BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Ujaran kebencian (*hate speech*) adalah bentuk komunikasi yang mengekspresikan kebencian, diskriminasi, atau permusuhan terhadap individu maupun kelompok berdasarkan atribut identitas tertentu, seperti ras, agama, jenis kelamin, atau gender [1, 2]. Ekspresi tersebut tidak hanya disampaikan melalui bahasa lisan atau tulisan, tetapi dapat muncul dalam bentuk simbol, gambar, dan tindakan yang berpotensi menyinggung atau merugikan kelompok tertentu [3]. Dampaknya dapat berbeda bergantung pada kelompok yang menjadi sasaran, bentuk ekspresi, dan tingkat kerugian yang ditimbulkan [4]. Selain itu, setiap negara memiliki regulasi yang berbeda dalam menangani ujaran kebencian, termasuk ketentuan mengenai hasutan terhadap kekerasan atau genosida yang umumnya berfokus pada pencegahan dampak sosial yang lebih luas, sementara aturan yang lebih umum cenderung menitikberatkan pada pencegahan diskriminasi dan berbagai bentuk kerugian lainnya [5].

Pesatnya perkembangan media sosial telah mendorong perubahan dalam pola komunikasi masyarakat, khususnya dalam proses interaksi dan penyampaian pendapat. Berbagai platform digital seperti X (*Twitter*), *Instagram*, *Facebook*, dan *YouTube* memungkinkan setiap individu untuk menyampaikan opini secara bebas dan cepat [6]. Namun, kebebasan ini juga memunculkan berbagai permasalahan serius, salah satunya adalah meningkatnya penyebaran ujaran kebencian. Melalui fitur seperti *retweet*, *hashtag*, dan *sharing* dapat memperluas jangkauan unggahan negatif dalam waktu singkat sehingga moderasi secara manual menjadi semakin sulit dilakukan.

Indonesia memiliki aturan yang mengatur tentang ujaran kebencian, salah satunya melalui Pasal 28 ayat (2) dari Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) yang disaahkan pada Tahun 2024. Ketentuan dan ancaman sanksinya perlu dicantumkan sesuai versi regulasi terbaru serta didukung rujukan resmi [7]. Salah satu upaya untuk mengurangi penyebaran konten tersebut adalah melakukan identifikasi secara dini. Karena pengguna media sosial menghasilkan unggahan dalam jumlah besar dan berlangsung hampir secara waktu nyata, penyaringan manual saja tidak memadai [7]. Oleh sebab itu, penerapan sistem

otomatis menjadi alternatif yang relevan untuk mengelola dan menganalisis volume konten yang besar dalam waktu yang relatif singkat.

Instansi pemerintah seperti Kementerian Komunikasi dan Digital Republik Indonesia serta Kepolisian Negara Republik Indonesia dapat memanfaatkan sistem otomatis untuk melakukan pengawasan terhadap aktivitas pada platform digital. Selain itu, penyedia layanan media sosial juga dapat mengimplementasikan sistem serupa guna mendukung proses moderasi konten. Tingginya jumlah konten yang berpotensi mengandung ujaran kebencian, disertai keterbatasan sumber daya dalam moderasi manual, menjadikan pengembangan sistem deteksi ujaran kebencian sebagai solusi yang penting [8]. Salah satu metode yang sering diterapkan untuk membantu proses deteksi tersebut adalah *machine learning* atau pembelajaran mesin, yang mampu melakukan klasifikasi secara efektif dan dapat diterapkan pada skala data yang besar [8].

Pembelajaran mesin sekarang sudah menjadi instrumen yang krusial untuk mengidentifikasi ujaran kebencian di platform media sosial [9]. Kajian terbaru mengeksplorasi berbagai algoritma, strategi rekayasa fitur, serta pengaturan untuk bahasa tertentu guna meningkatkan akurasi dalam deteksi dan moderasi yang beretika [10]. Beberapa model yang umum diterapkan meliputi *K-Nearest Neighbors, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machines*, dan *Random Forest* [11, 12]. Dengan meningkatnya keperluan untuk memahami pola bahasa yang lebih kompleks dan terperinci, pendekatan yang memanfaatkan *deep neural networks* dan *transfer learning* mulai memperlihatkan hasil klasifikasi yang lebih baik dibandingkan dengan pendekatan konvensional [13]. Model yang berlandaskan pada IndoBERT seperti *BERT* dan *RoBERTa* telah terbukti menunjukkan kemampuan luar biasa dalam menangkap konteks bahasa serta memberikan hasil yang sangat baik dalam pengelompokan teks [14, 15]. Namun demikian, model IndoBERT memiliki keterbatasan dalam menangkap fitur lokal seperti pola kata atau ungkapan tertentu yang biasa timbul dalam ujaran kebencian.

Di sisi lainnya, CNN diakui sangat efektif dalam mengambil fitur lokal dari teks dan mampu mengenali pola-pola linguistik secara spesifik [16]. Sering kali, mereka menunjukkan performa yang unggul dibandingkan dengan pendekatan konvensional, khususnya dalam menangani ujaran kebencian yang lebih kompleks atau halus [16, 17]. CNN merupakan jenis algoritma *deep learning* yang khusus untuk meneliti data yang memiliki format menyerupai kisi, contohnya gambar. CNN bisa digunakan untuk berbagai macam tugas, seperti mengategorikan gambar,

mendeteksi objek, dan mengenali suara berkat kemampuannya dalam mempelajari dan mengekstrak secara otomatis fitur penting dari data asal [18, 19]. Oleh karena itu, integrasi antara IndoBERT dan *CNN* dalam bentuk model *Hybrid* IndoBERT CNN menjadi pendekatan yang menjanjikan, karena mampu menggabungkan keunggulan pemahaman konteks global dan ekstraksi fitur lokal secara bersamaan.

Meskipun model *deep learning* mampu memberikan performa yang tinggi, sebagian besar model tersebut bersifat *black box* dan sulit untuk diinterpretasikan [20]. Hal ini menjadi tantangan dalam implementasi sistem deteksi ujaran kebencian, terutama pada konteks moderasi konten yang membutuhkan transparansi dan kejelasan alasan di balik suatu keputusan. Oleh sebab itu, diperlukan pendekatan *Explainable Generative AI* untuk meningkatkan transparansi dan interpretabilitas proses pengambilan keputusan model.

Salah satu pendekatan yang banyak digunakan untuk meningkatkan interpretabilitas model adalah *Local Interpretable Model-Agnostic Explanations* (LIME). Metode ini mampu mengidentifikasi fitur-fitur yang memberikan kontribusi terhadap hasil prediksi model. Informasi tersebut selanjutnya dapat dimanfaatkan dalam pendekatan *Explainable Generative AI* untuk menghasilkan penjelasan naratif yang lebih mudah dipahami oleh manusia [21]. Dengan demikian, hasil klasifikasi tidak hanya menunjukkan kategori ujaran kebencian maupun non-ujaran kebencian, tetapi juga disertai penjelasan mengenai faktor-faktor yang memengaruhi keputusan model.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pendekatan Integrasi *Hybrid* IndoBERT CNN dan *Explainable Generative AI* untuk Deteksi serta Interpretasi Ujaran Kebencian pada Platform X. Pendekatan ini diharapkan mampu menciptakan sistem yang mampu mempertahankan akurasi klasifikasi sekaligus meningkatkan transparansi dalam menjelaskan dasar pembentukan keputusan model.

1.2 Rumusan Masalah

Berdasarkan hasil identifikasi permasalahan yang telah diuraikan pada latar belakang, rumusan masalah dalam penelitian ini dapat dinyatakan sebagai berikut:

1. Bagaimana merancang model *Hybrid* IndoBERT CNN untuk deteksi ujaran kebencian?

2. Bagaimana kinerja model *Hybrid* IndoBERT CNN dibandingkan dengan

model CNN dan IndoBERT dalam deteksi ujaran kebencian pada platform X?

3. Bagaimana mengimplementasikan *Explainable Generative AI* untuk menghasilkan penjelasan naratif terhadap hasil klasifikasi ujaran kebencian?

1.3 Batasan Permasalahan

Batasan dalam penelitian ini adalah:

1. Data yang digunakan dalam penelitian ini terbatas pada *dataset* atau kumpulan komentar dari*Twitter/X* yang tersedia secara publik melalui repositori *GitHub* milik Muhammad Okky Ibrohim [22]. *Dataset* tersebut dipublikasikan beserta pedoman anotasi yang digunakan dalam proses pelabelannya. Sebelum dipublikasikan, data telah melalui proses kurasi dan anotasi yang mendalam, termasuk verifikasi oleh lebih dari satu penilai berdasarkan panduan anotasi yang sistematis serta tahapan validasi bertingkat untuk memastikan konsistensi dan keandalan label yang dihasilkan. *Dataset* ini telah banyak dipakai dalam berbagai penelitian sebelumnya yang berfokus pada deteksi ujaran kebencian, menunjukkan tingkat kualitas yang cukup sehingga layak digunakan sebagai sumber data dalam penelitian ini.
2. Klasifikasi pada dataset dibagi menjadi positif dan negatif.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Merancang model *Hybrid* IndoBERT CNN untuk deteksi ujaran kebencian.
2. Membandingkan performa model CNN, IndoBERT, dan *Hybrid* IndoBERT CNN menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.
3. Mengimplementasikan *Explainable Generative AI* untuk menghasilkan penjelasan naratif terhadap hasil klasifikasi.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini meliputi:

1. Penelitian ini memberikan kontribusi terhadap pengembangan ilmu pengetahuan di bidang *Natural Language Processing* (NLP), khususnya dalam pemanfaatan model *Hybrid* IndoBERT CNN untuk meningkatkan efektivitas klasifikasi teks. Selain itu, penelitian ini juga memperkaya kajian terkait *Explainable Generative AI* untuk interpretasi hasil model.
2. Penelitian ini mendukung pengembangan sistem moderasi konten pada platform media sosial melalui penerapan deteksi ujaran kebencian secara otomatis. Integrasi *Explainable Generative AI* memungkinkan sistem menyajikan penjelasan naratif mengenai alasan di balik setiap prediksi sehingga dapat membantu pengelola platform maupun lembaga terkait dalam memahami dasar pengambilan keputusan model.
3. Penelitian ini menghasilkan model integrasi *Hybrid* IndoBERT CNN yang menggabungkan kemampuan IndoBERT dalam memahami konteks bahasa dengan kemampuan CNN dalam mengekstraksi fitur lokal, serta mengintegrasikannya dengan *Explainable Generative AI*. Integrasi tersebut menghasilkan sistem yang tidak hanya melakukan klasifikasi ujaran kebencian, tetapi juga menyajikan interpretasi hasil prediksi secara naratif.
4. Penelitian ini menyediakan penjelasan naratif berdasarkan hasil interpretasi LIME sehingga alasan di balik prediksi model dapat dipahami oleh pengguna non-teknis. Dengan demikian, sistem tidak hanya memberikan hasil klasifikasi, tetapi juga memberikan informasi yang mendukung pemahaman terhadap dasar prediksi model dan mendukung literasi digital dalam memahami karakteristik konten yang terindikasi sebagai ujaran kebencian.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini disusun ke dalam lima bab sebagai berikut.

1. Bab 1 PENDAHULUAN
   Bab ini berisi latar belakang masalah, rumusan masalah, batasan permasalahan, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
2. Bab 2 LANDASAN TEORI
   Bab ini berisi landasan teori yang mendukung penelitian, meliputi konsep

ujaran kebencian (*hate speech*), Platform X, *Natural Language Processing* (NLP), *data preprocessing*, *deep learning*, IndoBERT, *Convolutional Neural Network* (CNN), *Hybrid* IndoBERT CNN, *Explainable Generative AI*, metrik evaluasi, dataset yang digunakan, serta penelitian terdahulu yang menjadi dasar pengembangan penelitian.

3. Bab 3 METODOLOGI PENELITIAN
   Bab ini menjelaskan metodologi penelitian yang meliputi studi literatur, pengumpulan data, pemrosesan data, perancangan model *Hybrid* IndoBERT CNN, optimasi hiperparameter menggunakan Optuna, serta metode evaluasi model.

4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil implementasi dan pembahasan penelitian yang meliputi spesifikasi sistem, deskripsi dataset, tahap *preprocessing*, perancangan model, proses pelatihan dan optimasi hiperparameter, implementasi LIME dan *Explainable Generative AI*, serta hasil pengujian dan evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*.

5. Bab 5 KESIMPULAN DAN SARAN
   Bab ini berisi simpulan yang diperoleh berdasarkan hasil penelitian serta saran yang dapat dijadikan acuan untuk pengembangan penelitian selanjutnya.