

BAB I

PENDAHULUAN

1.1 Latar Belakang

Di era digital saat ini, media sosial telah bertransformasi menjadi ruang publik yang sangat dinamis, tempat masyarakat secara aktif menyuarakan pandangan mereka mengenai berbagai isu strategis nasional. Fenomena ini menciptakan aliran data percakapan digital yang masif dan terjadi secara terus-menerus. Namun, derasnya arus informasi tersebut memicu tantangan baru; metode pemantauan konvensional berbasis peninjauan manual kini kewalahan dalam menangkap esensi emosi, sarkasme, dan konteks tersirat dari setiap opini yang dilontarkan [1]. Ketidakmampuan tersebut berdampak langsung pada lambatnya proses evaluasi kebijakan serta besarnya risiko kekeliruan dalam memetakan suara masyarakat secara utuh. Sayangnya, di tengah kompleksitas bahasa digital yang ada, instrumen pemantauan yang digunakan saat ini umumnya masih bertumpu pada pendekatan tradisional seperti analisis leksikon statis atau pembelajaran mesin konvensional [1]. Metode-metode lama ini memiliki keterbatasan besar karena cenderung kaku dan gagal memahami struktur bahasa tidak baku serta variasi dialek lokal yang justru menjadi corak utama komunikasi masyarakat di media sosial [1].

Penelitian terdahulu yang pertama membahas tentang perkembangan platform media sosial telah menggeser paradigma komunikasi publik menjadi lebih interaktif dan masif, khususnya di Indonesia yang mencatatkan penetrasi pengguna aktif TikTok lebih dari 160 juta dengan rata-rata durasi penggunaan 44 jam per bulan [2]. Data kuantitatif ini menjadi fondasi penting untuk memahami bahwa

kolom komentar TikTok bukan sekadar fitur interaksi, melainkan repositori opini publik yang autentik dan *real-time*. Kelebihan dari penelitian ini adalah sumber data tersebut terletak pada akurasi statistik makro yang terverifikasi secara berkala, sehingga dapat dijadikan acuan valid dalam menentukan urgensi penelitian berbasis platform digital [2]. Kekurangan dari penelitian ini adalah data tersebut hanya menyajikan metrik penggunaan agregat tanpa memberikan gambaran mendalam mengenai dinamika percakapan, pola sentimen, atau konteks linguistik yang terbentuk di tingkat komentar [2].

Penelitian terdahulu yang kedua membahas tentang fenomena judi *online* yang telah menjadi krisis multidimensi di Indonesia dengan jumlah pelaku yang melonjak drastis hingga 8,8 juta pada tahun 2024 serta perputaran dana yang mencapai ratusan triliun rupiah, sehingga mendorong pemerintah mengintensifkan instrumen pemblokiran dan penindakan hukum [3]. Kelebihan dari penelitian ini berupa pemetaan data intelijen dan dampak sosio-ekonomi yang komprehensif, sehingga memperkuat argumen bahwa judi *online* merupakan isu nasional yang memerlukan respons berbasis bukti [3]. Namun, penelitian ini memiliki kekurangan karena lebih berfokus pada aspek penegakan hukum, aliran dana, dan dampak demografis, sehingga mengabaikan perspektif publik yang terekam secara organik di ruang digital. Selain itu, kekurangan lainnya adalah belum dilakukannya analisis sentimen masyarakat terhadap isu judi *online* yang berkembang pada platform media sosial, khususnya TikTok yang memiliki tingkat interaksi pengguna yang tinggi. Ketiadaan analisis terhadap opini publik menyebabkan belum tersedianya informasi mengenai kecenderungan sikap masyarakat, baik yang mendukung,

menolak, maupun bersikap netral terhadap fenomena judi *online*. Oleh karena itu, diperlukan penelitian yang mampu mengidentifikasi dan mengklasifikasikan sentimen masyarakat melalui komentar pengguna media sosial dengan memanfaatkan metode pemrosesan bahasa alami (Natural Language Processing/NLP), seperti fine-tuning IndoBERT, sehingga dapat memberikan gambaran yang lebih mendalam mengenai respons publik terhadap isu judi *online* [3].

Penelitian terdahulu yang ketiga membahas tentang untuk merespons kebutuhan pemetaan opini publik secara otomatis, pendekatan *Natural Language Processing* (NLP) telah banyak diadopsi, termasuk pengembangan model *deep learning* yang diklaim mampu menangani karakteristik bahasa informal dan slang pada media sosial [4]. Kelebihan dari pendekatan tersebut terbukti dalam kemampuan menangkap pola sekuensial dan konteks lokal yang sulit diolah oleh metode berbasis kamus statis, sehingga menghasilkan akurasi klasifikasi yang lebih konsisten pada teks tidak baku [4]. Kekurangannya terletak pada cakupan domain yang masih umum, di mana model belum diuji secara spesifik pada *dataset* berisiko tinggi dengan polarisasi emosional kuat seperti isu judi *online*, serta belum mempertimbangkan karakteristik unik platform TikTok yang didominasi oleh format komentar sangat pendek dan campuran kode bahasa [4].

Penelitian terdahulu yang keempat membahas tentang algoritma pembelajaran mesin konvensional seperti *Naïve Bayes* dan *Support Vector Machine* (SVM) memang banyak digunakan untuk klasifikasi sentimen teks Indonesia dan memberikan hasil yang cukup baik pada data terstruktur atau semi-terstruktur [5].

Kelebihan dari penelitian ini adalah efisiensi komputasi yang tinggi, kemudahan implementasi, serta transparansi dalam proses pengambilan keputusan berbasis fitur TF-IDF atau *n-gram* [5]. Namun, kekurangannya adalah ketergantungan pada rekayasa fitur manual yang membuat model sangat rentan terhadap *noise*, sarkasme, dan variasi linguistik tidak baku yang umum ditemukan pada komentar media sosial, sehingga performa metrik *F1-score* sering menurun drastis ketika dihadapkan pada data nyata yang kompleks [5].

Penelitian terdahulu yang kelima membahas tentang penelitian yang menerapkan arsitektur Transformer berbasis IndoBERT pada komentar TikTok telah menunjukkan potensi besar dalam memetakan persepsi publik terhadap isu kebijakan atau adopsi teknologi di Indonesia [6]. Kelebihan utama dari studi terdahulu tersebut adalah keberhasilan membuktikan bahwa model pra-latih monolingual mampu mengungguli pendekatan multibahasa dalam menangkap nuansa semantik dan slang khas Indonesia pada platform video pendek [6]. Kekurangannya adalah fokus penelitian masih terbatas pada topik komersial atau layanan publik yang relatif netral, belum menyentuh domain sensitif dengan tingkat *noise* dan bias kelas yang ekstrem, serta belum menyertakan strategi *fine-tuning* yang dioptimalkan khusus untuk menjaga stabilitas representasi vektor pada data yang sangat tidak seimbang [6].

Penelitian terdahulu yang keenam membahas tentang identifikasi kesenjangan tersebut, penelitian ini mengajukan penerapan *fine-tuning* model IndoBERT-base-p1 yang secara khusus dikembangkan untuk korpus bahasa Indonesia dengan arsitektur *encoder* 12 lapisan dan mekanisme perhatian

bidireksional [7]. Kelebihan model ini terletak pada kapasitasnya yang telah tervalidasi secara empiris dalam memahami sintaksis, konteks global, dan variasi dialek lokal tanpa memerlukan pelatihan dari nol, sehingga sangat efisien secara komputasi untuk tugas klasifikasi spesifik [7]. Kekurangannya adalah performa optimal hanya dapat dicapai apabila model diadaptasi menggunakan *dataset* beranotasi berkualitas tinggi dengan protokol pelabelan yang ketat, yang menjadi alasan mengapa penelitian ini merancang *pipeline* validasi *dataset*, penanganan ketidakseimbangan kelas, dan evaluasi metrik makro yang terstruktur sebelum implementasi akhir [7].

Berdasarkan identifikasi kesenjangan penelitian dan urgensi kebutuhan pemetaan opini publik tersebut, maka penelitian ini difokuskan pada pengembangan model klasifikasi yang adaptif, terukur, dan sesuai dengan karakteristik linguistik data komentar TikTok. Pemilihan judul “Analisis Sentimen Komentar TikTok Terhadap Judi *Online* Di Indonesia Menggunakan *Fine-Tuning* IndoBERT” dilatarbelakangi oleh tiga pertimbangan utama: (1) volume data yang masif dan dinamis menuntut pendekatan otomatis berbasis *deep learning* yang mampu memproses skala besar secara efisien; (2) dominasi bahasa informal, slang, dan *code-mixing* pada platform TikTok memerlukan model pra-latih monolingual yang telah tervalidasi secara empiris untuk konteks bahasa Indonesia; serta (3) domain judi *online* yang sensitif dan belum banyak dieksplorasi secara komputasional di Indonesia, sehingga hasil analisis dapat menjadi instrumen pendukung kebijakan berbasis data (*evidence-based policy*). Dengan demikian, penelitian ini tidak hanya bertujuan menghasilkan sistem analisis sentimen yang

akurat, tetapi juga menyediakan kontribusi metodologis dalam adaptasi arsitektur *Transformer* untuk platform media sosial berbasis video pendek serta referensi empiris bagi pengembangan *Natural Language Processing* (NLP) domain-spesifik di Indonesia.

1.2 Rumusan Masalah

Rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana merancang dan mengimplementasikan proses *fine-tuning* model IndoBERT untuk klasifikasi sentimen pada komentar TikTok berbahasa Indonesia terkait judi *online*?
2. Bagaimana tingkat performa model IndoBERT hasil *fine-tuning* apabila diukur menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*?
3. Bagaimana distribusi dan pola sentimen publik terhadap isu judi *online* di Indonesia berdasarkan hasil klasifikasi model pada *dataset* 99.065 komentar TikTok?

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Input Data membahas tentang hanya memproses 99.065 komentar teks TikTok bertopik judi *online* berbahasa Indonesia. Data non-teks (video, audio, gambar, *caption*) tidak diproses.
2. Model Pemrosesan membahas tentang menggunakan arsitektur IndoBERT-*base-p1* dengan teknik *fine-tuning*. Penelitian tidak mencakup komparasi eksperimental dengan model pembelajaran mesin lain.

3. *Output Sistem* membahas tentang menghasilkan klasifikasi sentimen ke dalam tiga kelas: Positif, Netral, dan Negatif. Analisis emosi granular atau subjek-objek spesifik tidak dilakukan.
4. Filter Linguistik membahas tentang teks yang diterima mencakup variasi formal, informal, slang, dan *code-mixing*. Komentar berbahasa asing murni dan terindikasi pola otomatis (*bot*) disaring pada tahap *preprocessing*.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Merancang dan mengimplementasikan *pipeline fine-tuning* model IndoBERT untuk klasifikasi sentimen pada komentar TikTok berbahasa Indonesia terkait judi *online*.
2. Mengukur dan mengevaluasi tingkat performa model IndoBERT hasil *fine-tuning* menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*.
3. Mengidentifikasi distribusi dan menganalisis pola sentimen publik terhadap isu judi *online* di Indonesia berdasarkan hasil klasifikasi model pada *dataset* 99.065 komentar TikTok.

1.5 Manfaat Penelitian

Manfaat penelitian ini dilihat dari dua aspek yaitu manfaat teoritis dan manfaat praktis sebagai berikut:

A. Manfaat Teoritis

1. Memperkaya referensi akademik mengenai penerapan arsitektur *Transformer* (IndoBERT) pada teks media sosial informal berbahasa Indonesia, khususnya untuk domain isu sosial sensitif.

2. Menyediakan korpus *dataset* berukuran 99.065 komentar TikTok bertopik judi *online* yang telah dilabeli, yang dapat dijadikan *benchmark* terbuka untuk penelitian *Natural Language Processing* (NLP) selanjutnya.
3. Memberikan bukti empiris mengenai efektivitas teknik *fine-tuning* dalam menangani tantangan linguistik khas media sosial Indonesia, seperti slang, singkatan, dan *code-mixing*.

B. Manfaat Praktis

1. Bagi Pemerintah dan Regulator
Menyediakan pemetaan sentimen publik berbasis data sebagai instrumen pendukung kebijakan (*evidence-based policy*) untuk mengevaluasi efektivitas komunikasi publik dan strategi pemberantasan judi *online*.
2. Bagi Akademisi dan Peneliti Sistem Informasi
Menyediakan panduan metodologis dan konfigurasi *pipeline* yang dapat direplikasi (*reproducible research*) untuk pengembangan sistem analisis sentimen pada isu-isu publik lainnya.
3. Bagi Masyarakat dan Organisasi Sipil
Meningkatkan kesadaran publik mengenai persepsi masyarakat terhadap dampak judi *online* berdasarkan data autentik, serta menjadi dasar advokasi berbasis data untuk pencegahan dan edukasi digital.