

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pulau Madura, yang juga dikenal sebagai Pulau Garam, terletak di sebelah utara Jawa Timur, Indonesia. Pulau ini memiliki luas kurang lebih 5.379 km² dan jumlah penduduk lebih dari 4 juta jiwa pada tahun 2023. Pulau Madura dikenal sebagai '*Hidden Paradise*' atau surga tersembunyi karena daya tariknya yang beragam, antara lain wisata alam, religi, sejarah, dan budaya[1]. Mengenai pariwisata, Madura menawarkan keindahan alam yang memukau, salah satunya pantai yang eksotis, yang mampu menarik perhatian para wisatawan yang ingin mencari pengalaman istimewa.

Meskipun memiliki potensi besar dalam pengembangan sektor pariwisata, Madura masih menghadapi beberapa tantangan yang signifikan. Menurut Kabar Madura, jumlah wisatawan yang mengunjungi Pamekasan pada tahun 2023 mengalami penurunan dibandingkan tahun sebelumnya. Pada tahun 2022, tercatat 198.118 wisatawan nusantara (winus) dan 17 wisatawan mancanegara (wisman) mengunjungi daerah tersebut. Namun, pada tahun 2023, jumlah wisatawan nusantara turun menjadi 151.875, sementara tidak ada kunjungan dari wisatawan mancanegara. Penurunan ini disebabkan oleh beberapa faktor, diantaranya kurangnya promosi yang efektif untuk menarik perhatian wisatawan domestik maupun mancanegara. Selain itu, infrastruktur yang belum memadai seperti jalan, transportasi, dan fasilitas umum juga menjadi hambatan dalam pengembangan pariwisata. Ulasan negatif dari wisatawan, yang sering kali mencatat minimnya fasilitas dan pelayanan yang kurang baik, juga turut mempengaruhi citra Pulau Madura sebagai destinasi wisata. Tantangan-tantangan ini perlu segera diatasi untuk mengoptimalkan potensi pariwisata Madura dan menjadikannya tujuan wisata unggulan[2].

Analisis sentimen menggunakan analisis teks untuk memperoleh berbagai sumber data dari internet dan berbagai platform media sosial. Tujuannya adalah untuk memperoleh hasil *review* dan opini dari pengguna yang terdapat pada platform tersebut[3]. Pada topik ini, analisis sentimen menjadi alat penting untuk

memahami perasaan dan opini yang telah diberikan oleh wisatawan yang telah berkunjung ke berbagai tempat wisata. Melalui analisis ini, pengelola pariwisata diharapkan dapat mengetahui aspek positif dan negatif yang dirasakan oleh wisatawan. Hal ini tentu sangat berguna untuk mengevaluasi berbagai aspek dari objek pariwisata, seperti fasilitas yang mungkin dinilai kurang memadai, layanan yang tidak memenuhi harapan dan lain sebagainya. Dengan demikian, pengelola dapat mengambil langkah-langkah yang tepat untuk meningkatkan kualitas dan pengalaman wisata secara keseluruhan, memastikan bahwa setiap kunjungan meninggalkan kesan yang baik dan dapat memenuhi ekspektasi para wisatawan.

Berdasarkan penelitian sebelumnya, sudah ada peneliti yang melakukan analisis sentimen dengan menggunakan metode *Random Forest*. Akan tetapi, judul yang diangkat berbeda dengan penelitian yang saat ini. Pada penelitian tersebut, diperoleh hasil akurasi sebesar 82%, precision sebesar 86% dan recall sebesar 86%[4]. Sementara itu, pada penelitian lainnya yang telah dilakukan oleh [5] dengan judul *Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Reviews in Bangkalan Regency*, memperoleh akurasi sebesar 70.22% untuk ulasan berbahasa Indonesia. Hasil dari kedua penelitian diatas, menunjukkan bahwa metode *Random Forest* memberikan hasil akurasi lebih tinggi dibandingkan dengan metode *Support Vector Machine*. Selain memberikan hasil akurasi yang tinggi, metode *Random Forest* juga mampu menangani data yang hilang, mengimplementasikan secara efisien pada dataset yang sangat besar, menangani data yang hilang, serta menggabungkan hasil dari banyak decision tree yang cenderung lebih stabil dan kurang rentan terhadap *overfitting*[6]. Namun, di balik beberapa kelebihan tersebut, *Random Forest* juga memiliki kelemahan, yaitu pada jumlah fitur yang sangat besar dapat menyebabkan penurunan performa model, yang tentu akan mempengaruhi akurasi hasil pengklasifikasian data[7].

Random Forest adalah salah satu metode *machine learning* yang tergolong dalam kategori *supervised learning* (pembelajaran diawasi). Umumnya, *Random Forest* digunakan untuk mengklasifikasi data dalam jumlah yang besar. Klasifikasi dengan *Random Forest* dilakukan melalui penggabungan pohon keputusan (*decision tree*) dengan melakukan *training* pada data sampel. Penentuan hasil klasifikasi dari *Random Forest* ini diambil berdasarkan hasil voting dari pohon-

pohon yang terbentuk. Pengacakan dilakukan dengan dua cara, yaitu pengambilan sampel data secara acak untuk membentuk sampel *bootstrap* seperti yang dilakukan dalam metode *bagging* (*bootstrap aggregating*), dan yang kedua dengan cara pemilihan fitur secara acak untuk menghasilkan pohon keputusan individu[8]

Kumpulan data yang dianalisis mungkin berisi ratusan fitur, banyak di antaranya tidak relevan atau berulang di dalam dokumen, sehingga mempengaruhi kinerja algoritma klasifikasi dari segi efisiensi waktu dan keakuratan hasil klasifikasi. Oleh karena itu, diperlukan suatu metode yang dapat mereduksi dimensi fitur, sehingga dapat memilih fitur-fitur penting yang mewakili dokumen dan meningkatkan kinerja klasifikasi[9]. Pada penelitian ini, terdapat tiga seleksi fitur yang dipakai, antara lain *Information Gain* dan *Chi-Square*. *Information Gain* merupakan salah satu metode yang digunakan untuk pemilihan fitur (*feature selection*) dalam *machine learning* dan juga data mining. Penggunaan teknik ini dapat mereduksi dimensi fitur dengan cara mengukur nilai reduksi *entropy* sebelum pemisahan dikurangi dengan nilai *entropy* setelah pemisahan, sehingga dapat menentukan atribut apa saja yang akan digunakan atau dibuang[10]. Seleksi fitur *Chi-Square* adalah metode yang digunakan untuk menentukan fitur mana yang paling relevan atau signifikan dalam mempengaruhi variabel target, terutama dalam kasus di mana variabel prediktor dan variabel target bersifat kategorikal. *Chi-Square* melakukan uji independensi antara dua peristiwa, yaitu peristiwa munculnya suatu kata dan peristiwa munculnya suatu kelas[11]. Seleksi fitur adalah metode untuk mengukur ketergantungan antara variabel input (fitur) dan variabel output (target). Metode ini dapat menangkap hubungan linear maupun non-linear dengan menilai seberapa banyak informasi yang diberikan suatu fitur tentang target yang diprediksi[12].

Berdasarkan latar belakang dan permasalahan yang muncul, penulis akan mengajukan penelitian dengan judul “Penerapan Metode *Random Forest* Dalam Analisis Sentimen Ulasan Destinasi Wisata Madura Dengan *Ensemble Feature Selection*”. Penulis berharap penelitian ini dapat menghasilkan akurasi yang tinggi dalam analisis sentimen tersebut.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Jumlah fitur yang besar dapat mempengaruhi kinerja metode *Random Forest* yang bisa menyebabkan penurunan performa model, maka dibutuhkan pendekatan *ensemble* seleksi fitur yang terdiri dari *Information Gain* dan *Chi-Square*.

1.2.2 Metode Usulan

Mengklasifikasikan data ulasan Wisata Madura pada *Google Maps* dengan menggunakan metode *Random Forest* dan penambahan beberapa seleksi fitur diantaranya *Information Gain* dan *Chi-Square* serta *ensemble* seleksi fitur.

1.2.3 Pertanyaan Penelitian

Beberapa pertanyaan dari penelitian ini adalah sebagai berikut:

1. Berapa besar hasil akurasi serta waktu terbaik yang didapat dari metode *Random Forest* tanpa penggunaan seleksi fitur?
2. Berapa besar hasil akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan masing-masing seleksi fitur?
3. Berapa besar hasil akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan *ensemble* seleksi fitur?
4. Berapa besar hasil akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan masing-masing seleksi fitur serta tambahan metode SMOTE?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan Penelitian

1. Untuk mengetahui tingkat akurasi serta waktu terbaik yang didapat dari metode *Random Forest* tanpa penggunaan seleksi fitur.
2. Untuk mengetahui tingkat akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan masing-masing seleksi fitur.
3. Untuk mengetahui tingkat akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan *ensemble* seleksi fitur.

4. Untuk mengetahui tingkat akurasi serta waktu terbaik yang didapat dari metode *Random Forest* dengan menggunakan masing-masing seleksi fitur serta tambahan metode SMOTE.

1.3.2 Manfaat Penelitian

Beberapa manfaat yang didapat dari penelitian ini adalah sebagai berikut:

1. Mendapatkan wawasan yang lebih mendalam mengenai persepsi wisatawan terhadap destinasi wisata Madura.
2. Dapat mengetahui seberapa besar pengaruh metode *Random Forest* dan *ensemble* seleksi fitur *Information Gain* dan *Chi-Square* dalam klasifikasi sentimen wisatawan terhadap destinasi wisata Madura.
3. Berkontribusi pada literatur akademik dengan menyediakan studi kasus spesifik tentang analisis sentimen pada ulasan *Google Maps* menggunakan teknik *Machine Learning* yang dapat dijadikan referensi untuk penelitian selanjutnya.

1.4 Batasan Masalah

Beberapa batasan masalah yang terdapat pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan pada penelitian ini adalah penggabungan data sekunder dan primer yang sama-sama diambil dari ulasan *Google Maps*.
2. Data ulasan sekunder diambil oleh peneliti sebelumnya, yaitu pada tanggal 10 Nopember 2023, dan data primer diambil tanggal 11 Juni 2024.
3. Data yang dianalisis menggunakan data ulasan berbahasa Indonesia.
4. Data yang digunakan berjumlah 1932 ulasan dengan 1259 ulasan positif, 575 ulasan negatif dan 98 ulasan netral yang mencakup 7 wisata yang ada di Madura, di antaranya Pantai Lombang, Pantai Sembilan, Pantai Slopeng, Air Terjun Toroan, Bukit Jaddih, Bukit Palalangan, dan Pantai Talang Siring.
5. Pelabelan dilakukan secara manual serta dibagi menjadi 3, yaitu positif, negatif dan netral yang telah divalidasi oleh seorang pakar Guru Bahasa Indonesia.
6. Metode yang digunakan, yaitu *Random Forest* dan *ensemble* seleksi fitur yang terdiri dari *Information Gain* dan *Chi-Square*.

7. Bahasa pemrograman yang digunakan adalah *Python*.
8. *Splitting* data yang digunakan, yaitu 90:10 dan 80:20

1.5 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut:

1. BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang, perumusan masalah yang mencakup permasalahan, metode usulan, pertanyaan penelitian serta terdapat juga tujuan dan manfaat dari penelitian, batasan-batasan masalah dalam penelitian dan sistematika penulisan.

2. BAB II KAJIAN PUSTAKA

Bab ini berisi tentang penjelasan dari teori-teori pendukung yang digunakan sebagai bahan penunjang untuk memahami isi yang akan dibahas dalam penelitian ini.

3. BAB III METODE PENELITIAN

Bab ini berisi rancangan sistem yang dibangun berdasarkan metode penelitian yang dipakai, dimulai dari dataset, arsitektur, hingga skenario uji coba.

4. BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil implementasi pada sistem, hasil skenario uji coba dan hasil analisa klasifikasi.

5. BAB V KESIMPULAN DAN SARAN

Bab ini berisi hasil kesimpulan dari pembahasan yang telah dilakukan dan juga saran untuk penelitian selanjutnya.