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. 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 [1]. Indonesia menjadi salah satu negara dengan pertumbuhan pengguna media sosial yang paling signifikan, dengan lebih dari 217,53 juta pengguna aktif, menjadikannya negara pengguna media sosial terbesar keempat di dunia [2]. Dari total populasi nasional, lebih dari 160 juta orang menjadikan platform digital ini sebagai sarana utama untuk berinteraksi sehari-hari [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 (*hate speech*) secara daring [4]. Ujaran kebencian sendiri didefinisikan sebagai segala bentuk ekspresi yang menyebarkan, menghasut, mempromosikan, atau membenarkan kebencian berbasis ras, maupun intoleransi lainnya [5]. Dampak dari fenomena ini bersifat multidimensi, selain menurunkan kesehatan mental dan memicu gangguan kecemasan pada tingkat individu [6, 7], penyebaran narasi destruktif tersebut terbukti secara empiris berkorelasi dengan eskalasi konflik sosial di dunia nyata [8]. Pada akhirnya, besarnya volume interaksi tersebut turut memperbesar peluang penyebaran konten berbahaya, termasuk komentar toksik dan ujaran kebencian, yang penanganannya menjadi tantangan tersendiri pada teks media sosial berbahasa Indonesia [9].

Penetrasi media sosial yang tinggi ini tidak sejalan dengan kualitas keberadaban digital masyarakatnya. Laporan *Digital Civility Index* (DCI) oleh Microsoft menempatkan Indonesia pada peringkat 29 dari 32 negara dengan skor 76, yang menunjukkan tingginya risiko paparan konten negatif di ruang digital [10, 11]. Hal ini diperburuk oleh meningkatnya ujaran kebencian dalam beberapa tahun terakhir, terutama selama periode kontestasi politik. Temuan Aliansi Jurnalis

Independen (AJI) tahun 2024 menunjukkan bahwa 36,35% konten di platform X (sebelumnya Twitter) selama masa kampanye pemilu mengandung narasi kebencian yang merujuk terhadap kelompok rentan dan minoritas berbasis agama dan etnis [12]. Tingginya volume dan cepatnya penyebaran konten tersebut menunjukkan bahwa pendekatan pengawasan konvensional tidak lagi memadai.

Berbagai pendekatan telah dikembangkan untuk mendeteksi ujaran kebencian berbahasa Indonesia secara otomatis pada data platform X. Metode *machine learning* tradisional seperti *Support Vector Machine* (SVM), *Naive Bayes*, *Decision Tree*, dan *Logistic Regression* telah dilakukan, dengan hasil terbaik berupa *F1-score* makro sebesar 77% menggunakan *Logistic Regression* [9]. Pendekatan serupa dengan tambahan *XGBoost* dan *Random Forest* pada dataset yang sama menghasilkan akurasi tertinggi sebesar 83%, namun model dengan akurasi tinggi tersebut belum tentu menghasilkan penjelasan prediksi yang logis secara linguistik [13]. Keterbatasan metode tradisional muncul karena representasi fitur berbasis TF-IDF tidak mampu menangkap konteks makna kata secara menyeluruh, terutama pada kalimat yang mengandung sarkasme, makna tersirat, atau penggunaan *slang* khas Twitter.

Untuk mengatasi keterbatasan tersebut, pendekatan *deep learning* berbasis *transformer* mulai digunakan secara luas [14]. IndoBERT merupakan model bahasa berbasis BERT yang dilatih khusus pada korpus Indo4B berisi sekitar 4 miliar kata dari sumber domain umum seperti berita, Wikipedia, dan situs web berbahasa Indonesia [15]. IndoBERTweet dikembangkan dengan arsitektur BERT yang sama, tetapi dilatih menggunakan korpus lebih dari 409 juta token yang seluruhnya bersumber dari tweet berbahasa Indonesia [16]. Kedua model memiliki arsitektur yang identik, sehingga perbedaan utama keduanya terletak pada domain korpus pra-latihnya [15, 16]. IndoBERTweet dikombinasikan dengan BiLSTM mampu meningkatkan performa secara signifikan dibandingkan pendekatan tradisional pada dataset yang sama [17]. Namun demikian, penelitian-penelitian tersebut hanya berfokus pada peningkatan performa klasifikasi tanpa menyertakan mekanisme penjelasan atas keputusan yang diambil oleh model [17–19].

Keunggulan model pra-latih domain spesifik atas model domain umum pada data yang sesuai domainnya telah ditunjukkan pada penelitian sebelumnya [16, 20], sehingga perbandingan performa klasifikasi saja belum cukup untuk memberikan pemahaman baru. Persoalan yang belum banyak dikaji adalah apakah keunggulan tersebut juga disertai dasar pengambilan keputusan yang relevan secara linguistik, mengingat model dengan akurasi tinggi belum tentu menghasilkan

penjelasan prediksi yang logis [13]. Untuk menjawab persoalan tersebut, penelitian ini membandingkan penjelasan prediksi IndoBERT dan IndoBERTweet terhadap masukan yang sama [13, 21]. Dengan arsitektur yang identik tetapi korpus pra-latih yang berbeda, perbedaan maupun kesamaan pola keputusan kedua model dapat memberikan indikasi mengenai pengaruh representasi yang dipelajari selama tahap pra-pelatihan.

Namun, pemeriksaan dasar pengambilan keputusan tersebut tidak dapat dilakukan hanya dari hasil prediksi akhir, karena model berbasis *transformer* bersifat *black box* sehingga tidak dapat menjelaskan alasan di balik setiap prediksi yang dihasilkan [22]. Model kompleks yang bersifat *blackbox* menuntut penggunaan teknik *explainability* untuk memahami proses pengambilan keputusan [23]. Ketiadaan transparansi model juga dapat mengurangi kepercayaan pengguna terhadap alasan mengapa suatu konten ditandai sebagai ujaran kebencian [24]. Oleh karena itu, *Local Interpretable Model-Agnostic Explanations* (LIME) digunakan untuk mengevaluasi kualitas keputusan model [13]. Saputra dan Sibaroni [19] juga secara eksplisit merekomendasikan penerapan LIME pada model berbasis *transformer* sebagai arah penelitian selanjutnya.

Oleh karena itu, penelitian ini mengimplementasikan model IndoBERT dan IndoBERTweet untuk deteksi ujaran kebencian berbahasa Indonesia pada platform X. Setelah diperoleh konfigurasi terbaik dari masing-masing model, LIME diterapkan sebagai teknik *Explainable AI* untuk memberikan interpretasi terhadap prediksi yang dihasilkan serta menganalisis kata yang memengaruhi keputusan model. Dengan demikian, perbandingan pada penelitian ini diarahkan untuk memahami pengaruh domain korpus pra-latih, baik terhadap performa klasifikasi maupun terhadap dasar pengambilan keputusan model dalam mendeteksi ujaran kebencian pada teks berbahasa Indonesia.

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 berbahasa Indonesia di platform X serta penerapan LIME untuk menginterpretasi prediksi model?

2. Berapa tingkat performa model IndoBERT dan IndoBERTweet berdasarkan

metrik akurasi, *precision*, *recall*, dan *F1-score*, serta bagaimana hasil interpretasi prediksi yang dihasilkan melalui LIME?

1.3 Batasan Masalah

1. Penelitian ini hanya berfokus pada deteksi ujaran kebencian dalam bahasa Indonesia, khususnya yang ditemukan di platform media sosial X (sebelumnya Twitter). Bahasa informal dan penggunaan slang dalam teks X menjadi fokus utama dalam penelitian ini.
2. Penelitian ini hanya membandingkan dua model bahasa berbasis *transformer*, yaitu IndoBERT dan IndoBERTweet, yang memiliki arsitektur BERT-Base identik dan berbeda pada domain korpus pra-latihnya. Tidak dilakukan perbandingan dengan model lain di luar kedua model tersebut.
3. Penelitian ini hanya berfokus pada klasifikasi biner ujaran kebencian (*hate speech* versus *non-hate speech*) dan tidak mencakup klasifikasi kategori ujaran kebencian seperti target maupun tingkat keparahan.
4. Analisis *Explainable AI* pada penelitian ini menggunakan metode LIME yang diterapkan pada konfigurasi terbaik dari masing-masing model untuk menginterpretasikan prediksi yang dihasilkan.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, tujuan penelitian adalah sebagai berikut.

1. Mengimplementasikan model IndoBERT dan IndoBERTweet dalam deteksi ujaran kebencian berbahasa Indonesia di platform X serta menerapkan LIME untuk interpretasi prediksi model.
2. Mengukur performa model IndoBERT dan IndoBERTweet berdasarkan metrik akurasi, *precision*, *recall*, dan *F1-score*, serta menganalisis interpretasi prediksi yang dihasilkan menggunakan LIME.

1.5 Manfaat Penelitian

Manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut.

1. Memberikan bukti empiris mengenai pengaruh domain korpus pra-latih terhadap performa model IndoBERT dan IndoBERTweet dalam mendeteksi ujaran kebencian berbahasa Indonesia.
2. Memberikan interpretasi terhadap keputusan model melalui penerapan LIME sehingga faktor-faktor yang memengaruhi prediksi ujaran kebencian dapat dipahami dengan lebih baik.
3. Memberikan kontribusi dalam pengembangan sistem deteksi ujaran kebencian berbahasa Indonesia yang tidak hanya memiliki performa yang baik, tetapi juga lebih transparan dan mudah diinterpretasikan.

1.6 Sistematika Penulisan

Sistematika penulisan Skripsi ini adalah sebagai berikut.

1. Bab 1 PENDAHULUAN
   Berisi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, dan 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 penerapan LIME.
4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil eksperimen serta analisis perbandingan performa antara model IndoBERT dan IndoBERTweet berikut penjelasan LIME dari masing-masing model.
5. Bab 5 SIMPULAN DAN SARAN
   Bab ini berisi simpulan dari hasil penelitian yang menjawab rumusan masalah secara langsung, serta saran untuk penelitian lebih lanjut.