

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pariwisata mengalami pertumbuhan signifikan di era digital, ditandai dengan meningkatnya penggunaan platform seperti Google Maps, TikTok, Instagram, dan Twitter (X) sebagai media wisatawan berbagi opini dan informasi destinasi. Ulasan digital ini membentuk *big data* yang bernilai untuk analisis sentimen, disampaikan (Bibi *et al.*, 2025), serta dipertegas oleh (Masson *et al.*, 2025) terkait pentingnya konten sosial multibahasa dalam memahami persepsi wisatawan. Di Indonesia, digitalisasi pariwisata semakin relevan seiring tingginya kunjungan wisatawan. Menurut Badan Pusat Statistik Tahun 2023, kunjungan wisatawan mancanegara pasca-pandemi COVID-19 meningkat menjadi 10,41 juta. Namun, keberhasilan destinasi tidak hanya bergantung pada jumlah kunjungan, tetapi juga pada persepsi dan kepuasan wisatawan yang terekam dalam ulasan daring sebagai dasar evaluasi dan strategi pengelolaan destinasi (Iswari & Afriliana, 2025).

Salah satu destinasi wisata alam populer di Jawa Timur adalah Air Terjun Tumpak Sewu yang terletak di Kabupaten Lumajang. Pada platform Google Maps memiliki *rating* 4,8 dengan 8000+ ulasan, sering disebut “*Niagara-nya Indonesia*” karena keindahan air terjunnya yang menyerupai tirai melingkar. Menunjukkan ragam komentar multibahasa dan bercampur (*code-mixing*), misalnya “air terjunnya indah, *but the track is a bit slippery*,” mencerminkan perubahan pola komunikasi wisatawan digital. Fenomena penggunaan campuran bahasa menandakan adanya pergeseran perilaku komunikasi digital wisatawan modern.

Selain Google Maps, TikTok juga menjadi sumber data penting untuk memahami persepsi wisatawan karena *engagement* yang tinggi dan komentar berbasis pengalaman langsung. Menurut (H. Chen *et al.*, 2023), TikTok membentuk ekosistem informasi baru melalui *user-generated content*. Pada penelitian ini, tiga video TikTok terkait wisata Air Terjun Tumpak Sewu menghasilkan lebih dari 11.000+ komentar umumnya pendek, emosional, penuh slang, campuran bahasa, dan emoji, memberi tantangan tambahan analisis sentimen multibahasa. Perbedaan karakteristik antara Google Maps dan TikTok menjadikan integrasi keduanya memberikan gambaran opini wisatawan yang lebih kaya dan komprehensif.

Namun, analisis sentimen multibahasa memiliki tantangan besar, terutama karena variasi struktur kalimat, informalitas, dan keragaman ekspresi yang tidak dapat ditangkap secara optimal oleh metode tradisional seperti TF-IDF dengan model klasifikasi klasik seperti *Decision Tree*, atau *Naïve Bayes* yang hanya mengandalkan frekuensi kata tanpa pemahaman semantik (Batiuk & Dosyn, 2025). Selain itu, ketidakseimbangan dataset juga menjadi masalah umum karena dominasi kelas tertentu dapat menyebabkan bias model. Untuk mengatasinya, digunakan teknik penyeimbangan data seperti *oversampling* (SMOTE) yang menambah sampel kelas minoritas dan *undersampling* (RUS) yang mengurangi sampel kelas mayoritas. Perbandingan kedua metode ini penting untuk menentukan pendekatan paling efektif dalam meningkatkan performa klasifikasi.

Salah satu pendekatan yang terbukti efektif adalah model berbasis *transformer*, seperti *Bidirectional Encoder Representations from Transformers* (BERT) dan turunannya. *Sentence-BERT* (SBERT) dikembangkan oleh (Reimers & Gurevych, 2019) untuk menghasilkan representasi kalimat (*sentence embeddings*) berdimensi tetap yang dapat mempertahankan relasi semantik antar kalimat. SBERT memungkinkan analisis sentimen dilakukan lebih akurat pada teks multibahasa karena mampu memahami konteks linguistik dan *code-mixing* yang sering sulit diproses oleh metode berbasis kata sederhana.

Dalam proses klasifikasi, algoritma *Extreme Gradient Boosting* (XGBoost) yang diperkenalkan oleh (T. Chen & Guestrin, 2016) merupakan algoritma *ensemble learning* yang terkenal karena efisiensinya dan performa tinggi pada data berfitur kompleks. Integrasi antara SBERT dan XGBoost menghasilkan model yang mampu menggabungkan keunggulan representasi semantik mendalam dari SBERT dan kekuatan klasifikasi dari XGBoost. Hal ini menunjukkan bahwa kombinasi ekstraksi fitur SBERT dan model XGBoost mampu meningkatkan akurasi klasifikasi secara signifikan.

Beberapa penelitian sebelumnya menunjukkan bahwa pendekatan modern berbasis *transformer* menghasilkan performa lebih unggul dalam analisis sentimen wisata. Seperti (Iswari & Afriliana, 2025) mengembangkan *aspect-based sentiment analysis* menggunakan SBERT dan mencatat peningkatan *F1-score* hingga 0,84. (Af'idah *et al.*, 2022) menerapkan kombinasi *Word2Vec* dan *Bi-LSTM* pada ulasan

wisata di Bali dengan akurasi mencapai 96,86%, meskipun dataset bersifat monolingual. (Yulianti & Nissa, 2024) menunjukkan bahwa *fine-tuning Multilingual BERT* pada ulasan berbahasa Indonesia meningkatkan performa analisis sentimen dibandingkan metode tradisional seperti *Word2Vec-CNN-XGBoost*. Temuan-temuan tersebut menguatkan bahwa penggunaan model berbasis *transformer* memberikan hasil yang lebih unggul, terutama ketika dikombinasikan dengan algoritma klasifikasi efisien seperti XGBoost.

Dalam konteks wisata Air Terjun Tumpak Sewu, tantangan muncul dari variasi bahasa, gaya penulisan, platform yang berbeda (Google Maps vs TikTok), serta ketidakseimbangan jumlah ulasan antar bahasa. (Masson *et al.*, 2025) menekankan bahwa analisis sentimen multibahasa membutuhkan strategi optimal untuk menghadapi keterbatasan anotasi dan keberagaman data.. (Bibi *et al.*, 2025) juga menyatakan bahwa sebagian besar riset masih berfokus pada data berbahasa Inggris, sehingga penelitian terhadap ulasan multibahasa masih menjadi kesenjangan penelitian yang signifikan.

Berdasarkan tinjauan literatur tersebut, terdapat beberapa kesenjangan penelitian (*research gap*) berupa:

1. Penelitian analisis sentimen di Indonesia masih terbatas pada satu bahasa atau domain umum, belum banyak meneliti ulasan wisata multibahasa.
2. Penerapan ekstraksi fitur SBERT dengan model XGBoost dalam konteks pariwisata lokal seperti Tumpak Sewu masih terbatas.
3. Sebagian penelitian berfokus pada akurasi tanpa mempertimbangkan perbandingan teknik penyeimbangan data pada dataset tidak seimbang.
4. Dataset Google Maps dan TikTok masih jarang dimanfaatkan secara mendalam dalam penelitian akademik.

Model tradisional dalam analisis sentimen memiliki keterbatasan dalam menangkap nuansa *linguistik* seperti bahasa informal, *code-mixing*, dan ekspresi emosional, sehingga sering gagal memahami makna sebenarnya. Oleh karena itu, diperlukan representasi semantik mendalam seperti SBERT yang mampu memahami konteks lintas bahasa. Selain itu, penelitian sebelumnya masih berfokus pada domain umum, sementara data dari Google Maps dan TikTok sebagai sumber ulasan autentik wisatawan masih jarang dimanfaatkan. Sesuai permasalahan dan

kesenjangan tersebut, penelitian ini mengusulkan penggunaan SBERT sebagai metode ekstraksi fitur yang dikombinasikan dengan XGBoost sebagai model klasifikasi, serta membandingkan teknik penyeimbangan data SMOTE dan RUS pada ulasan multibahasa dan *code-mixing* terkait wisata Air Terjun Tumpak Sewu. Pendekatan berbasis *SBERT-SMOTE/RUS-XGBoost* diupayakan mampu meningkatkan akurasi dan efisiensi klasifikasi, sekaligus memberikan kontribusi dalam pengembangan NLP multibahasa serta mendukung pengambilan keputusan berbasis data bagi pengelola destinasi wisata.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Analisis sentimen pada ulasan Google Maps dan TikTok menghadapi berbagai tantangan karena sifat datanya yang multibahasa, *code-mixing*, serta banyak mengandung ekspresi informal seperti *slang*, emoji, dan struktur kalimat tidak baku. (Masson *et al.*, 2025), menegaskan bahwa analisis multibahasa membutuhkan strategi khusus untuk menangani keragaman *linguistik* dan keterbatasan anotasi, sementara. (Bibi *et al.*, 2025) menunjukkan bahwa ulasan wisata memiliki karakter bahasa yang kompleks sehingga sulit diproses oleh sistem otomatis. Metode tradisional seperti TF-IDF dan *Naïve Bayes* kurang efektif hanya mengandalkan frekuensi kata tanpa memahami makna semantik (Batiuk & Dosyn, 2025).

Selain itu, ketidakseimbangan jumlah ulasan antar kelas sentimen serta perbedaan distribusi data antar bahasa dan platform (Google Maps dan TikTok) meningkatkan risiko bias klasifikasi dan menurunkan performa model. Oleh karena itu, diperlukan teknik penyeimbangan data seperti *oversampling* dan *undersampling* untuk mengurangi bias tersebut. Penelitian (Batiuk & Dosyn, 2025) menunjukkan bahwa performa model klasik dapat ditingkatkan melalui *embedding modern*, sehingga permasalahan utama penelitian ini adalah bagaimana mengatasi keterbatasan metode tradisional pada data multibahasa dan *code-mixing* serta bagaimana memanfaatkan representasi semantik mendalam seperti SBERT untuk meningkatkan akurasi dan stabilitas klasifikasi ulasan wisata dari kedua platform.

1.2.2 Metode Usulan

Penelitian ini mengusulkan penggunaan *Sentence-BERT* (SBERT) sebagai ekstraksi fitur utama untuk menangkap makna semantik ulasan multibahasa Google Maps dan TikTok. SBERT menghasilkan *sentence embeddings* berdimensi tetap mampu merepresentasikan konteks lintas bahasa secara akurat. Embedding tersebut kemudian diklasifikasikan menggunakan model XGBoost dipilih karena efisiensinya memproses data berfitur kompleks *gradient boosting* (Batiuk & Dosyn, 2025).

Untuk mengatasi ketidakseimbangan data, penelitian ini juga menerapkan dan membandingkan dua teknik penyeimbangan data, yaitu SMOTE (*oversampling*) dan RUS (*undersampling*), guna memperoleh distribusi data lebih seimbang serta meningkatkan kinerja model klasifikasi.

1.2.3 Pertanyaan Penelitian

1. Bagaimana performa kombinasi ekstraksi fitur berbasis SBERT dan model XGBoost dalam meningkatkan akurasi dan F1-score pada analisis sentimen multibahasa untuk ulasan Google Maps dan TikTok, baik pada data asli (*non-translate*) maupun data terjemahan Bahasa Indonesia?
2. Bagaimana perbandingan efektivitas teknik penyeimbangan data SMOTE dan *Random Under Sampling* (RUS) dalam meningkatkan kinerja model XGBoost pada analisis sentimen multibahasa?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

Tujuan utama penelitian ini adalah mengevaluasi kinerja pendekatan modern dalam analisis sentimen multibahasa pada ulasan wisata dari Google Maps dan TikTok. Secara khusus, penelitian ini bertujuan untuk:

1. Menganalisis performa kombinasi ekstraksi fitur berbasis SBERT dan model XGBoost dalam meningkatkan akurasi dan F1-score pada klasifikasi sentimen multibahasa, baik pada data asli (*non-translate*) maupun data yang telah diterjemahkan ke Bahasa Indonesia.

2. Membandingkan efektivitas penyeimbangan data SMOTE dan *Random Under Sampling* (RUS) dalam meningkatkan kinerja model XGBoost pada analisis sentimen multibahasa.

1.3.2 Manfaat Penelitian

Secara teoretis, penelitian ini berkontribusi pada pengembangan bidang *Artificial Intelligence* (AI) dan *Natural Language Processing* (NLP) melalui penerapan *Sentence-BERT* (SBERT) yang mampu menangkap makna semantik lintas bahasa secara lebih akurat. Integrasi SBERT dengan XGBoost memperkaya kajian analisis sentimen multibahasa, khususnya pada ulasan wisata yang bersifat heterogen, informal, dan mengandung *code-mixing*. Selain itu, penelitian ini juga memberikan kontribusi dalam pemilihan teknik penyeimbangan data yang optimal melalui perbandingan metode SMOTE dan *Random Under Sampling* (RUS) meningkatkan performa model dataset tidak seimbang.

Secara praktis, penelitian ini menghasilkan model analisis sentimen otomatis yang dapat dimanfaatkan oleh pengelola destinasi wisata untuk memahami persepsi wisatawan berdasarkan ulasan Google Maps dan TikTok. Hasil analisis sentimen dapat digunakan sebagai dasar pengambilan keputusan berbasis data dalam memantau reputasi digital, mengidentifikasi permasalahan layanan, merumuskan strategi promosi dan pengembangan fasilitas yang lebih tepat sasaran, sehingga mendukung peningkatan kualitas layanan dan daya saing destinasi pariwisata.

1.4 Batasan Masalah

Ruang lingkup penelitian ini dibatasi sebagai berikut:

1. Data penelitian berasal dari hasil *scraping* ulasan wisata Air Terjun Tumpak Sewu dari dua platform, yaitu Google Maps *Reviews* dan *TikTok Comments*, dengan rentang waktu pengambilan data tahun 2021–2025.
2. Analisis dilakukan pada ulasan multibahasa, meliputi Bahasa Indonesia, Bahasa Inggris, campuran bahasa (*code-mixing*), serta ekspresi informal seperti slang, emoji, dan struktur kalimat tidak baku.
3. Klasifikasi sentimen dibatasi pada tiga kategori kelas, yaitu positif, netral, dan negatif.

4. Tahapan *pre-processing* yang diterapkan mencakup *cleaning*, *case folding*, *normalization*, *tokenizing*, *negation handling*, *stopword removal*, dan *stemming*.
5. Ekstraksi fitur menggunakan *Sentence-BERT* (SBERT) untuk menghasilkan *sentence embeddings* representasi semantik multibahasa.
6. Penyeimbangan data menggunakan metode SMOTE (*Synthetic Minority Over-sampling Technique*) dan RUS (*Random Under Sampling*)
7. Model klasifikasi yang digunakan adalah XGBoost serta kinerjanya dibandingkan pada data asli (*non-translate*) serta data yang telah diterjemahkan ke Bahasa Indonesia.
8. Evaluasi model dilakukan menggunakan metrik akurasi, *precision*, *recall*, dan *F1-score* untuk mengukur performa klasifikasi secara keseluruhan.
9. Penelitian berfokus pemodelan, eksperimen, analisis hasil klasifikasi, sehingga tidak mencakup pengembangan aplikasi atau antarmuka web.

1.5 Sistematika Penulisan Skripsi

Sistematika penulisan disusun untuk memberikan gambaran terstruktur mengenai isi dan arah penelitian, meliputi:

1. Bab I Pendahuluan

Menguraikan konteks umum penelitian melalui uraian latar belakang yang memuat urgensi, fenomena, serta kesenjangan penelitian yang mendasari topik. Selain itu, bab ini mencakup perumusan masalah, metode usulan, tujuan dan manfaat penelitian, serta batasan masalah yang mendefinisikan ruang lingkup dan fokus penelitian.

2. Bab II Tinjauan Pustaka

Memaparkan dasar teori dan konsep penelitian, seperti *Machine Learning*, *Natural Language Processing* (NLP), analisis sentimen, deskripsi objek wisata Air Terjun Tumpak Sewu, pengambilan data ulasan Google Maps *Reviews* dan *TikTok Comments* melalui *scraping* (*Apify*), *text pre-processing*, model *Sentence-BERT* (SBERT), metode penyeimbangan data, algoritma klasifikasi digunakan XGBoost (*Extreme Gradient Boosting*) serta evaluasi kinerja.tinjauan penelitian terdahulu sebagai dasar teoretis dan identifikasi kesenjangan penelitian.

3. Bab III Metodologi Penelitian

Menjelaskan alur penelitian mulai dari pengumpulan dan *scraping* data ulasan Google Maps *Reviews* dan *TikTok Comments*, tahapan *text pre-processing*, pelabelan data, pembentukan *embedding* menggunakan SBERT, penerapan metode penyeimbangan data pada data *training*, implementasi model klasifikasi XGBoost, arsitektur sistem, serta evaluasi yang digunakan seperti akurasi, *precision*, *recall*, dan *F1-score*. Serta memuat skenario uji yang dirancang untuk membandingkan performa analisis sentimen multibahasa model pada data asli (*non-translate*) dan data *translate*.

4. Bab IV Hasil dan Pembahasan

Menampilkan hasil implementasi ekstraksi fitur SBERT dan model XGBoost sebagai klasifikasi sentimen pada dataset ulasan wisata multibahasa Air Terjun Tumpak Sewu. Disajikan analisis terhadap performa model, interpretasi hasil klasifikasi, pembahasan mengenai efektivitas model dalam konteks data multibahasa, serta implementasi skenario uji coba yang sudah dirancang sebelumnya akan di jawab pada bab ini. Pembahasan dilakukan dengan membandingkan hasil penelitian terhadap teori dan studi terdahulu untuk menunjukkan kontribusi ilmiah yang dihasilkan.

5. Bab V Penutup

berisi kesimpulan yang merangkum hasil penelitian serta jawaban terhadap pertanyaan penelitian. Selain itu, disertakan saran untuk penelitian lanjutan, seperti perluasan dataset, penambahan bahasa lain, atau penerapan ekstraksi fitur SBERT dan model XGBoost pada domain pariwisata yang berbeda.