BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Akselerasi teknologi informasi membawa pergeseran signifikan dalam fungsi media sosial, yang kini tidak lagi sekadar berperan sebagai platform komunikasi, melainkan telah berkembang menjadi ekosistem informasi global yang turut membentuk wacana publik dan dinamika sosial masyarakat modern. Berdasarkan laporan Statista menunjukkan bahwa negara Indonesia mengalami peningkatan signifikan dalam penggunaan media sosial, dengan lebih dari 217,53 juta pengguna aktif pada tahun 2022. Hal ini menjadikan Indonesia sebagai negara pengguna media sosial terbesar keempat di dunia, setelah negara Tiongkok, India, dan Amerika Serikat [1]. Selain itu, laporan We Are Social dan Kepios pada tahun 2024 menunjukkan bahwa pengguna media sosial secara global mencapai 5,04 miliar pada Januari 2024, setara dengan 62,3% dari total populasi dunia, dengan pertumbuhan 266 juta pengguna baru dalam satu tahun terakhir [2]. Di Indonesia, dari total populasi sekitar 272 juta jiwa, lebih dari setengahnya atau lebih kurang 160 juta orang telah menjadikan media sosial sebagai sarana utama untuk berinteraksi [3]. Namun, pertumbuhan masif ini juga diiringi oleh peningkatan penyebaran konten berbahaya di ruang digital. Studi oleh Joseph B. Walther menunjukkan bahwa 42-67% pemuda mengawasi ujaran kebencian secara daring dan 21% menjadi korban atas tulisan atau ucapan yang penuh kebencian dan merendahkan di internet [4]. Fenomena ini semakin mengkhawatirkan di Indonesia, di mana laporan Kementerian Komunikasi dan Informatika mencatat hampir 1,4 juta konten negatif di media sosial ditemukan, dengan ujaran kebencian menjadi 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]. Dalam konteks digital, ujaran kebencian tidak hanya terbatas pada serangan verbal eksplisit, tetapi juga mencakup bentuk komunikasi implisit yang mengandung stereotip negatif, dehumanisasi, atau marginalisasi kelompok tertentu [6]. Dampak ujaran kebencian sangat luas dan multidimensi, mulai dari tingkat individu yang melibatkan penurunan kesehatan mental, kecemasan, depresi, gangguan stres pasca-

trauma (PTSD), dan penurunan harga diri, hingga masyarakat yang dapat memicu polarisasi kelompok, memperdalam divisi sosial, menormalisasi diskriminasi, dan bahkan meningkatkan risiko kekerasan fisik terhadap kelompok yang menjadi sasaran [7]. Bukti empiris menunjukkan bahwa ujaran kebencian di media sosial sering mendahului eskalasi kekerasan di dunia nyata, seperti yang terjadi dalam krisis Rohingya di Myanmar, di mana konten ujaran kebencian di platform digital berkontribusi pada kekerasan dan pembersihan etnis terhadap minoritas [8]. Dengan kata lain, ujaran kebencian di media sosial tidak hanya memicu konflik sosial, tetapi juga mampu membentuk opini publik serta memperkuat persepsi dan sentimen negatif terhadap individu maupun kelompok tertentu.

Karakteristik ujaran kebencian pada media sosial Indonesia memiliki keunikan tersendiri yang membuatnya lebih kompleks untuk diidentifikasi dan ditangani secara otomatis. Tidak seperti ujaran kebencian konvensional yang disampaikan secara eksplisit dan langsung, 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, dan emoji sebagai pengganti kata, serta pencampuran bahasa Indonesia dengan bahasa Inggris atau bahasa daerah, yang dikenal dengan *code-mixing* [9–11]. Fenomena *code-mixing* bahasa Indonesia dan bahasa Inggris sangat umum ditemukan di platform media sosial, di mana pengguna secara fleksibel mengombinasikan kosakata, frasa, atau klausa dari kedua bahasa dalam satu teks atau percakapan. Penelitian menunjukkan bahwa penggunaan media sosial secara signifikan mendorong praktik penggunaan teks *code-mixing* Indonesia–Inggris, khususnya di kalangan Generasi Z, yang berfungsi untuk menekankan pesan, mengekspresikan identitas, meningkatkan keterlibatan audiens, serta membangun relasi dan menyampaikan informasi secara lebih menarik melalui beragam bentuk kata, frasa, pengulangan, idiom, dan klausa [12, 13].

Menghadapi kompleksitas dan volume ujaran kebencian yang terus meningkat di media sosial, pendekatan otomatis berbasis *machine learning* dan *deep learning* telah menjadi solusi dominan yang diadopsi oleh peneliti dan praktisi. Kemajuan pesat dalam *Natural Language Processing*, khususnya melalui arsitektur *Transformer*, secara signifikan meningkatkan akurasi sistem deteksi ujaran kebencian dan berkontribusi terhadap upaya menjaga ruang digital yang aman dan inklusif [5]. Tinjauan komprehensif terhadap lebih dari 100 studi menunjukkan evolusi metode dari pendekatan klasik seperti *Support Vector Machine* (SVM) hingga model neural seperti *Convolutional Neural Network*

(CNN) ataupun *Long Short-Term Memory* (LSTM) [14, 15]. Dalam konteks multilingual, model berbasis BERT, termasuk multilingual BERT dan XLM-RoBERTa, menunjukkan performa kompetitif sekitar 90% nilai *F1-score* pada skenario lintas bahasa dan teks *code-mixing*, khususnya pada bahasa Spanyol dan bahasa Inggris [16–20]. Sejumlah studi juga melaporkan capaian empiris yang kuat, seperti *F1-score* 94% pada LSTM berbasis GloVe untuk teks Indonesia, kemampuan generalisasi *zero-shot* dan *few-shot* pada korpus Spanyol–Inggris, degradasi performa 4–7% pada teks *code-mixing* dengan bahasa daerah Indonesia, serta akurasi 99,97% pada model BLOOM untuk teks Indonesia–Inggris [10, 17, 20, 21]. Meskipun demikian, sebagian besar pendekatan masih bergantung pada tantangan metodologis validasi model atau ketersediaan sumber daya pada spesifik bahasa yang membatasi skalabilitas dan generalisasi lintas bahasa [22, 23].

Selain keterbatasan sumber daya, pendekatan deteksi ujaran kebencian yang ada saat ini umumnya hanya menghasilkan label klasifikasi biner tanpa mampu mengidentifikasi secara eksplisit bagian teks mana yang secara spesifik berkontribusi terhadap kandungan kebencian tersebut. Beberapa penelitian menunjukkan bahwa deteksi di tingkat *span* atau potongan teks masih sangat jarang dieksplorasi dalam literatur ujaran kebencian [14, 24]. Penelitian oleh Zhou et al. membuktikan bahwa identifikasi *span* atau token ujaran kebencian (satu kata yang merupakan bagian dari *hate speech*) secara presisi jauh lebih informatif dibandingkan klasifikasi biner, namun masih terbatas pada teks berbahasa Inggris dan belum mencakup teks *code-mixing* yang bersifat informal [25]. Selain itu, kompetisi HASOC FIRE 2023 turut menegaskan bahwa deteksi *span* merupakan celah riset yang mendesak, terutama pada teks *code-mixing* yang masih sangat minim digarap [26]. Oleh karena itu, diperlukan pendekatan terpadu yang tidak hanya mampu mendeteksi ujaran kebencian di tingkat kalimat, tetapi juga melokalisasi di tingkat token ujaran kebencian sebagai bentuk *explainability* sistem.

Maka dari itu, penelitian ini mengusulkan pendekatan dengan memanfaatkan model BLOOM (*BigScience Large Open-science Open-access Multilingual Language Model*), dengan variasi model 560 parameter (BLOOM-560M) sebagai fondasi untuk seluruh proses deteksi ujaran kebencian pada teks *code-mixing* Indonesia-Inggris. BLOOM dipilih karena merupakan model bahasa besar multilingual berbasis arsitektur *Transformer* yang dilatih pada 46 bahasa natural termasuk bahasa Indonesia dan Inggris, tersedia secara *open source*, serta memiliki kemampuan *transfer learning* yang kuat untuk bahasa dengan sumber daya rendah [27]. Penelitian ini sekaligus menjawab kesenjangan yang belum

diatasi oleh penelitian sebelumnya, di mana analisis deteksi ujaran kebencian umumnya masih terbatas pada teks yang bersifat formal atau semi-formal dan belum mencakup bahasa gaul atau kata *slang*, maupun singkatan tidak baku yang justru mendominasi percakapan di media sosial [17]. Oleh sebab itu, penggunaan model BLOOM-560M berbasis *Multi-Task Learning* (MTL) dibangun untuk menggabungkan deteksi ujaran kebencian di tingkat kalimat dan identifikasi *hate span* di tingkat token, yaitu segmen token berurutan dalam suatu kalimat yang secara langsung menjadi alasan penyebab kandungan ujaran kebencian [25, 28, 29]. Tahap identifikasi tingkat token tersebut menggunakan skema pelabelan BIO (*Begin-Inside-Outside*), yakni format anotasi urutan yang lazim digunakan dalam tugas *Named Entity Recognition* untuk menandai awal (B), kelanjutan (I), dan bagian di luar entitas target (O) dari suatu segmen yang relevan [30, 31]. Label-label BIO ini dihasilkan secara otomatis melalui TF-IDF *Differential Lexicon* yang dibangun sepenuhnya dari dataset sendiri tanpa bergantung pada sumber eksternal, di mana *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan ukuran statistik yang menilai ciri khas suatu kata pada sebuah dokumen relatif terhadap keseluruhan korpus, dan selisih skor TF-IDF antara korpus *hate* dan *non-hate* dimanfaatkan untuk mengidentifikasi kata-kata yang paling diskriminatif sebagai leksikon ofensif [5, 32–34]. Pendekatan ini memungkinkan model untuk secara bersamaan mengklasifikasikan teks pada tingkat kalimat dan melokalisasi token yang berkontribusi terhadap kandungan kebencian, melalui pelatihan model BLOOM-560M berbasis MTL yang dikombinasikan dengan *hyperparameter tuning* berbasis Optuna dan terbukti secara signifikan meningkatkan akurasi, presisi, *recall*, dan *F1-score* model [35, 36].

Dengan kata lain sebagai kesimpulan, penelitian ini hadir sebagai upaya mengisi celah yang selama ini belum terjawab secara sistematis dan belum tersedia pada penelitian sebelumnya, yaitu perluasan cakupan teks dengan unsur *slang* atau bahasa gaul dan singkatan tidak baku, serta sistem deteksi yang tidak sekadar mengklasifikasikan ujaran kebencian pada tingkat kalimat, melainkan juga mampu mengidentifikasi *hate span* pada tingkat token yang menjadi penyebab ujaran kebencian sebagai bentuk *explainability* pada teks *code-mixing* bahasa Indonesia-Inggris yang bersifat informal. Maka dari itu, penelitian ini diharapkan dapat memberikan kontribusi nyata pada pengembangan sistem deteksi ujaran kebencian yang lebih adaptif dan transparan, sekaligus membuka jalan bagi penelitian lanjutan dalam penanganan konten berbahaya pada konteks bahasa *code-mixing* yang masih relatif minim dieksplorasi secara menyeluruh dalam beragam kombinasi bahasa.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah diutarakan pada latar belakang, rumusan masalah dalam penelitian ini dirumuskan sebagai berikut.

1. Bagaimana performa model BLOOM-560M yang dilatih dalam kerangka *Multi-Task Learning* dalam tingkat kalimat untuk mendeteksi ujaran kebencian pada teks *code-mixing* bahasa Indonesia dan bahasa Inggris yang mencakup kata *slang* atau bahasa gaul dan singkatan tidak baku berdasarkan metrik evaluasi akurasi dan *F1-score*?
2. Bagaimana performa model BLOOM-560M yang dilatih dalam kerangka *Multi-Task Learning* dalam tingkat token untuk mengidentifikasi *hate span* pada teks *code-mixing* bahasa Indonesia dan bahasa Inggris melalui skema pelabelan BIO otomatis berbasis TF-IDF *Differential Lexicon* berdasarkan metrik evaluasi *Span-F1*?

1.3 Batasan Permasalahan

Dalam upaya mencapai hasil terukur dan dapat dipertanggungjawabkan secara akademik, penelitian ini menetapkan batasan permasalahan sebagai berikut.

1. Penelitian ini dibatasi pada deteksi dan identifikasi *span* ujaran kebencian dalam teks *code-mixing* bahasa Indonesia dan bahasa Inggris, yang mencakup kata *slang* dan singkatan tidak baku. Namun tidak mencakup emoji, bahasa daerah (seperti Bahasa Jawa, Sunda, Batak, dan variannya), maupun konten berbahaya lainnya seperti misinformasi atau konten seksual eksplisit yang berada di luar cakupan ujaran kebencian. Sistem yang dikembangkan menghasilkan klasifikasi biner antara *hate speech* dan *non-hate speech* tanpa membedakan kategori atau target spesifik seperti ras, agama, gender, etnisitas, atau orientasi seksual, sebagai konsekuensi langsung dari skema label biner pada kedua dataset sumber yang digunakan.
2. Dataset yang digunakan terdiri dari 40.000 sampel dengan komposisi seimbang antara kelas *hate speech* dan *non-hate speech* yang terdiri atas dua komponen, yaitu data bahasa Indonesia dan bahasa Inggris yang bersumber dari platform Hugging Face, serta data *code-mixed* Indonesia-Inggris yang dibangkitkan secara sintetis menggunakan LLM dengan model Claude

Sonnet 4.6 karena keterbatasan sumber data berlabel publik untuk ragam bahasa ini. Seluruh proses pembersihan dan pembagian dataset dilakukan secara mandiri tanpa memanfaatkan dataset ataupun kamus eksternal.

3. Proses pelabelan BIO untuk deteksi tingkat token dilakukan secara otomatis menggunakan TF-IDF *Differential Lexicon* yang dibangun sepenuhnya dari dataset sendiri dengan penetapan ambang batas persentil serta pemakaian *unigram* dan *bigram* untuk tiap ragam bahasa, sehingga pendekatan ini tergolong dalam *weak supervision* yang valid secara akademis dan konsisten dengan komitmen penelitian untuk tidak bergantung pada sumber daya pihak ketiga. Pendekatan *Differential Lexicon* menyeleksi *term* berdasarkan asosiasi frekuensi statistik dengan label *hate*, bukan muatan semantiknya, sehingga sebagian *term* terpilih tidak selalu ofensif secara leksikal. Risiko ini dimitigasi dengan membatasi penerapan *lexicon* hanya pada dokumen berlabel *hate*, sehingga tidak mencemari label token pada dokumen *non-hate* meskipun berpotensi *noise* pada label *span* di dalam dokumen *hate* tetap menjadi batasan inheren pendekatan *weak supervision* ini. Mekanisme pelabelan token juga murni berbasis kecocokan *string* terhadap *lexicon* tanpa mempertimbangkan makna kontekstual, kecuali filter negasi sederhana pada kata sebelumnya, sehingga ambiguitas seperti sarkasme tidak tertangani.

1.4 Tujuan Penelitian

Berlandaskan pada rumusan masalah yang telah dipaparkan sebelumnya, penelitian ini memiliki tujuan yang hendak dicapai dan menjawab tiap rumusan masalah secara berurutan sebagai berikut.

1. Menerapkan dan mengevaluasi performa model BLOOM-560M yang dilatih dalam kerangka *Multi-Task Learning* dalam tingkat kalimat untuk mendeteksi ujaran kebencian pada teks *code-mixing* bahasa Indonesia dan bahasa Inggris yang mencakup kata *slang* atau bahasa gaul dan singkatan tidak baku berdasarkan metrik evaluasi akurasi dan *F1-score*.
2. Menerapkan dan mengevaluasi performa model BLOOM-560M yang dilatih dalam kerangka *Multi-Task Learning* dalam tingkat token untuk mengidentifikasi *hate span* pada teks *code-mixing* bahasa Indonesia dan bahasa Inggris melalui skema pelabelan BIO otomatis berbasis TF-IDF *Differential Lexicon* berdasarkan metrik evaluasi *Span-F1*.

1.5 Manfaat Penelitian

Dengan tercapainya tujuan penelitian yang telah ditetapkan, diharapkan penelitian ini mampu memberikan implikasi positif bagi berbagai pihak yang berkaitan. Manfaat yang diproyeksikan dari penelitian ini adalah sebagai berikut.

1. Memberikan kontribusi ilmiah berupa *pipeline* dua tahap yang terintegrasi dan belum pernah dieksplorasi sebelumnya pada konteks *code-mixing* bahasa Indonesia dan bahasa Inggris yang bersifat informal, yang mencakup deteksi ujaran kebencian tingkat kalimat dan identifikasi *hate span* tingkat token dalam pelatihan *Multi-Task Learning* (MTL) pada model BLOOM-560M, beserta optimasi *hyperparameter* menggunakan Optuna untuk skenario MTL tersebut, sehingga kontribusi ini dapat menjadi landasan metodologis bagi penelitian selanjutnya di bidang moderasi konten multilingual maupun optimasi sistem MTL pada konteks bahasa *code-mixing*.

2. Menyediakan dataset ujaran kebencian Indonesia-Inggris yang mencakup variasi linguistik informal secara komprehensif, terdiri atas gabungan data autentik dari Hugging Face serta komponen *code-mixed* yang dibangkitkan secara sintetis menggunakan LLM dengan model Claude Sonnet 4.6, yang secara keseluruhan dapat menjadi sumber daya terbuka bagi komunitas riset *Natural Language Processing* untuk pengembangan model deteksi dan moderasi konten pada ragam bahasa campuran.

3. Menawarkan pendekatan *weak supervision* berbasis TF-IDF *Differential Lexicon* yang dibangun per ragam bahasa tanpa ketergantungan pada kamus eksternal atau anotasi manual untuk menghasilkan label BIO otomatis pada dataset ujaran kebencian di berbagai konteks bahasa campuran yang kekurangan sumber daya anotasi di tingkat token.

4. Memberikan kontribusi praktis bagi pemangku kebijakan, platform digital, dan penyelenggara moderasi konten di Indonesia dengan menghadirkan kerangka sistem deteksi ujaran kebencian yang transparan, di mana setiap keputusan klasifikasi disertai identifikasi token penyebab (*hate span*) yang dapat diverifikasi secara eksplisit, sehingga mendukung pengembangan mekanisme moderasi konten nasional yang lebih akuntabel dan konsisten dengan prinsip transparansi algoritmik dalam tata kelola platform digital.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini disusun untuk memberikan gambaran yang terstruktur mengenai alur pembahasan penelitian dari awal hingga akhir. Adapun susunan penulisan skripsi ini terdiri dari lima bab antara lain sebagai berikut.

1. Bab 1 PENDAHULUAN
   Bab ini memuat uraian mengenai latar belakang masalah, rumusan masalah, batasan, tujuan, serta manfaat yang diharapkan dari penelitian ini. Pada bab ini juga dijelaskan gambaran umum permasalahan yang melandasi penelitian dan disertai data pendukung, sehingga dapat memberikan pemahaman awal mengenai konteks dan arah penelitian yang dilakukan.

2. Bab 2 LANDASAN TEORI
   Bab ini menyajikan kajian pustaka dan dasar teoritis yang relevan dengan topik penelitian. Pembahasannya mencakup teori, konsep, algoritma, serta arsitektur model yang digunakan, yang diperoleh dari berbagai sumber ilmiah seperti buku dan jurnal bereputasi. Bab ini bertujuan sebagai landasan konseptual dalam mendukung metode dan analisis penelitian.

3. Bab 3 METODOLOGI PENELITIAN
   Bab ini menjelaskan tahapan dan pendekatan penelitian yang digunakan secara sistematis. Uraian pada bab ini meliputi alur penelitian, metode pengumpulan dan pengolahan data, perancangan sistem atau model, serta tahapan implementasi yang dilakukan untuk mencapai tujuan penelitian.

4. Bab 4 HASIL DAN DISKUSI
   Bab ini menyajikan hasil penelitian yang diperoleh dari penerapan metodologi yang telah dirancang. Bab ini membahas hasil pengujian, evaluasi kinerja sistem atau model, serta analisis terhadap temuan penelitian. Pembahasan dilakukan untuk menunjukkan keterkaitan antara hasil yang diperoleh dengan rumusan masalah dan tujuan penelitian.

5. Bab 5 SIMPULAN DAN SARAN
   Bab ini berisi kesimpulan yang dirangkum berdasarkan hasil dan pembahasan penelitian. Selain itu, bab ini juga menguraikan keterbatasan penelitian serta saran yang dapat dijadikan acuan untuk pengembangan atau penelitian selanjutnya di bidang yang relevan.