalam penelitian deteksi ujaran kebencian multibahasa sebelumnya [17].

Melalui pendekatan tersebut, penelitian ini diharapkan dapat memberikan pemahaman yang lebih komprehensif mengenai efektivitas model transformer multibahasa dalam menangani karakteristik bahasa informal dan *code-mixed* yang umum ditemukan pada media sosial, sekaligus memberikan kontribusi metodologis melalui kombinasi strategi *fine-tuning*, metode *ensemble* berlapis, dan analisis interpretabilitas yang belum pernah dieksplorasi secara terpadu pada konteks teks *multilingual* Bahasa Indonesia, Bahasa Inggris, dan *code-mixed* dalam satu penelitian.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah diidentifikasi, rumusan masalah dalam penelitian ini dirumuskan sebagai berikut:

1. Bagaimana performa model XLM-RoBERTa dan BLOOM-560M yang di-*fine-tune* dalam mendeteksi ujaran kebencian pada teks media sosial *multilingual* yang mencakup Bahasa Indonesia, Bahasa Inggris, dan teks *code-mixed*?
2. Bagaimana pengaruh berbagai strategi *fine-tuning* terhadap performa model dalam mendeteksi ujaran kebencian pada data *multilingual*?
3. Bagaimana pengaruh penerapan metode *ensemble* terhadap performa deteksi ujaran kebencian dibandingkan dengan penggunaan model tunggal?

1.3 Batasan Permasalahan

Agar penelitian tetap terfokus dan terarah, batasan permasalahan dalam penelitian ini ditetapkan sebagai berikut:

1. Penelitian difokuskan pada klasifikasi biner ujaran kebencian dan non-ujaran kebencian pada teks media sosial.
2. Dataset yang digunakan terdiri atas teks berbahasa Indonesia, Bahasa Inggris, dan *code-mixed* antara kedua bahasa tersebut.
3. Model utama yang digunakan dalam penelitian ini adalah XLM-RoBERTa dan BLOOM-560M.
4. Eksperimen mencakup evaluasi berbagai strategi *fine-tuning* serta metode *ensemble* untuk meningkatkan performa model.
5. Evaluasi dilakukan menggunakan metrik klasifikasi, dan penelitian ini tidak mencakup implementasi sistem dalam lingkungan produksi maupun sistem *real-time*.

1.4 Tujuan Penelitian

Sejalan dengan rumusan masalah yang telah ditetapkan, tujuan penelitian ini adalah sebagai berikut:

1. Membandingkan dan mengevaluasi performa model XLM-RoBERTa dan BLOOM-560M yang di-*fine-tune* dalam mendeteksi ujaran kebencian pada teks media sosial *multilingual* yang mencakup Bahasa Indonesia, Bahasa Inggris, dan teks *code-mixed*.
2. Menganalisis pengaruh berbagai strategi *fine-tuning* terhadap performa model dalam mendeteksi ujaran kebencian pada data *multilingual*.
3. Mengevaluasi efektivitas metode *ensemble* dalam meningkatkan performa deteksi ujaran kebencian dibandingkan dengan penggunaan model tunggal.

1.5 Kontribusi Penelitian

Kontribusi yang dihasilkan dalam penelitian ini adalah sebagai berikut:

1. Menyusun dan mengonstruksi dataset *multilingual* yang terdiri atas teks Bahasa Indonesia, Bahasa Inggris, dan *code-mixed* dengan distribusi kelas yang seimbang untuk tugas deteksi ujaran kebencian, dengan memanfaatkan sumber data yang tersedia secara publik serta data sintetis yang dihasilkan melalui model bahasa besar.
2. Menyajikan evaluasi komparatif antara model XLM-RoBERTa dan BLOOM-560M pada tugas deteksi ujaran kebencian *multilingual*.
3. Mengevaluasi berbagai strategi *fine-tuning* untuk mengetahui pengaruhnya terhadap performa model transformer multibahasa.
4. Mengevaluasi efektivitas metode *ensemble* dalam meningkatkan performa deteksi ujaran kebencian dibandingkan penggunaan model tunggal.
5. Memberikan analisis terhadap kemampuan model transformer multibahasa dalam menangani karakteristik bahasa informal dan *code-mixed* pada media sosial.

1.6 Manfaat Penelitian

Adapun manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Memberikan kontribusi ilmiah dalam pengembangan metode deteksi ujaran kebencian pada teks media sosial multibahasa, khususnya teks *code-mixed* Bahasa Indonesia dan Bahasa Inggris, melalui perbandingan model transformer multibahasa, evaluasi strategi *fine-tuning*, serta penerapan metode *ensemble*.
2. Menyediakan analisis komparatif terhadap performa model bahasa besar multibahasa, yaitu BLOOM-560M dan XLM-RoBERTa, dalam menangani variasi linguistik informal pada data media sosial.
3. Menjadi referensi bagi penelitian selanjutnya dalam pengembangan sistem deteksi ujaran kebencian yang lebih *robust* terhadap penggunaan bahasa campuran, variasi bahasa informal, serta keterbatasan ketersediaan data berlabel untuk teks *code-mixed*.

1.7 Sistematika Penulisan

1. **Bab 1 PENDAHULUAN:** Memuat latar belakang, rumusan masalah, batasan masalah, tujuan, manfaat, dan sistematika penulisan.
2. **Bab 2 LANDASAN TEORI:** Membahas teori-teori pendukung serta penelitian terdahulu yang relevan dengan topik penelitian.
3. **Bab 3 METODOLOGI PENELITIAN:** Menjelaskan tahapan penelitian, meliputi deskripsi model XLM-RoBERTa dan BLOOM-560M, sumber dan karakteristik dataset, proses persiapan data, strategi *fine-tuning* yang digunakan, metode *ensemble* yang diterapkan, serta skema pelatihan dan evaluasi model.
4. **Bab 4 HASIL DAN DISKUSI:** Menyajikan hasil eksperimen beserta evaluasi performa model menggunakan metrik akurasi dan F1-Score, analisis *confusion matrix*, evaluasi per bahasa, perbandingan dengan penelitian terdahulu, serta pembahasan terhadap hasil yang diperoleh.
5. **Bab 5 KESIMPULAN DAN SARAN:** Berisi kesimpulan dari penelitian yang telah dilakukan, keterbatasan penelitian, serta rekomendasi untuk penelitian selanjutnya.