BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi digital telah mengubah berbagai aspek kehidupan, meliputi ekonomi, sosial, dan pendidikan. Kemajuan ini membawa tantangan baru, termasuk meningkatnya ancaman konten ujaran kebencian (*hate speech*) pada platform digital. Ujaran kebencian merujuk pada ekspresi berbasis teks, gambar, atau simbol yang menyebarkan kebencian, diskriminasi, atau permusuhan terhadap individu maupun kelompok berdasarkan atribut seperti ras, agama, atau etnis, sering kali memanfaatkan anonimitas di dunia maya [1, 2]. Dampaknya mencakup gangguan emosional, polarisasi sosial, dan dalam kasus tertentu, eskalasi konflik di dunia nyata.

Pendekatan *machine learning* telah banyak diterapkan dalam mendeteksi ujaran kebencian pada platform digital. Sejak awal kemunculan *Natural Language Processing* (NLP), metode tradisional seperti Bag-of-Words dan TF-IDF—serta penggunaan kamus khusus—telah digunakan untuk menganalisis teks [3, 4]. Pendekatan ini kemudian berkembang dengan penerapan algoritma klasifikasi tradisional seperti Support Vector Machine (SVM) yang menggunakan representasi fitur sederhana [5], dan untuk mengurangi kompleksitas vektor fitur yang tinggi, teknik reduksi dimensi seperti Principal Component Analysis (PCA) sering diterapkan untuk menghasilkan representasi yang lebih ringkas dan efisien, tetapi kurang mampu menangani analisis kompleks [6].

Platform media sosial, khususnya Platform X (Twitter), menjadi saluran utama penyebaran ujaran kebecian di Indonesia [7]. Dengan basis pengguna yang besar dan interaksi yang intens, konten agresif dapat menyebar dengan cepat melalui fitur retweet dan hashtag [8]. Penelitian terkini menunjukkan adanya peningkatan signifikan dalam paparan ujaran kebencian daring di Indonesia, dengan rasio yang meningkat sepuluh kali lipat dalam dua tahun terakhir, terutama menargetkan kelompok minoritas agama dan etnis [9]. Fenomena ini menimbulkan kekhawatiran besar terkait keamanan, privasi, serta kualitas interaksi di ruang digital, khususnya bagi generasi muda yang merupakan pengguna aktif media sosial.

Maraknya ujaran kebencian menuntut pendekatan yang lebih kompleks

untuk mendeteksi dan mencegah perilaku tersebut secara efektif. Salah satu teknik yang semakin mendapat perhatian adalah penggunaan model *Bidirectional Encoder Representations from Transformers* (BERT). Model ini mampu memahami konteks bahasa secara lebih mendalam dibandingkan metode tradisional, sehingga lebih akurat dalam mengidentifikasi kalimat ujaran kebencian yang rumit [10, 11]. Studi menunjukkan efektivitas BERT dalam mendeteksi ujaran kebencian, mencapai *F1-measure* 92% untuk bahasa Inggris pada dataset Davidson dan *macro-F1* 0,78 untuk bahasa Indonesia pada IndoToxic2024 [12, 9]. Selain itu, penelitian lainnya telah menunjukkan bahwa pendekatan berbasis BERT mampu meningkatkan performa deteksi *hate-speech* dengan tingkat akurasi yang lebih tinggi, menjadikannya solusi yang menjanjikan dalam analisis teks [13, 14, 15, 16].

Sebagai pengembangan dari pendekatan tersebut, penelitian ini mengusulkan model *hybrid* yang menggabungkan BERT dengan algoritma *Extreme Gradient Boosting* (XGBoost). Integrasi ini bertujuan untuk memanfaatkan kekuatan representasi semantik yang dihasilkan oleh BERT dan menggabungkannya dengan kemampuan klasifikasi XGBoost yang efisien dan cakap dalam menangani data berdimensi tinggi dan ketidakseimbangan kelas [17]. Dalam pendekatan ini, representasi vektor teks yang diekstraksi oleh BERT digunakan sebagai *input* bagi XGBoost dalam proses klasifikasi [18]. Dengan demikian, strategi *hybrid* menggabungkan pemahaman bahasa alami yang mendalam dengan performa klasifikasi yang tinggi, dan diharapkan mampu meningkatkan akurasi dalam mendeteksi ujaran kebencian, terutama dalam konteks sosial media berbahasa Indonesia seperti Platform X [17, 18, 19].

1.2 Rumusan Masalah

Berdasarkan pemaparan latar belakang masalah diatas, rumusan masalah yang dibahas dalam penulisan adalah sebagai berikut :

1. Bagaimana cara mendeteksi komentar tweet yang mengandung unsur ujaran kebencian.
2. Bagaimana performa model dalam membedakan komentar tweet berdasarkan *accuracy*, *precision*, *recall*, *F1-score*, dan *confffusion matrix* yang mengandung unsur ujaran kebencian.

1.3 Batasan Permasalahan

Pada bagian ini dijabarkan batasan yang diterapkan dalam penelitian ini agar pelaksanaan penelitian menjadi lebih terfokus kepada aspek-aspek berikut :

1. Penelitian hanya berfokus pada deteksi tweet ujaran kebencian.
2. Penelitian berfokus pada hasil model mengenai klasifikasi tweet ujaran kebencian.
3. Klasifikasi akan dibagi menjadi sentimen positif dan negatif.
4. Dataset yang digunakan terbatas pada hasil tweets yang diambil dari github.

1.4 Tujuan Penelitian

Berdasarkan permasalahan yang telah dirumuskan, berikut adalah tujuan yang telah dijabarkan :

1. Membuat model yang dapat mendeteksi komentar tweet yang mengandung unsur ujaran kebencian menggunakan pendekatan berbasis model *machine learning hybrid*.
2. Mengevaluasi performa model dalam membedakan komentar tweet yang mengandung ujaran kebencian berdasarkan *accuracy*, *precision*, *recall*, *F1-score*, dan *conffusion matrix*.

1.5 Manfaat Penelitian

Berdasarkan penelitian dalam implementasi BERT dan XGBoost untuk deteksi konten hate speech pada media sosial X, berikut adalah manfaat penelitian ini:

1. **Bagi Peneliti**
   (a) Menerapkan dan memperdalam pemahaman teori *machine learning* dan *Natural Language Processing* (NLP) melalui pengembangan serta evaluasi kombinasi algoritma BERT dan XGBoost untuk deteksi ujaran kebencian.

(b) Memberikan pengalaman praktis dalam mengintegrasikan model *hybrid* BERT dan XGBoost untuk menghasilkan analisis klasifikasi teks ujaran kebencian di media sosial.

2. **Bagi Pihak Lain**

(a) Mengidentifikasi konten ujaran kebencian secara lebih akurat dan efisien melalui pendekatan model *hybrid* BERT dan XGBoost.

(b) Menyediakan wawasan akademik mengenai penerapan model kombinasi BERT, XGBoost, dalam mendeteksi ujaran kebencian yang dapat menjadi dasar bagi penelitian lanjutan untuk mengoptimalkan akurasi dan performa model.

1.6 Sistematika Penulisan

Sistematika penulisan berupa struktur penulisan yang terdiri dari lima bab yang membahas aspek penting penelitian, yaitu :

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.
- Bab 2 LANDASAN TEORI
  Bab ini membahas teori-teori dan konsep-konsep yang relevan dengan topik penelitian, termasuk penelitian terdahulu yang mendukung dasar teori penelitian ini.
- Bab 3 METODOLOGI PENELITIAN
  Bab ini menjelaskan metode yang digunakan dalam penelitian, seperti pengumpulan data, perancangan sistem, teknik implementasi, serta tahapan-tahapan yang dilakukan dalam penelitian.
- Bab 4 HASIL DAN DISKUSI
  Bab ini menyajikan hasil penelitian, analisis data, serta penjelasan mengenai perbadingan hasil yang diberikan.
- Bab 5 SIMPULAN DAN SARAN
  Bab ini memuat kesimpulan dari hasil penelitian dan saran-saran yang dapat diberikan untuk pengembangan lebih lanjut atau penelitian selanjutnya.