BAB I
PENDAHULUAN

1.1 Latar Belakang

Penggunaan media sosial dan platform digital sebagai sarana komunikasi sekarang telah menjadi bagian dari kehidupan sehari-hari. Melalui berbagai forum online, kolom komentar, dan sosial media, siapa saja dapat dengan mudah menyampaikan pendapat, berbagi pengalaman, maupun berinteraksi dengan pengguna yang lain. Namun, kemudahan ini juga dapat membuka peluang akan terjadinya perilaku negatif seperti *cyberbullying* atau perundungan secara online.

Cyberbullying merupakan tindakan menyakiti orang lain melalui kata-kata kasar, ancaman, atau penghinaan yang disebarkan lewat dunia maya. Dampak dari *cyberbullying* tidak hanya dirasakan secara emosional, tetapi juga dapat memengaruhi kesehatan mental korban. Fenomena ini telah menjadi masalah yang semakin serius di berbagai belahan dunia, termasuk Indonesia. *Cyberbullying* merupakan bentuk perundungan yang bermigrasi dari dunia nyata ke dunia maya atau media sosial. Hal ini terjadi karena kemudahan akses dan jangkauan luas media sosial yang memungkinkan interaksi tanpa batas jarak, sehingga perilaku negatif seperti *bullying* pun lebih mudah terjadi dan menyebar. *Cyberbullying* dapat terjadi pada siapa saja, termasuk anak-anak, remaja, dan dewasa. Namun, anak-anak dan remaja cenderung menjadi sasaran yang lebih rentan karena kurangnya pengalaman dan kemampuan mereka untuk menghadapi tindakan tersebut [1].

Kondisi *cyberbullying* di Indonesia saat ini telah mencapai tahap yang mengkhawatirkan. Merujuk pada laporan UNICEF (2020), tercatat sebanyak 45% anak dan remaja pernah menjadi target perundungan daring. Data pendukung dari BNPT (2023) mempertegas kondisi tersebut dengan temuan bahwa 40% remaja mengalami intimidasi digital yang berujung pada stres dan trauma mendalam. Fenomena ini diperparah dengan temuan SAFEnet pada awal 2024 yang

menunjukkan adanya lonjakan kasus kekerasan berbasis gender daring, di mana mayoritas korban berada pada rentang usia produktif (18–25 tahun) dan kelompok usia di bawah 18 tahun [2].

Pemerintah Indonesia telah mengambil langkah hukum untuk menanggapi persoalan ini. Undang-Undang Informasi dan Transaksi Elektronik (UU ITE) menjadi dasar hukum yang melindungi masyarakat dari tindakan kejahatan digital, termasuk *cyberbullying*. Dalam Pasal 27 ayat (3) disebutkan bahwa setiap orang dilarang dengan sengaja mendistribusikan atau mentransmisikan informasi elektronik yang memuat penghinaan atau pencemaran nama baik. Sementara itu, Pasal 27 ayat (4) mengatur larangan atas penyebaran informasi yang mengandung ancaman atau pemerasan. Ketentuan pidana untuk pelanggaran tersebut diatur dalam Pasal 45 ayat (3), yang menyebutkan bahwa pelaku dapat dikenakan hukuman penjara paling lama 12 tahun dan denda paling banyak Rp. 2 miliar. Aturan ini juga diperkuat dengan pasal-pasal dalam Kitab Undang-Undang Hukum Pidana (KUHP) seperti Pasal 310 tentang pencemaran nama baik dan Pasal 311 tentang fitnah [3].

Pemanfaatan kecerdasan buatan, khususnya melalui teknik klasifikasi teks, muncul sebagai solusi teknologi yang paling efektif dalam memitigasi *cyberbullying*. Melalui pendekatan ini, sistem mampu melakukan ekstraksi dan kategorisasi komentar secara otomatis ke dalam kelas *bullying* atau *non-bullying*. Implementasi deteksi otomatis dinilai jauh lebih konsisten dan efisien dibandingkan metode moderasi manual yang sangat terbatas dari sisi waktu dan tenaga manusia dalam menghadapi volume data yang besar.

Salah satu algoritma *machine learning* yang paling sering digunakan untuk klasifikasi biner adalah *Support Vector Machine* (SVM). Algoritma ini bekerja dengan membentuk garis pemisah optimal antara dua kelas data. SVM dikenal efektif dalam menangani data teks berdimensi tinggi dan mampu memberikan hasil klasifikasi yang akurat, meski jumlah data pelatihan terbatas. Karena

kemampuannya memahami pola bahasa dan struktur kalimat, SVM sangat cocok digunakan dalam mendeteksi komentar-komentar yang mengandung unsur *bullying*.

Penelitian dilakukan oleh Rahman, Abdillah, dan Komarudin (2021) dengan judul "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine". Dalam penelitian ini, data dibagi menjadi dua label utama, yaitu komentar yang mengandung unsur *bullying* dan komentar *non-bullying*. Penelitian ini bertujuan untuk membandingkan tiga kernel SVM, yaitu linear, polynomial, dan RBF, pada dataset berjumlah 1.000 data (700 data latih dan 300 data uji). Hasil menunjukkan bahwa kernel RBF memberikan performa terbaik dengan akurasi sebesar 93%. Temuan ini mengindikasikan bahwa metode SVM efektif digunakan untuk deteksi komentar berunsur bullying, dan mampu membedakan secara akurat antara komentar yang bersifat menyerang dengan yang netral [4]. Penelitian yang dilakukan oleh Nauli, Berutu, Budiati, dan Maedjaja (2023) berjudul *"Klasifikasi Kalimat Perundungan pada Twitter Menggunakan Algoritma Support Vector Machine"* bertujuan untuk mengklasifikasikan komentar di media sosial ke dalam dua kategori utama, yaitu *bullying* dan *non-bullying*. Penelitian ini memanfaatkan algoritma Support Vector Machine (SVM) dengan empat jenis kernel, yaitu linear, polynomial, RBF, dan sigmoid. Sebanyak 2.752 tweet diproses menggunakan metode TF-IDF untuk mengubah teks menjadi representasi numerik. Hasil pengujian menunjukkan bahwa kernel linear memberikan performa terbaik dengan akurasi sebesar 89,47%. Temuan ini menguatkan bahwa SVM, khususnya dengan kernel linear, sangat efektif dalam membedakan komentar yang mengandung unsur perundungan dengan komentar yang bersifat netral atau tidak menyerang [5].

Fokus utama penelitian ini adalah mengembangkan sebuah sistem deteksi dini (*early detection system*) yang mampu mengidentifikasi konten bermasalah sebelum dampak negatifnya meluas secara masif. Pemilihan platform TikTok dan YouTube didasarkan pada karakteristik interaksinya yang sangat dinamis dengan

densitas komentar yang tinggi namun memiliki tingkat ketidakteraturan bahasa yang ekstrem. Penggunaan istilah *slang*, singkatan idiosinkratik, hingga modifikasi ortografis (*typo*) pada platform tersebut menciptakan dimensi data yang sangat tinggi. Karakteristik data yang kompleks ini menuntut penggunaan algoritma yang tangguh seperti *Support Vector Machine* (SVM). Melalui pemisahan *hyperplane* yang akurat, SVM mampu menangani ruang fitur berdimensi tinggi hasil ekstraksi TF-IDF, sehingga tetap presisi dalam membedakan antara opini netral dan tindakan perundungan (*cyberbullying*) meskipun dalam bahasa yang tidak baku.

Adapun perbedaan sistem yang dikembangkan dalam penelitian ini dibandingkan penelitian sebelumnya terletak pada implementasinya. Sistem ini dirancang untuk mendeteksi secara dini komentar *cyberbullying* dengan menggunakan input berupa link konten dari YouTube dan TikTok yang diolah melalui proses *scraping* menggunakan *plugin Instant Data Scraper*. Komentar yang diperoleh akan melalui tahapan *preprocessing* teks (dengan *library* Sastrawi), pembobotan fitur TF-IDF, dan diklasifikasikan menggunakan algoritma *Support Vector Machine* (SVM) ke dalam dua kategori utama, yaitu *bullying* dan *non-bullying*. Hasil klasifikasi ditampilkan dalam bentuk grafik distribusi dan daftar komentar berlabel, sehingga dapat membantu mengenali potensi *cyberbullying* secara dini. Dengan pendekatan ini, penelitian **"Klasifikasi Komentar Online untuk Deteksi Dini Cyberbullying Menggunakan Metode Support Vector Machine"** diharapkan mampu berkontribusi dalam menciptakan ruang digital yang lebih aman, khususnya bagi anak-anak dan remaja yang rentan terhadap perundungan daring.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, permasalahan dalam penelitian ini adalah bagaimana membangun model klasifikasi komentar online yang dapat mendeteksi secara dini indikasi *cyberbullying* menggunakan algoritma *Support Vector Machine* (SVM), sehingga mampu membedakan secara tepat antara komentar yang mengandung unsur *bullying* dan *non-bullying*.

1.3 Batasan Masalah

Berikut ini batasan masalah dari penelitan ini:

1. Penelitian ini difokuskan pada komentar online yang berasal dari platform media sosial YouTube dan TikTok, dan tidak mencakup platform media sosial lainnya.
2. Komentar yang dianalisis diklasifikasikan ke dalam dua kategori, yaitu komentar *bullying* dan komentar *non-bullying*.
3. Algoritma *Support Vector Machine* (SVM) digunakan dalam penelitian ini sebagai metode klasifikasi komentar online.
4. Dataset yang digunakan berjumlah 8.000 data komentar (4.000 data YouTube dan 4.000 data TikTok) dengan mayoritas menggunakan Bahasa Indonesia.
5. Sistem pada penelitian ini hanya berperan sebagai alat bantu deteksi dini *cyberbullying* dan tidak digunakan sebagai dasar penentuan keputusan akhir tanpa evaluasi lanjutan.
6. Penelitian ini dibatasi pada proses klasifikasi komentar yang berasal dari konten video pada platform YouTube dan TikTok. Sistem yang dikembangkan hanya dapat memproses tautan (link) konten video, sehingga komentar dari konten berbentuk gambar atau foto tidak termasuk dalam ruang lingkup klasifikasi.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengembangkan model klasifikasi komentar online yang mampu mendeteksi secara dini indikasi *cyberbullying* dengan kategori *bullying* dan *non-bullying*.
2. Menerapkan algoritma *Support Vector Machine* (SVM) dengan metode TF-IDF untuk membedakan komentar berdasarkan pola bahasa dan bobot kepentingannya dalam teks.

3. Menyajikan hasil klasifikasi dalam bentuk visualisasi grafik dan daftar komentar terklasifikasi sebagai alat bantu analisis untuk mendukung pencegahan dan mitigasi dini *cyberbullying* di media sosial.

1.5 Manfaat Penelitian

Berikut manfaat dari penelitian ini:

1. Memberikan solusi berbasis teknologi untuk deteksi dini berbagai bentuk komentar *cyberbullying*, melalui pengembangan model klasifikasi otomatis yang mampu membedakan komentar *bullying* dan *non-bullying*, guna mengurangi dampak psikologis terhadap korban.
2. Membantu pengguna media sosial dalam mengenali dan meninjau komentar berpotensi berbahaya secara otomatis, mendukung proses moderasi konten agar ruang digital tetap aman.
3. Menjadi referensi dan dasar pengembangan sistem deteksi komentar negatif berbasis *machine learning* di masa depan pada berbagai platform media sosial.