

BAB I PENDAHULUAN

1.1 Latar Belakang

Sektor pariwisata merupakan salah satu pilar penting dalam perekonomian Indonesia, dengan potensi besar untuk mendukung pertumbuhan dan perkembangan suatu negara. Sektor yang dimana sering diminati untuk mencari sensasi di luar aktivitas sehari-hari[1]. Salah satunya adalah di Kabupaten Sumenep yang dikenal memiliki potensi wisata yang melimpah, khususnya pada wisata pantai. Pantai menjadi daya tarik utama yang bisa mengidentifikasi faktor yang penting bagi pengunjung lokal maupun wisatawan mancanegara[2]. Di era digital saat ini, dimana lebih dari 78% penduduk Indonesia telah terhubung ke internet[3]. Penyampaian informasi terkait tempat wisata memerlukan media untuk menyampaikan. Kemudahan mencari atau melihat bagaimana pandangan yang diberikan pengunjung terhadap lokasi wisata yang ingin dikunjungi[1]. Proses pengambilan keputusan wisatawan sangat dipengaruhi oleh ulasan daring (*online reviews*) yang tersedia di berbagai platform seperti *Google Maps* [1][4][5][6]. Ulasan-ulasan ini menjadi sumber data yang kaya untuk memahami persepsi dan sentimen publik terhadap kualitas destinasi wisata[4]. Analisis terhadap data ini dapat memberikan wawasan berharga bagi pengelola wisata untuk meningkatkan kualitas layanan dan fasilitas.

Namun, data ulasan yang bersifat *User-Generated Content(UGC)* sering kali tidak terstruktur dan mengandung banyak tantangan, kemudahan dan kecepatan dalam berekspresi di media sosial seringkali menyebabkan pengguna lalai atau sengaja menyingkat kata, yang mengakibatkan banyaknya kesalahan ketik (*typo*), bahasa informal, dan singkatan yang unik untuk konteks Indonesia[7][8]. Analisis manual ulasan ini terbukti memakan waktu dan memiliki potensi untuk memperkenalkan bias[9]. Tantangan ilmu bahasa ini membuat analisis manual menjadi tidak efisien, sementara metode komputasi tradisional seringkali gagal menangkap nuansa makna dengan akurat, sehingga berpotensi menurunkan keakuratan klasifikasi sentimen[7]. Karena kesulitan menangani struktur ilmu bahasa yang canggih seperti ironi, sarkasme, dan negasi, serta salah menafsirkan arti dan hubungan makna antar kata[9].

Untuk mengatasi konteksitas bahasa ini, bidang *Natural Language Processing* (NLP) dan *Text Mining* telah berkembang pesat. Perkembangan model berbasis *Transformer* seperti *Bidirectional Encoder Representations from Transformers* (*BERT*) telah merevolusi cara mesin memahami konteks bahasa[10]. *IndoBERT*, sebagai model *BERT* yang telah dilatih secara khusus pada korpus besar Bahasa Indonesia, menunjukkan performa yang superior untuk berbagai tugas NLP, termasuk analisis sentimen, dan telah berhasil diterapkan pada domain pariwisata [11]. Akan tetapi, performa model canggih seperti *IndoBERT* seringkali masing sangat bergantung pada kualitas data masukan. Pentingnya pra-pemrosesan teks untuk mengubah teks mentah yang tidak terstruktur menjadi bentuk yang dapat dianalisis model sangat ditekankan. Adanya kesalahan ejaan dapat menyebabkan token tidak dikenali yang akan direpresentasikan sebagai token khusus oleh model[9] dan menurunkan akurasi model secara signifikan.

Oleh karena itu, diperlukan sebuah mekanisme koreksi ejaan cerdas sebagai bagian dari tahapan. Penelitian ini penggunaan metode *Damerau-Levenshtein Distance*, yang terbukti efektif untuk mengidentifikasi kandidat koreksi dengan menghitung jarak edit minimum antar *string*, termasuk operasi transposisi yang umum terjadi pada kesalahan ketik manusia[12][8]. Untuk memastikan koreksi yang terpilih sesuai dengan konteks kalimat, metode ini akan dikombinasikan dengan model probabilistik *N-Gram*. *N-Gram* akan menilai kandidat kata koreksi berdasarkan probabilitas kemunculannya dalam sebuah urutan kata, sehingga kata yang paling relevan secara kontekstual yang akan dipilih [13][14]. Kombinasi ini terbukti mampu meningkatkan akurasi klasifikasi secara signifikan[15].

Selain tantangan kualitas teks, kendala lain dalam pengembangan model untuk data berlabel. Melakukan pelabelan ribuan ulasan secara manual memakan waktu yang lama dan rentan terhadap inkonsistensi antar *annotator*. Untuk mengatasi tantangan efisiensi ini, penelitian menerapkan pendekatan pelabelan semi-otomatis menggunakan *Zero-Shot Classification* berbasis model *XNLI* [16]. Metode ini digunakan untuk memberikan label awal pada data ulasan secara otomatis, yang kemudian divalidasi kembali oleh manusia untuk memastikan akurasi label sebelum digunakan dalam pelatihan model[17].

Analisis sentimen yang hanya menghasilkan label positif atau negatif secara umum kurang memberikan wawasan yang dapat ditindaklanjuti. Oleh karena itu, penelitian ini menerapkan pendekatan analisis sentimen multi-aspek, yang memungkinkan identifikasi sentimen pada aspek-aspek pariwisata seperti daya tarik, aksesibilitas, fasilitas dan layanan pendukung[18][2]. Selanjutnya, keterbatasan atau ketidakseimbangan jumlah data berlabel menjadi tantangan umum yang dapat menyebabkan model menjadi bias[19]. Untuk mengatasi hal ini, teknik Teks Augmentasi akan diterapkan. Metode seperti *Synonym Replacement* dan *Back Translation* akan digunakan untuk memperbanyak data latih secara buatan/tiruan, yang terbukti mampu meningkatkan *F1-Score* dan akurasi model, terutama pada kelas minoritas[20][21].

Untuk memastikan seluruh tahapan penelitian berjalan secara sistematis dan terstruktur, penelitian ini mengadopsi kerangka kerja *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Kerangka kerja ini menyediakan panduan metodologis yang teruji, mulai dari pemahaman bisnis (*business understanding*) pemahaman dan persiapan data, pemodelan, evaluasi, hingga potensi implementasi (*deployment*), yang terbukti dapat meningkatkan kinerja terstruktur dalam proses klasifikasi[22].

Beberapa penelitian terdahulu telah menunjukkan keberhasilan komponen-komponen yang diusulkan dalam penelitian ini secara terpisah. Sebagai contoh, penelitian [23] berhasil menerapkan *fine-tuning IndoBERT* untuk analisis sentimen berbasis aspek pada ulasan aplikasi dan mencapai akurasi 94%, yang membuktikan kekuatan model ini untuk tugas klasifikasi teks berbahasa Indonesia. Di sisi lain, tantangan kualitas data seperti kesalahan ejaan telah dijawab [13] yang menunjukkan bahwa kombinasi *Damerau-Levenshtein Distance* dan *N-Gram* sangat efektif untuk membangun sistem koreksi ejaan, dengan kata yang benar berhasil muncul di 5 besar saran sebanyak 87%. Selain itu, untuk mengatasi keterbatasan data, penelitian [20] membuktikan bahwa teknik Teks Augmentasi seperti *Back Translation* mampu meningkatkan *F1-Score* pada kelas minoritas secara signifikan sebesar 8% pada data teks berbahasa Indonesia.

Meskipun demikian, terdapat celah penelitian (*research gap*) di mana ketiga pendekatan ini model *Transformer* canggih (*IndoBERT*), modul koreksi ejaan

cerdas, metode pelabelan *Zero-Shot* dan teknik teks augmentasi belum diintegrasikan secara menyeluruh dalam satu alur kerja untuk analisis sentimen ulasan pariwisata di Indonesia. Penelitian-penelitian sebelumnya cenderung berfokus pada salah satu aspek saja, baik pada pemodelan, perbaikan data, atau penambahan data. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut dengan mengintegrasikan *IndoBERT*, koreksi ejaan cerdas, teks augmentasi, dan analisis multi-aspek dalam kerangka kerja CRISP-DM, penelitian ini bertujuan untuk membangun sebuah model analisis sentimen yang akurat, komprehensif, dan dapat diandalkan untuk memberikan masukan strategis bagi pengembangan pariwisata pantai di Kabupaten Sumenep.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Ulasan wisatawan di platform *Google Maps* merupakan sumber data yang sangat berharga untuk mengevaluasi destinasi wisata pantai di Kabupaten Sumenep. Namun, data ini hadir dengan tantangan signifikan. Pertama, ulasan yang bersifat dibuat wisatawan seringkali mengandung bahasa informal, singkatan, dan terutama kesalahan ketik (*typo*) yang dapat menurunkan akurasi model analisis sentimen secara drastis dengan menyebabkan token tidak dikenali. Kedua, analisis sentimen yang hanya menghasilkan label umum (positif atau negatif) tidak cukup memberikan wawasan yang dapat ditindaklanjuti oleh pengelola wisata. Ketiga, keterbatasan jumlah data ulasan yang berlabel atau adanya ketidakseimbangan kelas sentimen berpotensi menghasilkan model yang bias dan tidak dapat melakukan pemerataan dengan baik. Oleh karena itu, diperlukan sebuah pendekatan sistematis untuk mengatasi ketiga tantangan ini secara bersamaan, membersihkan *noise* ilmu bahasa, melakukan analisis pada level yang lebih mendalam, dan memperkaya data latih.

1.2.2 Metode Usulan

Untuk menjawab tantangan tersebut, penelitian ini menerapkan sebuah alur kerja mendalam yang diadaptasi dalam kerangka CRISP-DM. Metode yang digunakan berpusat pada penggunaan model *Transformer IndoBERT* yang akan di *fine-tuning* untuk tugas analisis sentimen multi-aspek. Untuk mengatasi masalah kualitas data, akan diimplementasikan sebuah modul koreksi ejaan cerdas yang

mengkombinasikan *Demarau-Levenshtein Distance* untuk menghasilkan kandidat kata yang benar dan model *N-Gram* untuk memilih kandidat yang paling sesuai secara konteks. Guna mengatasi tantangan efisiensi pelabelan data, penelitian ini menerapkan metode pelabelan semi-otomatis menggunakan *Zero-Shot Classification* berbasis model *XNLI*, yang kemudian divalidasi secara manual. Selanjutnya, untuk mengatasi keterbatasan dan ketidakseimbangan data, teknik Teks Augmentasi (*Synonym Replacement* dan *Back Translation*) akan diterapkan pada data untuk meningkatkan kesetaraan data dan performa model.

1.2.3 Pertanyaan Penelitian

Berdasarkan uraian permasalahan dan metode yang diusulkan, pertanyaan penelitian utama adalah sebagai berikut:

1. Bagaimana pengaruh penerapan modul koreksi ejaan cerdas terhadap performa (akurasi, presisi, *recall*, dan *F1-Score*) model *IndoBERT* dibandingkan dengan model yang hanya menggunakan pra-pemrosesan standar?
2. Bagaimana pengaruh penambahan teknik teks augmentasi terhadap performa model *IndoBERT* yang telah dilengkapi dengan koreksi ejaan cerdas?
3. Pada model dengan performa terbaik dari hasil skenario pengujian, sejauh mana model tersebut mampu mengklasifikasikan sentimen secara akurat pada setiap aspek yang ditentukan (daya tarik, aksesibilitas, fasilitas, dan layanan pendukung)?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Berdasarkan rumusan masalah yang telah diuraikan, tujuan dari penelitian ini adalah:

1. Membangun model analisis sentimen dan mengukur sejauh mana pengaruh penerapan modul koreksi ejaan cerdas (*Damerau-Levenshtein Distance* dan *N-Gram*) dalam meningkatkan performa (*accuracy*, *precision*, *recall*, *F1-Score*) model *IndoBERT* dibandingkan dengan model *Baseline*.

2. Mengevaluasi dampak penambahan teknik teks augmentasi (*Synonym Replacement* dan *Back Translation*) terhadap peningkatan performa model *IndoBERT* yang telah dioptimalkan dengan modul koreksi ejaan cerdas.
3. Menerapkan model dengan performa terbaik dari hasil pengujian untuk menganalisis dan memetakan sentimen publik pada empat aspek pariwisata (daya tarik, aksesibilitas, fasilitas, dan layanan pendukung) di destinasi pantai Kabupaten Sumenep.

1.3.2 Manfaat

Pelaksanaan penelitian ini diharapkan dapat memberikan dampak positif yang signifikan, baik dari segi implementasi praktis di lapangan maupun pengembangan keilmuan di bidang teknologi informasi. Adapun rincian manfaat yang dapat diperoleh dari penelitian ini adalah sebagai berikut:

1. Menghasilkan sebuah model dan hasil analisis yang dapat menjadi dasar pengambilan keputusan berbasis data bagi para pemangku kepentingan dan sektor pariwisata Kabupaten Sumenep untuk meningkatkan kualitas layanan dan fasilitas.
2. Memberikan kontribusi akademis mengenai efektivitas penerapan kombinasi metode koreksi ejaan cerdas dan teks augmentasi sebagai pra-pemrosesan lanjutan untuk meningkatkan kinerja model *IndoBERT* pada data teks berbahasa indonesia yang bersifat informasi dan tidak terstruktur.

1.4 Batasan Masalah

Agar penelitian ini tetap fokus dan sesuai dengan tujuan yang telah ditetapkan, maka ditetapkan batasan-batasan masalah sebagai berikut:

1. Data yang digunakan terbatas pada ulasan publik berbahasa indonesia yang diperoleh dari *Google Maps* untuk ulasan wisata pantai di Kabupaten Sumenep.
2. Model *deep learning* utama yang digunakan adalah *IndoBERT*.
3. Analisis hanya berfokus pada data teks ulasan. Elemen lain seperti rating bintang dan gambar yang diunggah di *Google Maps* tidak akan diolah dalam penelitian ini.

4. Klasifikasi sentimen dibatasi pada empat aspek yang telah ditentukan yaitu daya tarik, aksesibilitas, fasilitas dan layanan pendukung dengan label sentimen positif atau negatif.

1.5 Sistematika Penulisan

Sistematika penelitian skripsi ini mengikuti aturan penulisan pada program studi Sistem Informasi Universitas Trunodjoyo Madura. Terdiri dari lima bab sebagai berikut:

BAB I PENDAHULUAN

Bab ini membahas terkait latar belakang, perumusan masalah, permasalahan penelitian, metode usulan, pertanyaan penelitian, tujuan penelitian, manfaat penelitian, batasan masalah dan sistematika penulisan skripsi.

BAB II TINJAUAN PUSTAKA

Menguraikan teori-teori yang menjadi landasan penelitian, seperti analisis sentimen, *IndoBERT*, *Damerau-Levenshtein Distance*, *N-Gram*, Teks Augmentasi, serta tinjauan penelitian terdahulu.

BAB III METODE PENELITIAN

Menjelaskan tahapan penelitian yang dilakukan secara sistematis menggunakan kerangka *Cross-Industry Standard Process for Data Mining* (CRISP-DM). mulai dari pengumpulan data hingga evaluasi model.

BAB IV HASIL DAN PEMBAHASAN

Bab ini memaparkan hasil implementasi dan analisis mendalam terhadap sistem yang dibangun. Pembahasan mencakup analisis karakteristik data, hasil tahapan pra-pemrosesan, proses pelatihan model, serta perbandingan kinerja model pada tiga skenario pengujian menggunakan metrik evaluasi.

BAB V PENUTUP

Bab ini menyajikan kesimpulan yang diperoleh dari hasil penelitian untuk menjawab pertanyaan penelitian, serta memberikan saran-saran memperbaiki bagi pengembangan penelitian selanjutnya atau perbaikan sistem di masa mendatang.

DAFTAR PUSTAKA

Memuat seluruh sumber referensi yang dikutip dalam penyusunan skripsi, baik berupa buku, jurnal ilmiah, prosiding konferensi, maupun sumber daring terpercaya yang relevan dengan topik penelitian.

LAMPIRAN

Berisi dokumen-dokumen pendukung penelitian yang tidak dimuat dalam laporan utama, seperti *Pseudocode* utama.