BAB I
PENDAHULUAN

1.1. Latar Belakang

Media sosial telah menjadi bagian yang tak terpisahkan dari kehidupan masyarakat modern. Hal ini memengaruhi berbagai aspek interaksi, komunikasi, dan perilaku [1]. Media sosial dikenal sebagai alat komunikasi yang efektif dan dapat mempermudah untuk berkomunikasi jarak jauh. Selain itu, media sosial dapat memberikan berbagai manfaat, seperti memperluas jaringan sosial, membangun komunitas daring, serta membuka peluang terbentuknya hubungan dan pertemanan baru [2]. Perkembangan tersebut menyebabkan penggunaan media sosial meningkat secara signifikan, terutama di Indonesia [3].

Berdasarkan *website* DataReportal, pengguna media sosial di Indonesia pada awal tahun 2025 mencapai 143 juta orang, yang setara dengan 50,2% dari jumlah penduduk Indonesia [4]. Angka ini menunjukkan peningkatan dibandingkan tahun sebelumnya, yakni pada tahun 2024 yang mencatat 139 juta pengguna [5]. Peningkatan ini mencerminkan bahwa media sosial semakin menjadi bagian penting dalam kehidupan sehari-hari masyarakat Indonesia. Namun, meningkatnya penggunaan media sosial juga menimbulkan permasalahan baru, salah satunya adalah tindakan *cyberbullying* atau perundungan di dunia maya [6]. Menurut laporan UNICEF Indonesia, terdapat 45% remaja dan dewasa muda berusia 14-24 tahun pernah mengalami tindakan *cyberbullying* [7]. Twitter merupakan salah satu sosial media yang memungkinkan penggunanya untuk memberikan opini bebas. Hal ini dapat memicu percakapan yang bersifat impulsif, bahkan tidak jarang menimbulkan komentar bernada negatif maupun ujaran kebencian yang dapat menyebabkan *cyberbullying* [8].

Cyberbullying dapat didefinisikan sebagai perilaku agresif yang dilakukan secara sengaja dan bertujuan untuk membahayakan serta menyakiti individu lain melalui perangkat elektonik [9]. Berbeda dengan konflik yang terjadi ketika dua pihak atau lebih mengalami pertikaian atas ketidaksesuaian dalam pendapat, tujuan, dan berbagai campur tangan dari pihak lain dalam mencapai suatu tujuan [10].

Tindakan *cyberbullying* dapat mencakup beberapa hal, seperti pengiriman pesan yang menyakitkan, penyebaran rumor atau kebohongan, pembagian informasi pribadi atau memalukan tanpa izin, hingga unggahan komentar atau gambar bernada merendahkan dengan tujuan menyakiti secara emosional atau sosial [11]. Bentuk-bentuk *cyberbullying* tersebut dapat berupa serangan langsung terhadap individu atau kelompok, peniruan identitas, pengucilan, maupun pembuatan akun palsu untuk mengecoh atau mengintimidasi korban [12]. Dampak dari *cyberbullying* dapat menyebabkan kerugian terhadap kesehatan fisik maupun mental pada korban [13]. Korban dapat mengalami gejala depresi, kecemasan, gangguan psikologis, dan gejala stres pasca-trauma [14]. Tidak hanya itu, korban *cyberbullying* dapat mengalami perubahan perilaku seperti agresivitas berlebihan, menarik diri dari lingkungan sosial, serta kesulitan membentuk hubungan baru [15]. Dalam kasus yang lebih ekstrem, *cyberbullying* bahkan dapat mendorong korban untuk memiliki pikiran atau perilaku bunuh diri, khususnya pada kalangan remaja [16]. Ide untuk melakukan bunuh diri pada kalangan remaja yang mengalami *cyberbullying* mencapai 38.4% dibandingkan remaja yang tidak mengalami *cyberbullying*, yaitu sebesar 16.6% [17]. Oleh karena itu, penting untuk mengembangkan sistem deteksi terhadap teks yang mengandung unsur *cyberbullying* berbahasa Indonesia yang mengandung kata-kata tidak baku sebagai bentuk pencegahan dan perlindungan bagi pengguna media sosial, khususnya pada aplikasi Twitter (X). Pada penelitian ini akan digunakan pendekatan *Natural Language Processing* (NLP) yang memungkinkan komputer untuk memproses dan memahami bahasa manusia. Berbagai teknik NLP telah digunakan secara luas pada berbagai penerapan, seperti analisis sentimen, maupun klasifikasi teks [18]. Dengan demikian, penerapan NLP dengan pendekatan klasifikasi teks ini dapat digunakan untuk mendeteksi *cyberbullying* pada sosial media X.

Beberapa penelitian terkait deteksi *cyberbullying* di Indonesia dengan menerapkan berbagai model *machine learning* telah dilakukan. Penelitian pertama dilakukan oleh [19] terkait deteksi *cyberbullying* pada Twitter menggunakan metode *Support Vector Machine* (SVM). Penelitian ini menggunakan algoritma C-SVC dan Nu-SVC dari metode SVM. Hasil dari penelitian ini menunjukkan bahwa

algoritma C-SVC lebih baik daripada Nu-SVC, dengan perbandingan 79.6% dan 70% pada tahap pengujian. Penelitian selanjutnya dilakukan oleh [20] yang bertujuan untuk mendeteksi *cyberbullying* pada komentar berbahasa Indonesia pada media sosial YouTube dengan menggunakan kombinasi model *Convolutional Neural Network* dengan *Long-Short Term Memory* (CNN-LSTM). Penelitian ini menerapkan *layer* CNN untuk melakukan ekstraksi fitur dan LSTM untuk mempelajari konteks urutan kata dalam teks. Proses *training* dilakukan dengan menggunakan percobaan beberapa kali terhadap layer CNN-LSTM, mulai dari 128, 256, 512, 1024. Hasil akurasi terbaik yang didapatkan adalah 0.9658 yang didapatkan dari mengatur 1024 *layer* CNN-LSTM. Kemudian, didapatkan hasil nilai *F1-score* dari model CNN-LSTM yaitu 0.841. Penelitian lain dilakukan oleh [21] yang mengimplementasikan model IndoBERT untuk deteksi *cyberbullying*. Pada penelitian ini, dilakukan penggabungan metode VADER Lexicon dan *Generative Large Language Model* untuk proses pemberian label pada data. Hasil penelitian ini menunjukkan bahwa model IndoBERT dengan penerapan *fine tuning* menunjukkan tingkat akurasi sebesar 96,7% dan paling unggul dibandingkan dengan model LSTM dan Bi-LSTM.

Beberapa metode yang telah disebutkan di atas, seperti SVM, CNN-LSTM, dan IndoBERT memiliki keterbatasan dalam memahami kalimat yang mengandung kata-kata yang tidak baku seperti singkatan atau *slang* (bahasa gaul) yang sering dijumpai pada platform media sosial seperti X. Dengan demikian, diperlukan sebuah model yang dapat mengatasi masalah tersebut. IndoBERTweet merupakan sebuah model bahasa yang telah dilatih menggunakan 26 juta tweet berbahasa Indonesia dengan penambahan 46% kosakata baru yang relevan dengan bahasa informal di media sosial, sehingga mampu memahami berbagai penulisan bahasa Indonesia yang tidak baku, seperti singkatan atau *slang* (bahasa gaul) dengan lebih baik [22]. Penelitian terkait keunggulan IndoBERTweet telah dilakukan oleh [23] dengan membandingkan model IndoBERT, IndoBERTweet, dan mBERT pada penerapan klasifikasi *multilabel* terkait ulasan mahasiswa. Pada penelitian ini, IndoBERTweet menunjukkan performa paling baik di antara model IndoBERT dan

mBERT dengan dengan nilai *F1 score* sebesar 0.8462, melebihi IndoBERT dengan nilai 0.8243 dan mBERT dengan nilai 0.8230

Kemudian, penelitian lain terkait penerapan *hybrid model* IndoBERTweet dan Bi-LSTM dilakukan oleh [24] untuk mendeteksi *hate speech* berbahasa Indonesia pada sosial media Twitter. Hasil penelitian ini menunjukkan bahwa kombinasi tersebut paling unggul dibandingkan dengan kombinasi model lainnya. Nilai akurasi pada penggabungan model IndoBERTweet dan BiLSTM adalah sebesar 93.7%, *recall* sebesar 92.9%, *precision* sebesar 93.8%, dan *F1-score* sebesar 93.3%. Sedangkan, kombinasi model lain yaitu IndoBERTweet dan CNN memiliki hasil nilai akurasi sebesar 88%, *recall* sebesar 87.7%, *precision* sebesar 87.5%, dan *F1-score* sebesar 87.6% Berdasarkan beberapa penelitian tersebut, dapat dibuktikan bahwa penerapan model IndoBERTweet dapat bekerja dengan baik pada *dataset* sosial media berbahasa Indonesia. Dengan demikian, penelitian ini akan menerapkan *hybrid model* pada IndoBERTweet dengan Bi-GRU. *Hybrid model* merupakan kombinasi dua atau lebih model yang digabungkan untuk meningkatkan akurasi suatu prediksi [25].

Bi-Directional Gated Recurrent Unit (Bi-GRU) merupakan salah satu tipe dari arsitektur *Recurrent Neural Network* (RNN) yang memproses sekuensi input secara dua arah [26]. Arsitektur dasar dari Bi-GRU adalah GRU yang bekerja dengan cara menggabungkan *gating mechanism* yang memungkinkan jaringan untuk secara selektif mengingat atau melupakan informasi dari *input* sebelumnya, sehingga model dapat menangkap ketergantungan jangka panjang. Dengan demikian, Bi-GRU mampu menangkap informasi kontekstual secara efektif. Metode ini dipilih karena menerapkan mekanisme *gating* yang lebih sederhana dan tetap mempertahan kemampuannya untuk tetap efektif dalam mempelajari ketergantungan jangka panjang dengan lebih efisien secara komputasional dibandingkan dengan metode *Recurrent Neural Network* lainnya [27]. Penerapan *hybrid model* yang digabung dengan RNN terbukti efektif dalam meningkatkan performa klasifikasi [24]. Penelitian terkait penerapan Bi-GRU pada *hybrid model* telah dilakukan oleh [26] dengan penggabungan model BERT-BiGRU-Senti-GCN untuk menganalisis sentimen pelanggan pada *e-commerce*. Penelitian ini

menerapkan BERT sebagai *word embedding* dan Bi-GRU sebagai *feature extractor* yang memungkinkan untuk menangkap informasi dari hasil *word embedding* secara kontekstual dan efektif. Hasil penelitian menunjukkan bahwa kombinasi model BERT-BiGRU-GCN memiliki performa yang paling unggul dibandingkan dengan kombinasi model lainnya. Dengan demikian, penerapan kombinasi model IndoBERTweet dan Bi-GRU diharapkan dapat menjadi solusi yang efektif untuk deteksi *cyberbullying* pada sosial media, khususnya X atau twitter.

Penelitian ini difokuskan pada pengaplikasian *hybrid model* IndoBERTweet yang dikombinasikan dengan *Bi-Directional Gated Recurrent Unit* (Bi-GRU) untuk mendeteksi *cyberbullying* pada sosial media X berbahasa Indonesia. Pada penelitian ini akan diterapkan model IndoBERTweet pada bagian *word embedding* secara kontekstual agar dapat memberikan pemahaman konteks pada bahasa informal yang sering dijumpai pada berbagai sosial media. Sementara itu, Bi-GRU digunakan sebagai lapisan lanjutan yang dapat menangkap konteks urutan kata sehingga informasi yang dihasilkan dari proses *embedding* dengan IndoBERTweet dapat diolah secara lebih mendalam. Hal ini membuat model mampu memahami makna kalimat secara utuh, termasuk dalam mendeteksi indikasi *cyberbullying* pada data teks sosial media, khususnya X. Kombinasi kedua model tersebut diharapkan dapat mengatasi keterbatasan pendekatan konvensional serta meningkatkan kemampuan model dalam mengenali pola bahasa yang kompleks secara kontekstual. Hal ini bertujuan agar dapat menghasilkan model deteksi *cyberbullying* berbahasa Indonesia yang lebih akurat, adaptif terhadap variasi bahasa di media sosial khususya X, serta dapat menjadi dasar pengembangan sistem penyaringan konten *cyberbullying* otomatis yang lebih efektif.

Penelitian ini juga menerapkan tahap pemberian label pada data secara otomatis yang menjadi aspek penting dalam proses pembuatan model untuk mendeteksi *cyberbullying*. Otomatisasi label telah dilakukan pada penelitian sebelumnya dengan menggabungkan metode VADER lexicon dengan *Generative Large Language Model* (ChatGPT) untuk pemberian label pada *tweet cyberbullying*. VADER lexicon merupakan sebuah *tool* berbasis *rule-based* untuk klasifikasi sentimen dan *Generative Large Language Model* (ChatGPT) merupakan

sebuah model yang memiliki kemampuan untuk menganalisis dan memahami suatu kalimat keseluruhan secara kontekstual. Hasil penelitian menunjukkan bahwa pendekatan kedua metode pelabelan tersebut dapat meningkatkan kualitas label pada data [21]. Namun, pada penelitian [28] menunjukkan bahwa beberapa model *lexicon-based sentiment analysis*, yang memiliki karakteristik serupa dengan VADER, menghasilkan performa yang lebih rendah dibandingkan *Generative Large Language Model*. Oleh karena itu, penelitian ini akan menerapkan pendekatan otomatisasi label, yaitu menggunakan *Generative Large Language Modeling* dengan menerapkan dua jenis model yang berbeda. Dengan demikian, pendekatan tersebut diharapkan dapat menghasilkan penentuan label yang lebih akurat, konsisten, dan efektif pada studi kasus *cyberbullying* yang memerlukan pemahaman konteks secara lebih mendalam.

Data mentah yang diambil dari Twitter atau X seringkali tidak memiliki distribusi kelas yang seimbang setelah dilakukan proses pemberian label. Ketidakseimbangan pada data dapat menyebabkan model mengalami kesulitan untuk mempelajari data karena kurangnya distribusi data pada suatu kelas minoritas [29]. Terdapat berbagai metode yang dapat menangani masalah ketidakseimbangan distribusi data, salah satunya yaitu *Focal Loss*. *Focal Loss* merupakan sebuah metode yang digunakan untuk meningkatkan sensitivitas model pada kelas minoritas dengan memfokuskan proses pelatihan pada sampel data yang sulit. Salah satu keunggulan dari metode ini adalah mampu mengatasi ketidakseimbangan pada data tanpa mengubah data asli, seperti *oversampling* dan *undersampling* [30]. Metode ini telah diterapkan pada beberapa penelitian, salah satunya yaitu penelitian [31] yang memiliki data dengan ketimpangan distribusi kelas yang sangat tinggi pada data klasifikasi teks. Penerapan *focal loss* pada penelitian tersebut mampu menghasilkan nilai *precision* lebih tinggi dibandingkan dengan metode *oversampling*. Maka dari itu, pada penelitian ini akan diterapkan metode *Focal Loss* sebagai *loss function* pada proses pelatihan model yang diharapkan dapat mempelajari kelas yang memiliki distribusi data yang rendah dan menghasilkan performa model yang lebih optimal.

Pada penelitian ini, penulis akan membuat sebuah laman sederhana untuk menerapkan hasil model yang telah dilakukan pada penelitian ini dalam dunia nyata. Laman sederhana ini memungkinkan pengguna untuk menulis sebuah *tweet*. Kemudian sistem akan menganalisis apakah *tweet* tersebut terindikasi *cyberbullying* yang terdiri dari beberapa kategori yaitu *sexual harrasment, body shaming,* kata-kata kasar, dan *hate speech,* atau tidak terindikasi *cyberbullying*. Jika *tweet* yang ditulis terindikasi *cyberbullying*, maka *tweet* tersebut tidak bisa terkirim dan menampilkan *pop out* dari salah satu kategori *cyberbullying* yang terdeteksi, yaitu *sexual harrasment, hate speech, body shaming,* atau kata-kata kasar. Selain itu, sistem dapat menampilkan jumlah *cyberbullying* yang pernah dilakukan oleh suatu *user* sebagai hasil dari analisis lanjutan. Dengan demikian, sistem *website* ini diharapkan dapat mengurangi dan mencegah persebaran konten yang mengandung *cyberbullying* di media sosial, serta dapat mengetahui pola *cyberbullying* yang pernah dilakukan oleh pengguna

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, pada bagian ini terdapat berbagai rumusan masalah berupa pertanyaan yang diharapkan dapat memandu alur penelitian pada studi kasus ini. Adapun beberapa rumusan masalah yang terkait dipaparkan sebagai berikut.

1. Bagaimana pengimplementasian model IndoBERTweet-BiGRU dapat mendeteksi teks yang mengandung unsur *cyberbullying?*
2. Bagaimana performa model yang dihasilkan berdasarkan penerapan eksplorasi dari berbagai eksperimen pada tahap *pre-processing* dan penanganan ketidakseimbangan kelas?
3. Bagaimana performa yang dihasilkan oleh metode *Focal Loss* sebagai *loss function* dalam menangani ketidakseimbangan pada data?
4. Bagaimana perancangan terkait *User Interface* (UI) berbasis *web* yang terintegrasi dengan model deteksi *cyberbullying*?

1.3. Batasan Masalah

Dalam penelitian ini, terdapat batasan masalah agar peneliti dapat memfokuskan pada permasalahan yang dikaji. Berikut merupakan batasan-batasan yang diterapkan pada penelitian ini:

1. Penelitian difokuskan pada identifikasi teks yang mengandung *cyberbullying* dalam bahasa Indonesia dengan berbagai kategori pada kata kunci, yaitu *hate speech*, *sexual harrasment*, *body shaming*, dan kata-kata kasar dengan jumlah 31 kata kunci.
2. Model yang digunakan merupakan kombinasi IndoBERTweet sebagai *contextual word embedding* dan BiGRU sebagai *feature extractor*.
3. Menggunakan data primer yang diperoleh dengan teknik *scraping* dari *platform* media sosial X dengan maksimal 1700 data terakhir pada setiap kata kunci.

1.4. Tujuan Penelitian

Berdasarkan permasalahan yang telah dipaparkan sebelumnya, tujuan penelitian ini adalah sebagai berikut.

1. Mengetahui hasil pengimplementasian model IndoBERTweet yang dikombinasikan dengan BiGRU dalam mendeteksi teks yang mengandung *cyberbullying.*
2. Menghasilkan model dengan performa terbaik berdasarkan eksperimen pada tahap *pre-processing* dan penanganan ketidakseimbangan kelas yang mampu mendeteksi teks yang mengandung *cyberbullying.*
3. Mengetahui performa pada pengimplementasian metode *Focal Loss* sebagai *loss function* untuk menangani permasalahan terkait ketidakseimbangan distribusi pada data.
4. Merancang dan menghasilkan *User Interface* berbasis *website* yang terintegrasi dengan model deteksi *cyberbullying.*

1.5. Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan manfaat bagi berbagai pihak, baik dari segi teoritis maupun praktis, di antaranya yaitu:

1. Manfaat teoritis

   Memberikan pemahaman terkait penerapan model IndoBERTweet dan Bi-GRU dalam studi kasus deteksi *cyberbullying*. Selain itu, penelitian ini dapat memeperkaya literatur dalam pengklasifikasian teks, terutama pada penggunaan IndoBERTweet dan Bi-GRU sebagai alat pemrosesan teks pada teknik *Natural Language Processing* berbahasa Indonesia.

2. Manfaat praktis

   a. Bagi penulis, penelitian ini diharapkan dapat menjadi landasan untuk meningkatkan pemahaman dan keterampilan penulis dalam mengembangkan kombinasi model IndoBERTweet dan Bi-GRU untuk mendeteksi teks *cyberbullying* berbahasa Indonesia pada *platform* media sosial.

   b. Bagi peneliti, penelitian ini dapat menjadi referensi serta dasar pengembangan metode atau pendekatan baru dalam deteksi *cyberbullying* berbahasa Indonesia.

   c. Bagi pengembang aplikasi, hasil penelitian ini dapat dijadikan acuan dalam mengintegrasikan sistem pendeteksi *cyberbullying* ke dalam *platform* berbasis *web* atau media sosial.