BAB I
PENDAHULUAN

1.1. Latar Belakang

Pesatnya jangkauan internet di Indonesia telah mengubah cara masyarakat berkomunikasi, namun juga memunculkan berbagai perilaku negatif di ruang digital[1]. Salah satu fenomena yang kini menjadi perhatian serius adalah Kekerasan Berbasis Gender Online (KBGO), khususnya di media sosial Instagram[2][3]. Berbeda dengan ujaran kebencian secara umum, kekerasan berbasis gender memiliki karakteristik yang lebih destruktif karena melibatkan pelecehan verbal, objektifikasi, hingga intimidasi yang menyerang identitas gender perempuan[4]. Fenomena ini bukan sekadar perilaku toksik biasa, melainkan ancaman nyata terhadap partisipasi aman individu di ruang publik digital.

Upaya penanganan KBGO menghadapi tantangan besar akibat keterbatasan manual dan sulitnya memfilter konten secara spesifik [5][6]. Dampak dari sulitnya memfilter konten tersebut tidak hanya berhenti pada masalah administratif, tetapi meluas menjadi masalah sosial yang berdampak nyata bagi korban[7]. Normalisasi pelecehan verbal di media sosial menciptakan iklim ketakutan bagi korban, sehingga diperlukan sistem pengawasan otomatis yang mampu membedakan opini kritis dengan ujaran kekerasan diskriminatif. Namun, karakteristik bahasa media sosial Indonesia yang penuh slang, sarkasme dan ambiguitas menjadi kendala tersendiri[8]. Metode klasifikasi tradisional kerap kehilangan konteks, sementara pendekatan *Deep Learning* seperti *Long Short-Term Memory* terbukti lebih unggul dalam mengenali pola kalimat (Pasuruan, U.M). Meski demikian, kekerasan berbasis gender menuntut pemahaman konteks linguistik yang lebih mendalam, karena makna ujaran sangan bergantung pada konteks kalimat [9].

Oleh karena itu dengan memanfaatkan arsitektur *Deep Learning* yang lebih sensitif terhadap konteks bahasa Indonesia, penelitian ini diharapkan mampu menghasilkan model deteksi yang lebih akurat dibandingkan penelitian-penelitian sebelumnya[10]. Melalui judul "Deteksi Ujaran Kekerasan Berbasis Gender Terhadap Perempuan di Media Sosial Instagram Menggunakan Algoritma IndoBERT", penelitian ini diharapkan mampu memberikan kontribusi nyata dalam menciptakan ekosistem media sosial yang lebih aman, inklusif, dan bermartabat.

1.1.1. Defenisi Operasional Variabel Penelitian

Dalam penelitian ini KBGO didefenisikan sebagai ujaran tertulis pada media sosial Instagram yang mengandung unsur merendahkan, melecehkan, mengancam atau kebencian terhadap perempuan, baik secara langsung maupun tersirat, serta ditujukan kepada individu maupun kelompok. Objek analisis berupa teks komentar. Komentar dikategorikan sebagai KBGO apabila memuat indikator bahasa kekerasan berbasis gender, sedangkan komentar kasar yang tidak berkaitan dengan gender diklasifikasikan sebagai Non-KBGO.

Tabel 1. Tabel Indikator

Kategori	Indikator	Kriteria	Contoh	Klasifikasi
Seksisme	Stereotip peran gender	Teks mengandung pelabelan negatif berdasarkan peran sosial gender	“Perempuan itu gacocok jadi pemimpin”	KBGO
Pelecehan seksual	Referensi seskual non-konsensual	Penyebutan aktivitas/atribut seksual tanpa persetujuan	“Liat aja pakaiannya, pantes digodain”	KBGO
Objektifikasi	Reduksi nilai pada tubuh	Individu direduksihanya pada aspek fisik/seksual	“Yang penting badannya, soal otak belakangan”	KBGO
Ancaman berbasis gender	Ancaman kekerasan seksual	Intimidasi eksplisit/implisit berbasis gender	“Perempuan kaya gitu harus dikasi Pelajaran”	KBGO
Kebencian berbasis gender	Generalisasi kolektif	Menyerang kelompok gender secara kolektif	“Semua Perempuan sama aja, sukabuat masalah”	KBGO
Penghinaan umum	Tanpa referensi gender	Ujaran kasar tanpa atribut gender	”Bodoh kau”	Non-KBGO

1.2. Rumusan Masalah

1. Bagaimana cara membangun arsitektur model *Deep Learning* yang mampu mengenali karakteristik unik dan konteks linguistik dari ujaran kekerasan berbasis gender di media sosial?
2. Sejauh mana penggunaan model *pre-trained* berbasis *Transformers* (seperti IndoBERT) dapat meningkatkan akurasi deteksi pada data teks yang mengandung bahasa *slang*, singkatan, dan sarkasme dibandingkan dengan metode konvensional?
3. Bagaimana efektivitas model dalam membedakan secara presisi antara ujaran kebencian umum dengan kekerasan yang secara spesifik menargetkan identitas gender?

1.3. Tujuan Penelitian

1. Mengembangkan sistem klasifikasi otomatis berbasis *Deep Learning* untuk mendeteksi ujaran kekerasan berbasis gender pada teks berbahasa Indonesia.
2. Mengevaluasi performa model melalui pengujian metrik *Accuracy, Precision, Recall,* dan *F1-Score* guna memastikan reliabilitas dalam mengatasi teks yang multitafsir.
3. Menghasilkan kontribusi berupa dataset terlabeli dan model komputasi yang dapat mengenali pola-pola objektifikasi serta intimidasi gender di platform media sosial.

1.4. Manfaat Penelitian

1. Bagi Peneliti: Sebagai sarana Penerapan teknologi NLP untuk menangani permasalahan sosial serta meningkatkan kompetensi dalam pengembangan model *Deep Learning*.
2. Bagi Masyarakat: Memberikan kontribusi dalam menciptakan ruang digital yang lebih aman, khususnya bagi perempuan dengan mengurangi dampak kekerasan verbal di media sosial.
3. Bagi Pengembang Platform : Menjadi referensi teknis dalam pengembangan sistem moderasi konten otomatis yang lebih cerdas dan sensitif terhadap isu gender.

1.5. Batasan Masalah

1. Sumber dan Jenis Data: Data diperoleh dari media sosial melalui teknik *crawling*, terbatas pada teks berbahasa Indonesia, termasuk bahasa slang atau tidak baku.
2. Format Konten: Analisis hanya mencakup data teks dan tidak melibatkan konten multimedia seperti gambar, audio dan video.

3. Ruang Lingkup Kategorisasi: Penelitian difokuskan pada ujaran kekerasan berbasis gender, kategori ujaran kebencian lain seperti SARA, politik atau *body shaming* tidak dianalisis kecuali memiliki keterkaitan langsung dengan kekerasan gender.

1.6. Aspek Etika dan Bias

1. Bias Gender dalam Data

   Konten Instagram berpotensi memuat bias sosial, dengan perempuan lebih sering menjadi sasaran pelecehan. Penelitian ini, meninimalkan bias melalui KBGO yang jelas.
2. Risiko Kesalahan Klasifikasi (*False Positive*)

   Model berpotensi mengklasifikasikan kritik, atau *satire* sebagai KBGO, sehingga hasil deteksi tidak bersifat mutlak dan digunakan sebagai alat bantu moderasi.
3. Dampak Sosial

   Kesalahan klasifikasi dapat memengaruhi kebebasan berekspresi, oleh karena itu sistem dikembangkan sebagai pendukung pengambilan keputusan.

1.7. Keterbaharuan

1. Berbeda dengan penelitian [5] yang membangun dataset ujaran kebencian secara multi-label, penelitian ini secara khusus memfokuskan kajian pada kekerasan berbasis gender yang menargetkan perempuan. Pendekatan spesifik ini memungkinkan perumusan defenisi operasional dan dataset terlabeli yang lebih terarah, sehingga deteksi KBGO menjadi lebih presisi terhadap ujaran yang ambigu dan tersirat.
2. Penelitian [11] masih menggunakan metode konvensional seperti *Naïve Bayes* dan SVM yang kurang mampu menangkap makna implisit. Penelitian ini menghadirkan kebaruan pada pendalaman pemahaman konteks ujaran KBGO melalui pemanfaatan representasi semantik IndoBERT, khususnya dalam membedakan ujaran kebencian umum dan kekerasan berbasis gender.
3. Penerapan LTSM yang memiliki keterbatasan dalam merepresentasikan hubungan semantik jangka panjang [12]. Sebagai kebaruan, penelitian ini menggunakan arsitektur *Transformer* berbasis IndoBERT dengan mekanisme *self-attention* serta melakukan pengujian *baseline* sehingga peningkatan performa dibuktikan melalui evaluasi metrik yang terukur.