BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan internet dan media sosial di Indonesia menunjukkan tren peningkatan yang pesat dan konsisten dari tahun ke tahun. Perkembangan media sosial ini memberikan ruang yang luas bagi penggunanya untuk berinteraksi, berbagi informasi, dan mengekspresikan diri [1]. Platform seperti X (dahulu Twitter) dan Instagram pun bertransformasi menjadi ruang publik digital utama bagi masyarakat untuk menyalurkan aspirasi dan kebebasan berpendapat. Kebebasan berekspresi di ruang digital ini dijamin secara konstitusional sebagai bagian dari hak asasi manusia, sehingga masyarakat merasa memiliki kebebasan yang luas dalam menyampaikan opininya di media sosial [2].

Namun, keterbukaan dan sifat anonimitas yang ditawarkan oleh platform-platform tersebut turut memicu munculnya fenomena ujaran kebencian (*hate speech*) secara masif. Ujaran kebencian merupakan bentuk komunikasi verbal maupun tulisan yang mengandung unsur penghinaan, provokasi, atau diskriminasi terhadap individu maupun kelompok berdasarkan suku, agama, ras, atau golongan tertentu, dan telah terbukti berdampak pada perpecahan sosial serta memicu konflik antarkelompok masyarakat [3]. Karakteristik ujaran kebencian di media sosial sangat bervariasi, mulai dari perdebatan teks yang panjang dan argumentatif hingga komentar-komentar pendek yang bersifat reaktif. Variasi ini menghasilkan volume data tekstual yang sangat masif dan tidak terstruktur (*unstructured data*) [4].

Volume data tekstual tersebut mengalir dalam bentuk ribuan komentar dan twit setiap detiknya, menjadikan moderasi konten yang dilakukan secara manual oleh manusia (*human moderation*) semakin sulit dilakukan, lambat dalam merespons, serta membutuhkan biaya operasional yang besar untuk mempekerjakan tenaga moderator dalam jumlah banyak. Ledakan volume data tersebut pada akhirnya membuat pendekatan moderasi manual mencapai titik jenuh dan tidak lagi efisien untuk membendung derasnya arus konten bermuatan ujaran kebencian [5].

Menjawab keterbatasan tersebut, pendekatan *Natural Language Processing* (NLP) dan algoritma klasifikasi *Machine Learning* menawarkan solusi otomatisasi yang mampu mengenali pola-pola kebahasaan kasar secara *real-time* dan pada skala data yang jauh lebih besar dibandingkan kapasitas manusia. Tinjauan sistematis

terhadap literatur NLP untuk deteksi ujaran kebencian menunjukkan bahwa pendekatan berbasis pembelajaran mesin, baik model klasik maupun arsitektur yang lebih mutakhir, secara konsisten mampu mengotomatisasi identifikasi teks toksik dengan tingkat akurasi yang terus meningkat seiring perkembangan teknik ekstraksi fitur dan pemodelan bahasa [6]. Studi bibliometrik terhadap tren riset deteksi ujaran kebencian turut mengonfirmasi bahwa algoritma klasifikasi seperti *Support Vector Machine* (SVM) dan *Naive Bayes* tetap menjadi fondasi metode yang banyak dikaji berdampingan dengan pendekatan *deep learning* dan model *transformer* yang lebih kompleks [5].

Di antara berbagai algoritma klasifikasi teks, Naive Bayes merupakan salah satu metode yang paling banyak diadopsi karena kesederhanaan konsepnya yang berbasis teorema probabilitas Bayes dengan asumsi independensi antarfitur. Keunggulan utama algoritma ini terletak pada kecepatan komputasinya yang sangat tinggi, sehingga sangat sesuai diterapkan pada data teks yang mengalir secara cepat dan kontinu di media sosial. Pengujian performa Naive Bayes pada klasifikasi teks berbahasa Indonesia menunjukkan bahwa proses pelatihan model dapat diselesaikan dalam hitungan detik dengan tetap menghasilkan akurasi yang kompetitif, membuktikan efisiensi sumber daya komputasi metode ini [7]. Efisiensi serupa juga dikonfirmasi pada perbandingan Naive Bayes dengan beberapa algoritma klasifikasi lain berbasis fitur TF-IDF untuk analisis sentimen ulasan produk, di mana Naive Bayes tetap menjadi salah satu metode paling ringan dari sisi komputasi [8], serta pada komparasi Naive Bayes dan SVM untuk analisis sentimen aplikasi berbasis media sosial [9]. Konsistensi keunggulan kecepatan dan kesederhanaan Naive Bayes ini juga tervalidasi pada berbagai konteks data Twitter yang berbeda-beda, mulai dari migrasi televisi digital, kebijakan kuota internet Kemdikbudristek pada masa pandemi, hingga sentimen publik terhadap ChatGPT, yang secara keseluruhan menunjukkan bahwa Naive Bayes tetap kompetitif meskipun umumnya sedikit di bawah SVM dari sisi akurasi [10–12].

Sebagai penantang, *Support Vector Machine* (SVM) menawarkan keunggulan yang berbeda. Ketika data teks ditransformasikan ke dalam representasi numerik menggunakan pembobotan *Term Frequency-Inverse Document Frequency* (TF-IDF), dimensi data yang dihasilkan menjadi sangat tinggi karena setiap kata unik dalam korpus akan direpresentasikan sebagai satu fitur tersendiri. SVM dikenal sangat andal dalam menangani permasalahan klasifikasi berdimensi tinggi semacam ini karena kemampuannya mencari *hyperplane* pemisah kelas secara optimal. Penerapan SVM dengan pembobotan TF-IDF terbukti efektif dalam

mendeteksi ujaran kebencian pada media sosial X dengan tingkat presisi yang tinggi [13], demikian pula pada klasifikasi pencemaran nama baik di media sosial X yang menunjukkan keunggulan SVM dalam menangani fitur teks berdimensi besar [14]. Kombinasi SVM dengan skema pembobotan TF-IDF N-Gram juga telah dibuktikan mampu meningkatkan akurasi klasifikasi teks secara umum [15]. Keunggulan SVM dalam menangani ruang fitur berdimensi tinggi dan data yang jarang (*sparse*) ini secara konsisten teramati pada berbagai domain klasifikasi teks lain di Indonesia, termasuk analisis sentimen penggunaan mobil listrik, evaluasi layanan BPJS, dan klasifikasi teks bencana pada media sosial, yang seluruhnya melaporkan SVM unggul dibandingkan Naive Bayes dari sisi akurasi [16–18].

Meskipun berbagai penelitian terdahulu telah membandingkan performa algoritma Naive Bayes dan Support Vector Machine (SVM) sebagaimana disebutkan pada paragraf sebelumnya, masih terdapat celah penelitian terkait skala data yang digunakan. Mayoritas penelitian tersebut masih memanfaatkan dataset dengan jumlah data yang relatif kecil, umumnya di bawah 2.000 data. Penggunaan dataset berskala kecil berpotensi menyebabkan model mengalami overfitting, yaitu kondisi ketika model terlalu menyesuaikan diri dengan pola pada data latih sehingga kemampuan generalisasinya terhadap data baru menjadi kurang optimal. Oleh karena itu, kemampuan kedua algoritma dalam menangani dataset berskala lebih besar masih memerlukan pengujian lebih lanjut untuk memperoleh gambaran performa yang lebih representatif [19].

Untuk mengatasi keterbatasan tersebut, penelitian ini memanfaatkan dataset teks berbahasa Indonesia yang terdiri atas 14.306 data berlabel dari repositori Hugging Face. Penggunaan dataset dalam jumlah besar menghasilkan representasi fitur berdimensi tinggi setelah dilakukan ekstraksi menggunakan metode TF-IDF, sehingga matriks fitur yang terbentuk bersifat sangat renggang (sparse) dan memiliki dimensi yang tinggi [20]. Oleh karena itu, penelitian ini menerapkan tahapan preprocessing, ekstraksi fitur TF-IDF, seleksi fitur, serta optimasi parameter sebagai bagian dari pipeline klasifikasi untuk menghasilkan proses pelatihan model yang lebih efisien dan konsisten. Seluruh tahapan tersebut diterapkan secara identik pada algoritma Naive Bayes dan Support Vector Machine (SVM) sehingga perbandingan performa kedua algoritma dilakukan pada kondisi eksperimen yang sama.

Berdasarkan uraian tersebut, penelitian ini bertujuan membandingkan performa algoritma Naive Bayes dan Support Vector Machine (SVM) dalam mengklasifikasikan ujaran kebencian (hate speech) pada teks media sosial

berbahasa Indonesia menggunakan dataset berskala besar. Perbandingan dilakukan berdasarkan metrik accuracy, precision, recall, dan F1-score, serta efisiensi waktu komputasi yang meliputi waktu pelatihan (training time) dan waktu pengujian (testing time). Dengan menggunakan pipeline pemrosesan data yang sama pada kedua algoritma, penelitian ini diharapkan dapat memberikan gambaran yang lebih objektif mengenai algoritma yang memiliki performa dan efisiensi komputasi terbaik dalam klasifikasi hate speech pada dataset berskala besar.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini dirancang untuk menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana penerapan algoritma Naive Bayes dan *Support Vector Machine* dalam mengklasifikasikan ujaran kebencian (*hate speech*) pada teks media sosial berbahasa Indonesia?
2. Bagaimana perbandingan performa antara algoritma Naive Bayes dan SVM berdasarkan metrik evaluasi (*accuracy, precision, recall*, dan *F1-score*) serta tingkat efisiensi waktu komputasi (waktu *training* dan *testing*)?

1.3 Batasan Permasalahan

Supaya ruang lingkup penelitian tetap jelas, batasan yang digunakan adalah sebagai berikut:

1. Data yang digunakan merupakan *dataset superset* sekunder berbahasa Indonesia berskala besar (terdiri dari 14.306 baris data) yang bersumber dari platform Hugging Face.
2. Penelitian ini berfokus murni pada pemrosesan teks (*Natural Language Processing*). Data berupa gambar, video, atau audio yang mungkin menyertai teks asli di media sosial tidak diikutsertakan.
3. Penelitian ini hanya melakukan klasifikasi biner (*binary classification*), yaitu mengategorikan teks ke dalam dua kelas utama: mengandung ujaran kebencian (*hate speech*) dan tidak mengandung ujaran kebencian (*non-hate speech*). Penelitian ini tidak mendeteksi atau mengklasifikasikan sub-kategori ujaran kebencian spesifik (seperti rasisme, seksisme, atau SARA).

1.4 Tujuan Penelitian

Mengacu pada rumusan masalah di atas, tujuan penelitian ini adalah sebagai berikut:

1. Mengimplementasikan dan menganalisis penerapan algoritma Naive Bayes dan *Support Vector Machine* (SVM) dalam mengklasifikasikan teks ujaran kebencian (*hate speech*) pada media sosial berbahasa Indonesia.
2. Membandingkan performa model klasifikasi antara algoritma Naive Bayes dan SVM berdasarkan metrik evaluasi (*accuracy, precision, recall*, dan *F1-score*) serta menganalisis tingkat efisiensi waktu komputasinya (waktu *training* dan *testing*).

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberi manfaat bagi beberapa pihak berikut:

1. Bagi Pengembangan Ilmu Pengetahuan dan Penelitian Selanjutnya: Hasil penelitian ini diharapkan dapat memperkaya kajian literatur di bidang *Natural Language Processing* (NLP), khususnya terkait analisis komparatif performa dan efisiensi waktu komputasi antara algoritma Naive Bayes dan *Support Vector Machine* (SVM). Temuan yang diperoleh dapat menjadi rujukan empiris untuk penelitian lanjutan mengenai klasifikasi *hate speech* menggunakan *dataset* berskala besar.
2. Bagi Pengembangan Sistem Moderasi Konten: Penelitian ini dapat memberikan panduan teknis bagi pengembang sistem moderasi konten otomatis dalam memilih algoritma klasifikasi yang paling optimal. Dengan adanya tolak ukur perbandingan antara tingkat akurasi dan kecepatan waktu prediksi (*testing*), pengembang dapat merancang sistem penyaringan konten ujaran kebencian yang tanggap dan efisien.
3. Bagi Organisasi atau Institusi : Penelitian ini dapat memberikan rujukan bagi organisasi, instansi pemerintah, atau lembaga pengamat ruang digital yang perlu memproses data teks dalam volume masif. Hasil komparasi ini mempermudah institusi dalam mengambil keputusan terkait pemilihan algoritma *Machine Learning* yang tidak hanya akurat, tetapi juga hemat sumber daya komputasi.

1.6 Sistematika Penulisan

Skripsi ini disusun dalam lima bab yang saling berkaitan agar alur pembahasannya mudah diikuti. Sistematika penulisannya adalah sebagai berikut:

1. BAB I PENDAHULUAN

   Bab pertama memuat pengantar penelitian, meliputi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

2. BAB II LANDASAN TEORI

   Bab kedua menyajikan landasan teori yang mendukung penelitian, mencakup klasifikasi teks, *Natural Language Processing*, metode *machine learning*, dan teknik evaluasi yang digunakan.

3. BAB III METODOLOGI PENELITIAN

   Bab ketiga menjelaskan metode penelitian, mulai dari pengumpulan data, *preprocessing*, ekstraksi fitur menggunakan TF-IDF, pembagian *dataset*, *cross validation*, *feature selection*, pelatihan model, *hyperparameter tuning*, hingga evaluasi.

4. BAB IV HASIL DAN PEMBAHASAN

   Bab keempat berisi hasil implementasi dan pengujian model, termasuk hasil *preprocessing*, ekstraksi fitur, pelatihan model, evaluasi dengan metrik *accuracy*, *precision*, *recall*, dan *F1-score*, serta analisis perbandingan performa kedua algoritma.

5. BAB V SIMPULAN DAN SARAN

   Bab kelima memuat simpulan penelitian dan saran yang dapat dipertimbangkan untuk pengembangan studi sejenis pada masa mendatang.