BAB I
PENDAHULUAN

Bab pendahuluan ini membahas mengenai latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan yang akan digunakan dalam dokumen skripsi ini yang berjudul Klasifikasi Komentar *Cyberbullying* Berbahasa Indonesia pada Media Sosial Menggunakan Metode DistilBERT.

1.1 Latar Belakang

Perkembangan teknologi internet telah mengubah pandangan komunikasi masyarakat Indonesia secara fundamental. Media sosial kini menjadi ruang utama masyarakat dalam memperoleh dan menyebarkan informasi, menggantikan sebagian besar interaksi tatap muka tradisional. Berdasarkan laporan Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pada tahun 2024 jumlah pengguna internet di Indonesia telah mencapai lebih dari 215 juta orang, atau sekitar 79,5% dari total populasi nasional (APJII, 2024). Temuan ini sejalan dengan penelitian Ashari dkk. (2023) yang menunjukkan bahwa aktivitas masyarakat Indonesia kini didominasi oleh penggunaan media sosial sebagai sarana komunikasi dan pertukaran informasi.

Media sosial telah menjadi salah satu ruang utama masyarakat Indonesia dalam berinteraksi, berbagi informasi, dan mengekspresikan opini di ruang digital. Kuss dan Griffiths (2017) menjelaskan bahwa media sosial memiliki peran penting dalam membentuk pola komunikasi masyarakat modern karena memungkinkan hubungan yang lebih cepat, luas, dan tidak terikat oleh batasan geografis. Di sisi lain, Hutabarat (2023) menyoroti bahwa media sosial berpotensi merenggangkan hubungan interpersonal yang sudah terjalin dan menurunkan intensitas interaksi tatap muka secara drastis. Terlepas dari dampak tersebut, laporan Kemp (2024) mencatat bahwa lebih dari 49,9% pengguna internet Indonesia aktif menggunakan media sosial setiap hari, dengan rata-rata durasi penggunaan mencapai lebih dari tiga jam per hari. Kondisi ini mencerminkan dominasi media sosial dalam aktivitas komunikasi digital masyarakat Indonesia di era saat ini.

Dengan seiring meningkatnya intensitas penggunaan media sosial, muncul pula fenomena sosial yang menjadi perhatian serius, yaitu *cyberbullying*. *Cyberbullying*

merupakan bentuk perundungan yang dilakukan melalui media elektronik atau platform daring, yang memiliki karakteristik dapat dilakukan secara anonim, tidak terbatas ruang dan waktu, serta berpotensi menyebar dengan cepat di ruang digital (Fazry & Apsari, 2021). Marlef dan Muda (2024) juga menegaskan bahwa *cyberbullying* kerap muncul melalui platform media sosial dan berdampak signifikan terhadap kondisi psikologis korban. Berdasarkan laporan UNICEF (2020), sekitar 45% dari 2.777 anak muda di Indonesia pernah mengalami bentuk *cyberbullying*. Tingginya angka kejadian ini tidak terlepas dari karakteristik unik media sosial sebagai ruang utama penyebaran *cyberbullying*.

Tingginya kasus *cyberbullying* di Indonesia juga dikonfirmasi oleh penelitian Efianingrum dkk. (2020) yang menemukan bahwa media sosial, khususnya kolom komentar, merupakan salah satu kanal utama terjadinya perundungan secara verbal, seperti hinaan, ujaran kebencian, pelecehan, dan ancaman. Fenomena *cyberbullying* juga membawa dampak psikologis yang serius bagi para korbannya. Berdasarkan penelitian Ningrum dan Amna (2020) yang menegaskan bahwa paparan berulang terhadap ujaran negatif dari *cyberbullying* dapat memperburuk kesehatan mental remaja, terutama pada masa perkembangan sosial dan emosional.

Bentuk komentar yang mengandung *cyberbullying* sering kali tidak muncul secara eksplisit, melainkan menggunakan bahasa sarkastik, plesetan, atau sindiran halus, sehingga menyulitkan proses deteksi dengan pendekatan berbasis *keyword matching* atau pencocokan kata kunci sederhana (Huang dkk., 2024). Padahal, fenomena ini berpotensi menimbulkan dampak yang luas dan masif, sehingga sangat diperlukan metode yang lebih cerdas dan akurat.

Maka dari itu, diperlukan metode yang efektif dan akurat untuk mengklasifikasikan komentar *cyberbullying* di media sosial. Salah satu bidang yang berperan penting dalam pengolahan teks berbahasa alami adalah *Natural Language Processing* (NLP), yang memungkinkan komputer memahami, memproses, dan menganalisis bahasa manusia (Amien, 2023). Kemampuan NLP dalam mengolah data teks yang tidak terstruktur menjadikannya sangat relevan untuk menangani karakteristik bahasa informal, sarkastik, dan kontekstual yang umum ditemukan pada komentar media sosial. Sejalan dengan tujuan tersebut, penerapan model berbasis NLP untuk klasifikasi komentar *cyberbullying* di media

sosial menjadi strategi penting dalam mendukung deteksi dini serta menciptakan ruang digital yang lebih aman dan sehat bagi pengguna.

Berbagai penelitian telah mengembangkan model klasifikasi teks menggunakan metode *machine learning*. Dalam penelitian terkait, Kusuma dkk. (2023) membandingkan *Naïve Bayes* (NB), *Support Vector Machine* (SVM), dan *Random Forest* (RF) untuk deteksi komentar *cyberbullying* Bahasa Indonesia. Hasilnya menunjukkan performa yang terbatas, di mana RF dan NB mencapai akurasi 77%, sementara SVM hanya 76%.

Seiring perkembangan, pendekatan *Deep Learning* (DL) juga menawarkan peningkatan. Penelitian oleh Lo dkk. (2024) membandingkan ML, yaitu *Random Forest* (RF) dan *Support Vector Machine* (SVM) dengan DL, yaitu *Convolutional Neural Networks* (CNN) dan *Recurrent Neural Networks* (RNN) untuk deteksi *cyberbullying*. Studi ini menemukan bahwa model DL, seperti RNN dengan nilai akurasi 84.71% dan CNN dengan nilai akurasi 84%, secara signifikan mengungguli model ML RF hanya 72%.

Kemajuan yang lebih signifikan muncul dengan hadirnya model berbasis *Transformer* (sebuah arsitektur yang diperkenalkan oleh Vaswani dkk. (2017)). *Transformer* dirancang untuk memahami konteks antar kata secara lebih efektif melalui mekanisme *Self-Attention*, sehingga mampu menangkap hubungan dalam kalimat secara lebih menyeluruh dibandingkan pendekatan DL. Salah satu pengembangan penting dari arsitektur ini adalah *Bidirectional Encoder Representations from Transformers* (BERT) yang dikembangkan oleh Devlin dkk. (2019). BERT memungkinkan kemampuan pemahaman konteks dua arah sehingga menghasilkan representasi bahasa yang jauh lebih kaya dan akurat.

Dalam konteks Bahasa Indonesia, BERT kemudian diadaptasi menjadi *Indonesian Bidirectional Encoder Representations from Transformers* (IndoBERT), yang dilatih menggunakan himpunan data teks (korpus) berskala besar Berbahasa Indonesia. Model ini menjadi salah satu yang paling unggul dalam berbagai tugas pemrosesan bahasa alami. Penelitian oleh Novandian dkk. (2024) mengungkapkan bahwa untuk deteksi *cyberbullying* Bahasa Indonesia, IndoBERT dengan nilai akurasi 96.7% unggul secara performa dibandingkan dengan model DL seperti Bi-LSTM dengan nilai akurasi 94.9%. Meskipun IndoBERT unggul secara akurasi dan menjadi dasar penelitian klasifikasi teks bahasa Indonesia, model ini memiliki kelemahan signifikan dalam hal efisiensi. Novandian dkk.

(2024) menemukan bahwa IndoBERT jauh lebih boros waktu dengan waktu run 59 menit 6 detik dibandingkan Bi-LSTM dengan waktu run 2 menit 57 detik.

Kesenjangan antara kebutuhan akan akurasi dan tantangan efisiensi komputasi mendorong pengembangan model berbasis *Transformer* yang lebih ringan. DistilBERT merupakan versi reduksi dari BERT yang dikembangkan melalui teknik *knowledge distillation*, yaitu proses transfer pengetahuan dari model besar ke model yang lebih kecil agar tetap mempertahankan performa yang kompetitif. Melalui pendekatan ini, DistilBERT mampu mempertahankan sebagian besar kemampuan representasi BERT dengan ukuran *hidden layer* dan jumlah parameter yang lebih kecil. Penelitian oleh Putri dkk. (2023) secara langsung membandingkan IndoBERT dengan DistilBERT. Penelitian Putri dkk. (2023) membuktikan bahwa DistilBERT lebih cepat dalam proses *training* dan lebih hemat memori GPU. Meskipun akurasi DistilBERT sedikit lebih rendah dengan nilai akurasi 85% dibandingkan BERT dengan nilai akurasi 89%, Putri dkk. (2023) menyimpulkan bahwa selisih performa tersebut tidak terpaut jauh, menjadikannya solusi yang seimbang antara performa dan efisiensi. Berdasarkan hal tersebut, DistilBERT dipilih untuk menyelesaikan masalah klasifikasi *cyberbullying* pada media sosial berbahasa Indonesia.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, rumusan masalah dalam penelitian ini adalah bagaimana cara mengklasifikasikan komentar *cyberbullying* berbahasa Indonesia pada media sosial menggunakan metode DistilBERT.

1.3 Tujuan dan Manfaat

Tujuan dari penelitian ini adalah mengimplementasikan model DistilBERT untuk menghasilkan model klasifikasi *cyberbullying* yang efisien secara komputasi serta memiliki performa tinggi berdasarkan metriks *accuracy, precision, recall,* dan *F1-score*. Sementara itu, manfaat dari penelitian ini adalah tersedianya model klasifikasi komentar *cyberbullying* berbahasa Indonesia yang dapat mendeteksi ujaran negatif secara otomatis. Model ini diharapkan dapat berkontribusi dalam upaya menciptakan ruang digital yang lebih sehat dan aman melalui deteksi dini terhadap komentar yang bersifat merundung atau menghina. Selain itu, hasil penelitian ini diharapkan dapat menjadi referensi bagi penelitian selanjutnya

dalam pengembangan sistem deteksi *cyberbullying* berbasis *deep learning* dan pemrosesan bahasa alami dalam konteks Bahasa Indonesia.

1.4 Ruang Lingkup

Penelitian skripsi ini memerlukan penentuan batasan ruang lingkup agar proses penelitian berlangsung secara terarah dan fokus, sehingga hasil yang diperoleh sesuai dengan tujuan yang telah ditetapkan. Adapun batasan ruang lingkup dalam penelitian ini ditetapkan sebagai berikut:

1. Ruang lingkup penelitian ini difokuskan pada komentar berbahasa Indonesia yang berasal dari platform media sosial. Media sosial dipilih karena merupakan ruang digital yang paling banyak digunakan oleh masyarakat untuk mengekspresikan opini, termasuk komentar yang berpotensi mengandung unsur *cyberbullying*. Platform media sosial yang digunakan dalam penelitian ini adalah *Instagram* dan *Twitter*.
2. *Dataset* yang digunakan dalam penelitian ini diperoleh dari tiga sumber publik di platform Kaggle, yaitu *dataset* yang disusun oleh Ramadhan (2024), Hanni (2021), dan Luqyana dkk. (2018). Ketiga *dataset* tersebut digabungkan dan melalui proses *cross-check* untuk menghindari duplikasi maupun inkonsistensi label.
3. *Dataset* diformat ke dalam bentuk *Comma Separated Values* (CSV**).**
4. Model yang digunakan adalah DistilBERT yang dikembangkan secara khusus untuk bahasa Indonesia oleh Wirawan (2024) melalui platform *HuggingFace*. Model ini menggunakan aktivasi *sigmoid*, menerapkan *Binary Cross Entropy* (BCE) sebagai fungsi *loss*, serta memanfaatkan AdamW sebagai fungsi optimasi.
5. Proses pencarian *hyperparameter* dilakukan menggunakan *Grid Search* dengan mengkombinasikan *learning rate*, *batch size*, *dropout*, dan *weight decay*, untuk memperoleh konfigurasi terbaik dalam meningkatkan performa model DistilBERT.

1.5 Sistematika Penulisan

Struktur penulisan laporan ini dibagi ke dalam lima bab utama yang membahas topik klasifikasi komentar *cyberbullying* di media sosial berbahasa Indonesia menggunakan metode DistilBERT. Pembagian bab ini dimaksudkan agar penulisan menjadi lebih terstruktur dan mudah dipahami. Berikut adalah ringkasan singkat dari masing-masing bab yang ada:

BAB I PENDAHULUAN

Bab ini membahas latar belakang, rumusan masalah, tujuan penelitian, manfaat, ruang lingkup, serta sistematika penulisan yang digunakan dalam penyusunan skripsi mengenai klasifikasi komentar *cyberbullying* berbahasa Indonesia menggunakan metode DistilBERT.

BAB II TINJAUAN PUSTAKA

Bab ini memaparkan landasan teori yang meliputi konsep media sosial, komentar *cyberbullying*, serta teori-teori pendukung terkait *natural language processing*, *deep learning*, dan model DistilBERT sebagai metode utama klasifikasi teks.

BAB III METODE PENELITIAN

Bab ini menjelaskan metodologi yang digunakan dalam penelitian, meliputi alur penelitian, pengumpulan dan pembagian data, tahapan *pre-processing*, tokenisasi, arsitektur model DistilBERT dengan *dense layer*, serta metode evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, *dan F1-score*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil eksperimen yang diperoleh dari proses pelatihan dan pengujian model, analisis performa berdasarkan metrik evaluasi, serta interpretasi terhadap hasil klasifikasi komentar *cyberbullying* yang dihasilkan oleh model.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian dan saran untuk pengembangan model atau penelitian selanjutnya yang berkaitan dengan deteksi komentar *cyberbullying* menggunakan pendekatan *deep learning*.