BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Era digital kontemporer ditandai oleh ledakan konten yang dibuat oleh pengguna (user-generated content), terutama pada platform media sosial. Fenomena ini sangat menonjol di Indonesia, di mana masyarakatnya dikenal sangat aktif di dunia maya. Per Januari 2025, Indonesia memiliki 143 juta identitas pengguna media sosial, setara dengan 50,2 persen dari total populasi [1]. Dalam ekosistem digital yang padat ini, TikTok muncul sebagai salah satu platform paling favorit dan berpengaruh di Indonesia, dengan jumlah pengguna mencapai 108 juta jiwa (usia 18+) dan menjadikannya pasar TikTok terbesar kedua secara global [2]. Tingkat keterlibatan pengguna juga sangat tinggi; publik Indonesia menghabiskan hampir 45 jam setiap bulan di TikTok, melampaui rata-rata global [3]. Bahkan, Indonesia menempati posisi keenam secara global dalam hal durasi penggunaan TikTok, tidak jauh di bawah negara-negara teratas seperti Finlandia (54 jam 37 menit), Bulgaria (46 jam 9 menit), dan Kroasia (45 jam 35 menit). Volume konten yang masif pada platform tersebut, khususnya di kolom komentar, menciptakan tantangan komputasi yang signifikan dalam hal moderasi konten, di mana proses manual terbukti tidak lagi memadai dan efektif secara skalabilitas.

Di tengah lautan data tekstual ini, muncul masalah sosial yang mendesak: penyebaran ujaran kebencian berbasis gender atau seksisme. Komentar yang merendahkan martabat berdasarkan gender telah menciptakan lingkungan digital yang toksik, terutama bagi perempuan. Fenomena ini bukan sekadar masalah etika, tetapi telah menjadi krisis terukur. Data dari Kementerian Pemberdayaan Perempuan dan Perlindungan Anak (KemenPPPA) menunjukkan lonjakan kasus Kekerasan Berbasis Gender Online (KBGO) sebesar 400% pada triwulan pertama 2024, yang berdampak serius pada kesehatan psikologis dan keamanan korban [4].

Pada tahap awal penelitian deteksi konten berbahaya, khususnya untuk Bahasa Indonesia, banyak peneliti mengandalkan algoritma machine learning konvensional. Sebagai contoh, studi fundamental oleh Ibrohim dan Budi (2019) membangun dataset ujaran kebencian dan mengujinya menggunakan Naive Bayes dengan representasi fitur TF-IDF [5]. Pendekatan seperti ini, yang mengandalkan feature engineering manual, menjadi baseline penting pada masanya. Namun,

keterbatasan utama dari metode ini adalah ketidakmampuannya untuk menangkap makna kontekstual yang mendalam. Model-model konvensional kesulitan membedakan antara kata yang digunakan secara harfiah dan yang digunakan dalam konteks sarkasme atau ironi, sehingga seringkali gagal memahami nuansa yang kompleks dalam percakapan di media sosial. Keterbatasan inilah yang memicu pengembangan model representasi bahasa yang lebih canggih, yang mampu belajar representasi bidireksional yang mendalam dengan mengkondisikan konteks kiri dan kanan di semua lapisan, seperti yang diperkenalkan oleh arsitektur Transformer [6].

Sebagai jawaban atas keterbatasan tersebut, penelitian di bidang NLP beralih ke model deep learning berbasis arsitektur Transformer. Model ini terbukti unggul karena kemampuannya memahami konteks kata secara mendalam. Penelitian oleh Kusuma dan Chowanda (2023) berhasil menerapkan IndoBERTweet yang dikombinasikan dengan BiLSTM untuk mendeteksi ujaran kebencian di Twitter dengan hasil akurasi 93.7% [7]. Studi lain oleh Darmawan et al. (2023) juga membuktikan keunggulan IndoBERT untuk klasifikasi ujaran kebencian multi-label dengan akurasi makro mencapai 88,23% [8], sementara Wijanarko et al. (2024) mencapai skor Macro F1 sebesar 0.718 menggunakan IndoBERTweet untuk tugas serupa [9]. Hasil-hasil ini mengonfirmasi bahwa model pra-latih seperti IndoBERT merupakan fondasi yang kuat untuk menganalisis teks berbahasa Indonesia di media sosial.

Namun, tinjauan terhadap literatur yang ada menunjukkan adanya beberapa celah penelitian (research gap) yang signifikan. Pertama, fokus utama riset di Indonesia lebih banyak tertuju pada deteksi ujaran kebencian (hate speech) secara umum [5], [7], [8], [9]. Padahal, seksisme memiliki karakteristik linguistik yang lebih subtil dan bernuansa, sehingga memerlukan pendekatan yang lebih spesifik. Kedua, platform yang dianalisis dalam penelitian-penelitian tersebut dominan adalah Twitter [7], sementara TikTok dengan karakteristik demografi dan linguistik komentarnya yang unik masih kurang dieksplorasi. Ketiga, sementara penelitian internasional telah fokus secara spesifik pada deteksi seksisme seperti yang dilakukan oleh Bremm et al. (2024) untuk bahasa Jerman [10] dan Fudulu et al. (2023) untuk bahasa Inggris [11] penerapan dan evaluasi model serupa pada konteks Bahasa Indonesia masih sangat terbatas.

Dampak dari konten seksis sangat serius, mulai dari gangguan psikologis, trauma, hingga pengucilan sosial [12]. Moderasi manual terbukti tidak memadai [13], sementara laporan dari pengguna sering tidak ditindaklanjuti dengan konsisten [14]. Oleh karena itu, pendekatan otomatis seperti Natural Language

Processing (NLP) menjadi penting dalam mendeteksi ujaran seksis di media sosial. Model berbasis transformer seperti BERT telah terbukti unggul dalam memahami nuansa ujaran kebencian dan seksisme, termasuk ketika ekspresinya bersifat implisit atau menggunakan ironi [15]. Untuk konteks bahasa Indonesia, model seperti IndoBERT yang dilatih secara khusus dengan korpus lokal menunjukkan performa yang sangat baik dalam tugas klasifikasi teks sosial media [16]. dengan model transformer seperti BERT menunjukkan hasil unggul dalam memahami nuansa ujaran seksis [15].

Berdasarkan celah penelitian tersebut, penelitian ini diajukan untuk menjawab tantangan yang ada dengan mengaplikasikan metode komputasi pada domain yang spesifik dan relevan. Platform TikTok dipilih karena posisinya yang sangat dominan dengan tingkat keterlibatan pengguna yang masif di Indonesia [1], namun karakteristik linguistik komentarnya dalam konteks deteksi konten berbahaya masih kurang tereksplorasi. Fokus pada seksisme diambil karena urgensinya sebagai salah satu bentuk Kekerasan Berbasis Gender Online (KBGO) yang dampaknya terus meningkat secara signifikan di Indonesia [4]. Terakhir, pemilihan model IndoBERT sebagai inti dari solusi teknologi didasarkan pada performanya yang telah terbukti andal dan unggul untuk berbagai tugas pemrosesan Bahasa Indonesia [7]. Dengan demikian, penelitian ini memberikan kontribusi unik dengan menguji dan mengevaluasi secara sistematis sebuah model state-of-the-art pada kombinasi masalah (seksisme), platform (TikTok), dan bahasa (Indonesia) yang belum banyak tersentuh oleh riset sebelumnya.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana cara mengidentifikasi komentar TikTok berbahasa Indonesia yang mengandung seksime?
2. Seberapa baik performa model IndoBERT dalam mendeteksi komentar seksis pada TikTok berdasarkan metrik evaluasi klasifikasi seperti akurasi, presisi, recall, dan f1-score?

1.3 Batasan Permasalahan

Agar penelitian ini lebih terarah dan fokus, maka ruang lingkup penelitian ini dibatasi sebagai berikut:

1. Sumber Data: Penelitian ini difokuskan pada analisis data teks dari komentar publik di platform TikTok berbahasa Indonesia. Aspek multimodal seperti analisis video atau audio tidak termasuk dalam cakupan penelitian.
2. Skema Pelabelan: Tugas klasifikasi yang dilakukan adalah klasifikasi biner, yaitu melabeli komentar sebagai ”seksis” atau ”tidak seksis”. Penelitian ini tidak mencakup klasifikasi multi-kelas untuk berbagai jenis kategori seksisme.
3. Arsitektur Model: Model utama yang diimplementasikan adalah versi pra-latih indobenchmark/indobert-base-p1. Penelitian ini berfokus pada proses fine-tuning terhadap model tersebut tanpa melakukan modifikasi pada arsitektur dasarnya.
4. Pra-Pemrosesan Teks: Proses pra-pemrosesan teks difokuskan pada normalisasi kata tidak baku (slang) dan pembersihan umum. Penelitian ini tidak menangani fenomena linguistik yang lebih kompleks seperti majas atau code-switching dengan bahasa daerah.
5. Lingkup Optimasi: Proses hyperparameter tuning difokuskan pada parameter kunci seperti learning rate, umlah epoch, batch size, droup out dan tidak mencakup eksplorasi seluruh kemungkinan hyperparameter lainnya.
6. Metrik Evaluasi: Kinerja model diukur menggunakan metrik evaluasi standar untuk klasifikasi biner, yaitu akurasi, presisi, recall, dan F1-score.
7. Penanganan Ambiguitas: Komentar yang bersifat ambigu atau sarkastik tetap menjadi bagian dari dataset, namun penelitian ini tidak melakukan analisis pragmatik mendalam untuk menangani kasus-kasus tersebut secara khusus.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mencapai hasil yang sesuai dengan rumusan masalah yang telah disampaikan sebelumnya. Adapun tujuan penelitian ini adalah sebagai berikut:

1. Mengidentifikasi komentar seksis pada video TikTok berbahasa Indonesia menggunakan model IndoBERT.
2. Mengukur validitas dan performa algoritma IndoBERT dalam mendeteksi komentar seksis pada platform TikTok berdasarkan metrix evaluasi yang didefinsiikan seperti accuracy, f1 score, predictiion dan recall

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sesuai dengan lingkup kerja yang telah ditentukan. Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Manfaat Teoritis
   (a) Hasil penelitian ini bisa menjadi referensi atau acuan bagi mahasiswa atau peneliti lain yang ingin melakukan penelitian serupa di masa depan, khususnya dalam bidang *Natural Language Processing* untuk Bahasa Indonesia.
   (b) Menambah pengetahuan tentang seberapa baik kemampuan model IndoBERT dalam mendeteksi komentar seksis yang menggunakan bahasa informal seperti di TikTok.
2. Manfaat Praktis
   (a) Model yang dikembangkan dalam penelitian ini dapat menjadi dasar bagi platform media sosial seperti TikTok untuk mengembangkan fitur deteksi komentar seksis secara otomatis.
   (b) Bagi pengembang *developer*, penelitian ini bisa menjadi contoh langkah-langkah teknis untuk membuat sistem sejenis.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian mengenai deteksi komentar seksis dari video Tiktok menggunakan Indobert disusun secara sistematis agar pembahasan tersaji secara runtut dan tersrtruktur. Adapun penulisan laporan ini adalah sebagai berikut:

1. Bab 1 PENDAHULUAN

   Bab 1 berisi latar belakang penelitian, rumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan penelitian.

2. Bab 2 LANDASAN TEORI

   Bab 2 berisi landasan teori penelitian. Pada bab ini dijelaskan semua istilah dan algortima yang digunakan dalam penelitian terkait deteksi komentar seksis menggunakan Indobert.

3. Bab 3 METODOLOGI PENELITIAN

   Bab 3 menjelaskan terkait metodologi penelitian. sistematika penelitian dijelaskan pada bab ini, termasuk flowchart sistem, sistematika setiap proses algortima berjalan dalam pembuatan model maupun pengumpulan data untuk mendukung model.

4. Bab 4 HASIL DAN DISKUSI

   Pada bab 4 dituliskan hasil pembuatan model dan analisis dengan hopotesa awal.

5. Bab 5 KESIMPULAN DAN SARAN

   Pada bab 5 dijelaskan hasil dari penelitiian yang dilakukan berdasarkan masalah, tujuan dan manfaat yang sudah ditentukan batasan-batasan pada bab sebelumnya.