

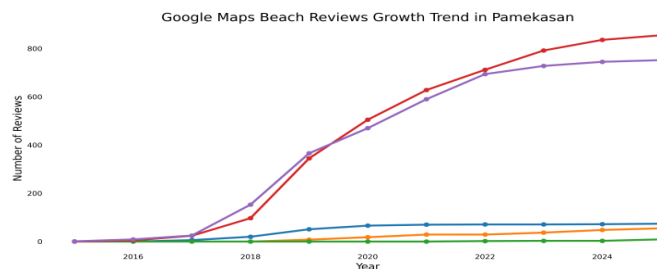
BAB I

PENDAHULUAN

1.1 Latar Belakang

Pariwisata merupakan salah satu sektor yang mendukung pertumbuhan ekonomi. Di era digital, wisatawan semakin banyak berbagi pengalaman melalui *platform online* yang telah menjadi sumber informasi utama (Ho *et al.*, 2025). Ulasan dan peringkat mencerminkan persepsi tentang kualitas layanan, fasilitas, dan pengalaman secara keseluruhan. Kabupaten Pamekasan, yang terletak di Pulau Madura memiliki potensi wisata pantai yang menawarkan daya tarik bagi wisatawan (Martono *et al.*, 2022). Wisata pantai yang ada di Kabupaten Pamekasan diantaranya yaitu, Pantai Talang Siring, Pantai Jumiang, Pantai Batu Kerbuy, Pantai Padelegan, dan Pantai Cemara Indah. Pantai-pantai ini sering dikunjungi karena ombaknya yang tenang, adanya kawasan batu karang serta lingkungannya yang bersih sebagai tempat yang nyaman untuk rekreasi. Pantai-pantai tersebut memiliki potensi besar untuk menjadi destinasi unggulan bagi wisatawan (Yuyun, 2022).

Potensi dari kawasan wisata tersebut belum ada evaluasi komprehensif tentang persepsi wisatawan yang dilakukan menggunakan data ulasan empiris. Pemahaman yang terbatas tentang sentimen wisatawan telah menyebabkan pengembangan destinasi yang kurang optimal, karena para pengelola seringkali bergantung pada asumsi atau survei konvensional dengan cakupan yang sempit (Velea *et al.*, 2024). Sebaliknya, ulasan daring yang melimpah menawarkan wawasan mendalam secara *real-time* tentang preferensi, keluhan, dan umpan balik wisatawan. Tren Pertumbuhan Ulasan Pantai di Google Maps di Pamekasan terdapat pada Gambar 1.1



Gambar 1.1 Jumlah Ulasan Wisatawan di Google Maps tentang Pantai-pantai di Kabupaten Pamekasan

Gambar 1 menunjukkan tren pertumbuhan jumlah ulasan wisatawan untuk lima destinasi pantai di Pamekasan. Hasilnya menunjukkan bahwa Pantai Talang Siring memiliki ulasan terbanyak, diikuti oleh Pantai Teluk Jumiang. Kedua pantai tersebut menunjukkan peningkatan ulasan yang signifikan dari tahun ke tahun. Sementara itu, Pantai Padelegan, Pantai Cemara Indah, dan Pantai Batu Kerbuy memiliki ulasan yang lebih sedikit, tetapi tetap menunjukkan interaksi wisatawan yang berkelanjutan dan karenanya memiliki potensi untuk pengembangan lebih lanjut (Yan *et al.*, 2025).

Perkembangan teknologi informasi dan komunikasi dapat meningkatkan penggunaan *platform* digital sebagai media penyampaian opini publik, khususnya dalam sektor pariwisata. Ulasan daring (*online review*) yang dihasilkan pengguna pada *platform* seperti Google Maps, TripAdvisor, TikTok, YouTube, dan Facebook menjadi sumber data yang sangat besar dalam menggambarkan persepsi wisatawan terhadap suatu destinasi. Data tersebut mengandung informasi penting, namun diperlukan analisis untuk penyajian data yang telah ada dalam melakukan identifikasi (Chaabi *et al.*, 2022). Analisis sentimen menjadi pendekatan yang dapat digunakan mengidentifikasi opini pengguna ke dalam kategori tertentu seperti positif, negatif, dan netral. Analisis sentimen secara umum melibatkan beberapa tahapan pada tahap *preprocessing*, teks dibersihkan dari *noise* seperti tanda baca, *stopword*, serta dilakukan normalisasi dan *stemming* agar data lebih terstruktur (Yan *et al.*, 2025).

Metode *Term Frequency–Inverse Document Frequency* (TF-IDF) digunakan untuk memberikan bobot pada kata berdasarkan tingkat kepentingannya dalam dokumen. Algoritma klasifikasi seperti *Multinomial Naive Bayes* (MNB) untuk menentukan kategori sentimen berdasarkan probabilitas kemunculan kata dalam setiap kelas (Guleria *et al.*, 2025). Peningkatan kualitas data teks dalam analisis sentimen menggunakan metode *Damerau-Levenshtein Distance* (DLD) berfungsi untuk memperbaiki kesalahan penulisan dengan menghitung jarak edit antar kata. Secara

keseluruhan integrasi antara *preprocessing*, DLD, TF-IDF, dan MNB membentuk suatu sistem analisis sentimen yang sistematis dan terstruktur (Karo *et al.*, 2025).

Penggunaan analisis sentimen telah banyak dilakukan namun, terdapat beberapa permasalahan dalam siklus kerjanya, terutama pada data teks berbahasa Indonesia. Permasalahan pertama adalah ketidakkonsistenan bahasa, hal ini menyebabkan kesulitan dalam proses normalisasi dan interpretasi makna (Dewi *et al.*, 2023). Permasalahan kedua adalah kesalahan penulisan yang besar pada data ulasan pengguna. Kesalahan ini menyebabkan variasi kata yang seharusnya sama menjadi dianggap berbeda oleh sistem, sehingga meningkatkan *lexical sparsity* dan menurunkan kualitas representasi fitur. Permasalahan ketiga adalah *noise* data, seperti keberadaan *stopword* dan kata yang tidak memiliki makna sentimen, dapat mengganggu proses pembobotan TF-IDF (Zahiroh *et al.*, 2025).

Permasalahan keempat adalah ketidakseimbangan data (*class imbalance*), di mana jumlah data pada satu kelas misalnya negatif jauh lebih banyak dibandingkan kelas lainnya. Hal ini dapat menyebabkan model cenderung bias terhadap kelas dominan. Selain itu, terdapat juga masalah pada tahap klasifikasi, yaitu keterbatasan algoritma dalam menangkap hubungan kompleks antar kata, serta tingginya dimensi fitur pada data teks (Utomo *et al.*, 2025). Penelitian sebelumnya menunjukkan bahwa kualitas *preprocessing* sangat memengaruhi akurasi analisis sentimen, terutama dalam konteks data ulasan pariwisata. Oleh karena itu, diperlukan pendekatan yang mampu mengatasi berbagai permasalahan tersebut secara terintegrasi.

Penelitian ini memfokuskan pada masalah utama yaitu ketidakkonsistenan teks akibat kesalahan penulisan dan variasi kata yang menyebabkan rendahnya akurasi klasifikasi sentimen. Masalah ini dipilih karena secara langsung memengaruhi kualitas representasi fitur dan performa model klasifikasi. Seperti contoh, kata “bagus”, “baguss”, dan “bgus” memiliki makna yang sama, namun dianggap sebagai fitur berbeda oleh sistem jika tidak dilakukan normalisasi. Hal ini menyebabkan distribusi kata menjadi tersebar dan probabilitas pada algoritma *Naive Bayes* menjadi tidak optimal. Akibatnya, model dapat menghasilkan prediksi yang kurang akurat (Khofifah

et al.,2022). Masalah ini menjadi krusial karena sebagian besar data ulasan oleh pengguna bersifat tidak terstruktur dan penuh dengan kesalahan penulisan. Oleh karena itu, diperlukan solusi yang dapat memperbaiki kualitas teks sebelum dilakukan proses klasifikasi.

Berbagai solusi telah dikembangkan untuk mengatasi permasalahan dalam analisis sentimen. Pertama, metode *preprocessing* standar seperti *cleansing*, *tokenizing*, *stopword removal*, dan *stemming* digunakan untuk membersihkan data teks. Kedua, metode pembobotan seperti TF-IDF digunakan untuk menentukan kata yang paling relevan dalam dokumen (Guleria *et al.*, 2025). Ketiga, algoritma klasifikasi seperti *Multinomial Naive Bayes*, *Support Vector Machine* (SVM), dan *Deep Learning* (LSTM, BERT) banyak dimanfaatkan dalam analisis sentimen (Catelli *et al.*, 2022). Keempat, pendekatan berbasis leksikon digunakan untuk menentukan polaritas kata berdasarkan kamus sentimen. Teknik *text augmentation* juga digunakan untuk meningkatkan jumlah dan variasi data (Chaabi *et al.*, 2022).

Kesalahan penulisan dikoreksi menggunakan metode *Damerau Levenshtein Distance* telah digunakan dalam beberapa penelitian. Metode ini mampu memperbaiki kata berdasarkan jarak edit minimum antar kata (Putra, 2024). Beragam metode yang digunakan analisis sentimen memiliki kekurangan dan kelebihan dalam penerapannya. Metode *preprocessing* standar mampu mengurangi *noise*, namun tidak sepenuhnya mengatasi kesalahan penulisan. TF-IDF efektif dalam pembobotan kata, tetapi sangat bergantung pada kualitas data awal (Liu, *et al.*, 2024). Algoritma SVM memiliki keunggulan dalam menangani data berdimensi tinggi, namun memerlukan *tuning* parameter yang kompleks dan memiliki biaya komputasi yang tinggi. Sementara itu, metode *deep learning* seperti BERT memiliki akurasi tinggi, tetapi membutuhkan data besar dan sumber daya komputasi yang tinggi (Catelli *et al.*, 2022).

Pendekatan leksikon mudah diimplementasikan, tetapi kurang fleksibel terhadap konteks bahasa. Selain itu, metode *Levenshtein Distance* hanya mempertimbangkan operasi dasar tanpa memperhitungkan transposisi karakter, sehingga kurang optimal dalam menangani kesalahan penulisan tertentu teks (Utomo

et al., 2025). Kelemahan utama dari solusi yang ada adalah belum adanya integrasi yang optimal antara normalisasi teks dan metode klasifikasi yang efisien. Berdasarkan kelemahan tersebut, penelitian ini menawarkan solusi berupa integrasi metode *Damerau-Levenshtein Distance*, TF-IDF, dan *Multinomial Naive Bayes* dalam satu kerangka kerja analisis sentimen (Wijaya dan Hermawan, 2023). Metode *Damerau-Levenshtein Distance* digunakan karena dapat menangani empat jenis kesalahan sekaligus, yaitu *insertion*, *deletion*, *substitution*, dan *transposition*, sehingga lebih akurat dalam memperbaiki kesalahan penulisan dibandingkan metode sebelumnya (Utomo *et al.*, 2025). Selanjutnya, TF-IDF digunakan karena terbukti efektif dalam merepresentasikan fitur teks dan meningkatkan kualitas pembobotan kata. Adapun *Multinomial Naive Bayes* dipilih karena memiliki performa yang baik dalam klasifikasi teks berdimensi tinggi, serta lebih efisien secara komputasi dibandingkan metode lain.

Penelitian menunjukkan bahwa MNB mampu mencapai akurasi tinggi dan stabil pada data teks dengan distribusi kata yang kompleks (Yan *et al.*, 2025). Keunggulan utama dari solusi ini adalah kemampuannya dalam meningkatkan konsistensi linguistik, mengurangi *sparsity*, serta menghasilkan model yang lebih akurat tanpa memerlukan sumber daya komputasi yang besar (Agustin *et al.*, 2025). Berdasarkan latar belakang permasalahan dan solusi yang digunakan penelitian ini dapat menghasilkan model analisis sentimen dengan hasil yang lebih baik dibandingkan metode sebelumnya. Secara terukur, target yang ingin dicapai adalah peningkatan akurasi klasifikasi hingga sekitar $\pm 89\%$, sebagaimana ditunjukkan dalam penelitian sebelumnya. Selain itu, terjadi peningkatan pada metrik evaluasi lain seperti *accuracy*, *recall*, dan *F1-score* yang lebih seimbang antar kelas. Dari segi efisiensi, metode yang diusulkan juga diharapkan mampu mengurangi waktu pemrosesan dibandingkan metode berbasis *deep learning*. Dengan demikian, penelitian ini tidak hanya berkontribusi dalam meningkatkan akurasi analisis sentimen, tetapi juga memberikan solusi yang praktis, efisien, dan dapat diterapkan pada data teks berbahasa Indonesia, khususnya dalam sektor pariwisata (Al Khadafi *et al.*, 2022).

1.2 Perumusan Masalah

1.2.1 Permasalahan (*Problem Statement*)

Permasalahan kualitas data teks dalam analisis sentimen telah banyak dibahas dalam penelitian terbaru. Ketidakteraturan bahasa pada data *user-generated content* seperti ulasan online menyebabkan tingginya *noise* dan variasi linguistik, yang berdampak langsung pada penurunan akurasi model klasifikasi. Penelitian menunjukkan bahwa ketidakkonsistenan kata dan keberadaan kata tidak relevan dapat mengganggu proses pembobotan TF-IDF sehingga menghasilkan representasi fitur yang kurang optimal (Fiddin *et al.*, 2025). Selain itu, kesalahan penulisan merupakan masalah signifikan dalam analisis sentimen berbasis teks.

Kesalahan ejaan dapat menyebabkan peningkatan *lexical sparsity*, yaitu kondisi di mana kata yang seharusnya sama dianggap berbeda oleh sistem, sehingga menurunkan performa model klasifikasi (Chaabi dan Ataa Allah, 2022). Hal ini diperkuat oleh hasil penelitian yang menunjukkan bahwa tanpa proses koreksi kata, akurasi *Naive Bayes* mengalami penurunan karena distribusi probabilitas kata menjadi tidak stabil (Yuyun *et al.*, 2025). Permasalahan umum yang sering muncul adalah ketidakseimbangan data (*class imbalance*). Distribusi data yang tidak seimbang menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga menurunkan kemampuan generalisasi model dalam mengenali kelas minoritas (Prasastio *et al.*, 2022).

Dari sisi algoritma, metode seperti *Support Vector Machine* (SVM) dan *deep learning* dapat memberikan performa tinggi, keduanya memiliki keterbatasan dalam hal kompleksitas komputasi dan kebutuhan data yang besar. Sementara itu, *Naive Bayes* tetap menjadi pilihan yang efektif untuk teks berdimensi tinggi, namun sangat bergantung pada kualitas *preprocessing* data (Aryanyi dan Suria, 2025). Permasalahan utama dalam analisis sentimen tidak hanya terletak pada algoritma klasifikasi, tetapi pada kualitas data teks yang diproses, khususnya terkait kesalahan penulisan, variasi bahasa, dan ketidakseimbangan data.

1.2.2 Metode Usulan

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pendekatan analisis sentimen berbasis integrasi beberapa metode, yaitu:

1. *Damerau Levenshtein Distance* (DLD) digunakan untuk memperbaiki kesalahan penulisan (*spelling correction*) dengan mempertimbangkan operasi edit berupa *insertion*, *deletion*, *substitution*, dan *transposition*, sehingga mampu meningkatkan konsistensi kata dalam teks.
2. *Text Preprocessing* yang meliputi *cleansing*, *case folding*, *tokenizing*, *stopword removal*, normalisasi, dan *stemming* untuk menghasilkan data teks yang bersih dan terstruktur.
3. TF-IDF (*Term Frequency–Inverse Document Frequency*) digunakan sebagai metode ekstraksi fitur untuk memberikan bobot pada kata berdasarkan tingkat kepentingannya dalam dokumen.
4. *Multinomial Naive Bayes* (MNB) digunakan sebagai algoritma klasifikasi sentimen karena memiliki performa yang baik dalam menangani data teks berdimensi tinggi dan berbasis probabilistik.

1.2.3 Pertanyaan Penelitian

1. Seberapa akurat metode *Multinomial Naive Bayes* dalam mengklasifikasikan sentimen ulasan wisata setelah diterapkan *preprocessing* dan pembobotan TF-IDF?
2. Bagaimana pengaruh penerapan *Damerau Levenshtein Distance* terhadap peningkatan akurasi analisis sentimen pada data ulasan wisata
3. Berapa besar peningkatan nilai *accuracy*, *recall*, dan *F1-score* setelah penerapan metode *Damerau Levenshtein Distance* dibandingkan tanpa proses koreksi kata pada analisis sentimen?

1.3 Tujuan & Manfaat

1.3.1 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode *Multinomial Naive Bayes* dengan pembobotan TF-IDF untuk memperoleh model analisis sentimen yang memiliki tingkat akurasi tinggi dalam mengklasifikasikan ulasan wisata ke dalam kategori positif, negatif, dan netral.
2. Mengetahui pengaruh penerapan metode *Damerau Levenshtein Distance* dalam proses *preprocessing* terhadap peningkatan kualitas data teks dan akurasi hasil klasifikasi sentimen.
3. Menganalisis peningkatan performa model analisis sentimen berdasarkan metrik evaluasi seperti *accuracy*, *recall*, dan *F1-score* setelah integrasi metode *preprocessing*, TF-IDF, dan *Multinomial Naive Bayes*.

1.3.2 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut

1. Menghasilkan model analisis sentimen yang lebih akurat dan stabil, sehingga dapat digunakan sebagai alat bantu dalam memahami opini wisatawan secara lebih objektif dan berbasis data.
2. Memberikan kontribusi dalam pengembangan metode analisis teks berbahasa Indonesia, khususnya dalam mengatasi permasalahan kesalahan penulisan dan variasi kata melalui pendekatan *Damerau Levenshtein Distance*.
3. Menyediakan dasar pengambilan keputusan berbasis data bagi pengelola destinasi wisata, sehingga dapat meningkatkan kualitas layanan, fasilitas, dan pengalaman wisatawan berdasarkan hasil analisis sentimen.

1.4 Batasan-batasan

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini terbatas pada ulasan wisata yang diperoleh dari platform digital seperti Google Maps, TripAdvisor, TikTok, YouTube, dan Facebook, dengan fokus pada destinasi wisata pantai di Kabupaten Pamekasan.
2. Metode yang digunakan dalam penelitian ini dibatasi pada proses *preprocessing* teks, penerapan algoritma *Damerau Levenshtein Distance* untuk koreksi kata, pembobotan TF-IDF, serta klasifikasi menggunakan *Multinomial Naive Bayes* tanpa membandingkan dengan metode *deep learning* secara mendalam.
3. Klasifikasi sentimen dalam penelitian ini hanya mencakup tiga kategori, yaitu positif, negatif, dan netral, dengan evaluasi performa menggunakan metrik *accuracy*, *recall*, dan *F1-score*.

1.5 Sistematika Proposal

BAB I PENDAHULUAN

Bab ini berisi latar belakang penelitian, perumusan masalah, tujuan dan manfaat penelitian, batasan penelitian, serta sistematika penulisan proposal skripsi.

BAB II TINJAUAN PUSTAKA

Bab ini membahas teori-teori yang relevan dengan penelitian, seperti analisis sentimen, *preprocessing* teks, algoritma *Damerau Levenshtein Distance*, TF-IDF, *Multinomial Naive Bayes*, serta penelitian terdahulu yang mendukung.

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan penelitian yang meliputi pengumpulan data, *preprocessing*, penerapan metode yang diusulkan, proses klasifikasi, metode evaluasi yang digunakan serta pemaparan skenario uji coba.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil implementasi sistem, pengujian model, analisis hasil berdasarkan metrik evaluasi, serta pembahasan terkait performa metode yang digunakan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian serta saran untuk pengembangan penelitian selanjutnya.