BAB I

PENDAHULUAN

1.1. Latar Belakang

Media sosial membuka ruang bagi publik untuk berbagi pendapat secara bebas, namun komentar yang dibagikan secara publik juga berpotensi mengandung konten *toxic* yang dapat merusak reputasi individu dan memicu berbagai bentuk konflik (Shahid dkk., 2024). *Toxic language* mencakup berbagai bentuk ujaran bermasalah, mulai dari *abusive* dan *offensive language* hingga *cyberbullying,* termasuk ujaran yang tidak selalu dapat diproses secara hukum tetapi bersifat merugikan secara sosial (Grotti, 2024).

Dalam konteks *Natural Language Processing* (NLP), deteksi *toxic comment* merupakan bagian dari analisis sentimen yang berfokus pada identifikasi sentimen negatif yang merugikan seperti penggunaan bahasa kasar pada sebuah komentar sebagaimana didukung oleh berbagai penelitian di ranah analisis sentimen (Shahid dkk., 2024).

Kajian deteksi *toxic comment* pada tahap awal banyak memanfaatkan metode *machine learning* tradisional yang bergantung pada fitur statistik teks. (Sulistiyawati dkk., 2022) menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan Naive Bayes untuk mendeteksi kata kasar berbahasa Indonesia namun cenderung mengalami *overfitting* dengan performa *f1-score testing* 84% dan masih terbatas pada skenario *monolingual*. Penelitian (Salsabila Azzahra dkk., 2021), menemukan bahwa model tradisional cenderung tidak bisa menangkap konteks sehingga sering gagal mendeteksi komentar *toxic. Best model*

mereka yaitu Linear SVM dengan *text processing* menghasilkan *f1-score* 76,67%, untuk mengatasi keterbatasan tersebut, penelitian kemudian beralih ke pendekatan *deep learning.*

Penelitian (Sushma dkk., 2025) membandingkan berbagai pendekatan dan hasilnya menunjukkan bahwa representasi *Bidirectional Encoder Representations from Transformers* (BERT) dengan kombinasi *Long Short-Term Memory* (LSTM) memberikan peningkatan performa yang cukup besar dengan akurasi 92,11% tetapi pada penelitian tersebut belum melakukan *fine-tuning* pada model BERT dan masih terbatas pada bahasa Inggris. Penelitian tersebut secara eksplisit menekankan bahwa *fine-tuning* transformer merupakan arah penelitian lanjutan yang penting untuk meningkatkan performa adaptabilitas model. Penelitian oleh (Nabiilah dkk., 2024) juga menunjukkan bahwa kombinasi *Indonesian* BERT (IndoBERT) dan *Multilingual* BERT (mBERT) mampu meningkatkan performa klasifikasi ujaran *toxic* dan menghasilkan *f1-score* 88,9% tetapi mengalami *overfitting* dan masih terbatas pada skenario *monolingual.* Temuan serupa juga ditunjukkan oleh (Made dkk., 2025) yang mengembangkan sistem deteksi komentar berbahaya menggunakan kombinasi *Bidirectional* LSTM (BiLSTM) dan FastText dan menghasilkan *f1-macro* 98% tetapi masih terbatas pada skenario *monolingual*.

Studi oleh (Neog & Baruah, 2025) menunjukkan bahwa model kombinasi antara *Convolutional Neural Network* (CNN), BiLSTM dan *Bidirectional Gated Recurrent Unit* (BiGRU) mampu memberikan performa yang lebih baik dengan *f1-score 93,47%*, tetapi masih terbatas pada skenario *monolingual.* Menariknya, pada

penelitian tersebut secara eksplisit menekankan dalam *future works* perlunya eksplorasi terhadap pendekatan *cross-lingual.*

Perkembangan penelitian deteksi *toxic comment* saat ini mulai bergeser dari pendekatan *monolingual* menuju pendekatan *cross-lingual*. Hal ini didorong oleh meningkatnya jumlah pengguna media sosial dengan latar belakang bahasa yang beragam, sehingga dalam satu platform sering ditemukan komentar dalam berbagai bahasa bahkan campuran beberapa bahasa sekaligus (Song, Huang, & Xiao, 2021). Kondisi tersebut menyebabkan model yang hanya dilatih menggunakan satu bahasa menjadi kurang efektif dalam menangkap informasi semantik lintas bahasa (Song, Huang, & Zhang, 2021).

Transisi menuju deteksi lintas bahasa didorong oleh ketersediaan model *pretrained* lintas bahasa berskala besar seperti *Cross-lingual Language Model with RoBERTa Architecture* (XLM-RoBERTa), yang memungkinkan transfer pengetahuan secara efektif antar bahasa tanpa perlu bergantung pada data hasil translasi (Song, Huang, & Xiao, 2021; Song, Huang, & Zhang, 2021; Waysi Naaman dkk., 2025; Xu dkk., 2024) Meskipun demikian, tantangan yang terus berlanjut dalam ranah ini adalah ketidakseimbangan kelas yang ekstrem antara sampel komentar *toxic* dan *non-toxic* (He & Garcia, 2009; Thabtah dkk., 2020) yang berisiko menurunkan performa model klasifikasi dan meningkatkan risiko kesalahan prediksi pada kelas minoritas (Song, Huang, & Zhang, 2021).

Penelitian (Song, Huang, & Xiao, 2021) melakukan komparasi teknik *imbalance data handling* pada dua model *cross-lingual* dalam konteks analisis sentimen yaitu deteksi *toxic comment* sebagai bentuk sentimen negatif yang

kemudian menjadi dasar tugas pada penelitian ini. Hasil terbaik penelitian tersebut berada pada *Model-Fusion-3* dengan *f1-macro 84,49%,* namun hasil terbaik pada model independennya dimiliki oleh XLM-RoBERTa dengan pendekatan *Loss-Fusion* yang menghasilkan *f1-macro* 83,89%. Penelitian ini menemukan bahwa metode fusi BCE Loss dan Focal Loss cenderung memberikan performa lebih baik dibandingkan hanya mengandalkan satu metode *loss* saja. Namun, penelitian tersebut belum melibatkan bahasa Indonesia dan masih mengandalkan strategi augmentasi berbasis translasi. Pada penelitian tersebut secara eksplisit menekankan bahwa penggunaan data asli bahasa target lebih disarankan dibandingkan hasil translasi. Penelitian tersebut juga merekomendasikan pengembangan lanjutan terhadap strategi *imbalance data handling*, khususnya melalui pendekatan *algorithm-level method* seperti melalui eksplorasi *adaptive loss weighting* atau penyesuaian rasio bobot fusi *loss* secara adaptif.

Penelitian (X. Liu dkk., 2024) mengusulkan *Dynamic-Recall Focal Loss* (DRFL) sebagai pengembangan Focal Loss untuk mengatasi *class imbalance* melalui mekanisme pembobotan *loss* yang disesuaikan secara dinamis berdasarkan nilai *recall* kelas*,* dan terbukti melampaui Focal Loss sebesar 2% pada akurasi dan *f1-macro*. Pendekatan ini dapat dikombinasikan dengan metode *loss* lain dengan mekanisme *adaptive loss weighting* untuk meningkatkan efektivitasnya dalam skenario *multi-komponen loss* yang menyesuaikan kontribusi setiap komponen *loss* secara dinamis selama pelatihan (Heydari dkk., 2019).

Pada penemuan lain, penelitian (Valizadehaslani dkk., 2022) menemukan bahwa strategi *fine-tuning* yang dikonfigurasi khusus untuk menangani *class*

imbalance dapat meningkatkan *f1-score* kelas minoritas sebesar 1,6% dibandingkan *fine-tuning vanilla*. Selain itu, pada tahap inferensi, *threshold tuning* dapat digunakan untuk mengoptimalkan batas keputusan guna meningkatkan performa kelas minoritas tanpa perlu melatih ulang model (Esposito dkk., 2021).

Permasalahan *class imbalance* pada deteksi *toxic* comment masih membuka ruang optimasi melalui *adaptive loss weighting* dan strategi *fine-tuning*. Penelitian ini menangani *class imbalance* pada level algoritma tanpa mengubah distribusi data. (Song, Huang, & Xiao, 2021) menunjukkan bahwa dua *loss* memberikan keseimbangan performa yang lebih baik serta mendorong eksplorasi *adaptive loss weighting*. Sementara itu, (X. Liu dkk., 2024) mengusulkan DRFL sebagai pengembangan dari Focal Loss, dan (Valizadehaslani dkk., 2022) menunjukkan bahwa *two-stage fine-tuning* efektif dalam meningkatkan F1 kelas minoritas. Berdasarkan temuan tersebut, penelitian ini mengeksplorasi berbagai konfigurasi *fusion loss* berbasis BCE, Focal Loss, dan DRFL serta menerapkan *two-stage fine-tuning* dan *threshold tuning* untuk mengoptimalkan performa kelas minoritas.

1.2. Rumusan Masalah

Berdasarkan latar belakang dan celah penelitian yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

a. Bagaimana mengatasi permasalahan data tidak seimbang dengan mengoptimasi *cross-lingual pretrained model* menggunakan pendekatan *algorithm-level method?*

b. Bagaimana pengaruh optimasi *algorithm-level method* pada *cross-lingual pretrained model* untuk deteksi komentar *toxic* dalam skenario bahasa Inggris dan bahasa Indonesia tanpa memanfaatkan dataset hasil translasi pada bahasa target?

1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini dijabarkan dalam beberapa poin di bawah, yaitu:

a. Menerapkan *adaptive fusion loss* dengan *two-stage fine-tuning* serta *threshold tuning* pada model XLM-RoBERTa untuk dianalisis pengaruhnya terhadap model dengan kondisi data tidak seimbang.
b. Mengevaluasi performa model XLM-RoBERTa yang dioptimasi dengan *algorithm-level method* dalam mendeteksi komentar *toxic* pada skenario *cross-lingual* bahasa Indonesia dan bahasa Inggris berbasis dataset asli bahasa target.

1.4. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

a. Memberikan alternatif strategi algoritma pelatihan model *deep learning* untuk penanganan *class imbalance* pada tugas klasifikasi teks.
b. Memberikan kontribusi ilmiah dalam pengembangan metode deteksi komentar *toxic cross-lingual* yang dapat menjadi referensi penelitian selanjutnya.

1.5. Batasan Masalah

Penelitian ini memiliki beberapa batasan agar ruang lingkup penelitian lebih terarah, yaitu sebagai berikut:

a. Penelitian ini hanya berfokus pada deteksi komentar *toxic* bahasa Inggris dan bahasa Indonesia sehingga belum mencakup skenario multibahasa.
b. Model tidak dirancang khusus untuk menangani masalah *code-mixed.*
c. Penelitian ini membatasi klasifikasi komentar hanya pada dua kelas, yaitu *toxic* dan *neutral,* serta belum mencakup klasifikasi multikelas.
d. Penelitian ini berfokus pada kontribusi metode sehingga belum masuk ke dalam penerapan dunia nyata.