

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pariwisata telah menjadi salah satu sektor utama dalam perekonomian masyarakat Sumenep yang mengundang perhatian besar dari wisatawan lokal maupun internasional. Perkembangan sektor pariwisata ini diikuti dengan peningkatan volume ulasan yang diberikan oleh para pengunjung terhadap destinasi pariwisata di Sumenep. Ulasan-ulasan ini mencerminkan pengalaman, preferensi, dan pandangan para pengunjung terhadap berbagai aspek pariwisata, termasuk objek wisata, akomodasi, makanan, dan aktivitas lainnya. Rating dalam ulasan juga menjadi faktor penentu wisatawan dalam menentukan perjalanan mereka. Masih banyak pengunjung situs yang beranggapan bahwa cukup dengan melihat peringkat ulasan, mereka sudah bisa menilai tingkat kepuasan terhadap suatu destinasi [1]. Oleh karena itu, analisis ulasan pariwisata menjadi penting dalam pemahaman dan pengembangan sektor pariwisata Sumenep. Dalam rangka memahami sentimen yang terkandung dalam ulasan-ulasan ini, penelitian ini mengusulkan penggunaan metode *Random Forest* dengan *Mutual Information* dan *Chi-Square*

Meskipun ulasan pariwisata memiliki peran penting dalam memengaruhi keputusan calon wisatawan, proses dalam menganalisis ulasan pariwisata di Sumenep tidak terlepas dari sejumlah masalah. Dalam analisis sentimen, banyak data dan atribut yang digunakan dalam proses klasifikasi. Keragaman dalam bahasa, gaya penulisan dan sentiment dalam ulasan turut memperumit proses analisis data, disamping itu kata-kata yang kurang relevan dalam dokumen juga dapat mengurangi nilai akurasi [1]. Volume besar ulasan yang diterima tentu membuat analisis manual menjadi tidak praktis, analisis sentimen secara otomatis juga menjadi tugas yang cukup rumit.

Masalah utama yang dihadapi adalah mengembangkan metode analisis sentimen yang efisien dan akurat untuk ulasan pariwisata di Sumenep. Beberapa solusi yang telah ada termasuk metode-metode tradisional seperti *naive bayes*, *support vector machines* (svm), dan algoritma *machine learning* lainnya. Kelemahan dari solusi-solusi tersebut meliputi kemampuan terbatas dalam memproses bahasa lokal, kesulitan dalam mengatasi data dalam volume besar, serta

kurangnya akurasi dalam mengklasifikasikan sentimen ulasan yang kompleks. Dalam jurnal “*Comparison of Machine Learning for Sentiment Analysis in Detecting Anxiety Based on Social Media Data*” menunjukkan kelemahan beberapa metode lain, seperti K-Nearest Neighbors (K-NN) yang meskipun memiliki precision tertinggi, mengalami penurunan efektivitas karena recall yang rendah. XG-Boost juga menunjukkan kelemahan dengan recall yang tinggi tetapi precision yang rendah, yang mengurangi akurasi keseluruhan dalam deteksi sentimen. Selain itu, Bernoulli dan Support Vector Classifier walaupun memiliki akurasi mendekati *Random Forest*, namun tidak mencapai keseimbangan yang optimal antara precision dan recall. Sebaliknya, *Random Forest* unggul dengan akurasi terbaik, mencapai 84,99% pada ekstraksi fitur Count-Vectorization dan 82,63% pada TF-IDF, serta memberikan keseimbangan yang lebih baik antara precision dan recall. Keunggulan ini membuat *Random Forest* pilihan yang efektif dalam analisis sentimen dibandingkan metode lain [2].

Selanjutnya solusi terpilih yang ditawarkan dalam penelitian ini adalah penggunaan metode *Random Forest* dengan *Mutual Information* dan *Chi-Square*. Kelebihan metode ini adalah kemampuannya dalam mengatasi variasi bahasa lokal, efisien dalam mengelola data besar, dan memiliki akurasi yang lebih baik dalam mengklasifikasikan sentimen ulasan yang kompleks [3]. Metode *Random Forest* unggul karena kemampuannya untuk menggabungkan prediksi dari beberapa pohon keputusan dan memberikan evaluasi akurasi serta diversitas yang lebih baik. *Random Forest* menunjukkan kapabilitas superior dalam memprediksi mortalitas jangka pendek dibandingkan metode tradisional seperti National Early Warning Score, menjadikannya alat prediksi yang lebih andal dan efektif [4]. Di sisi lain, oversampling *SMOTE* berguna untuk mengatasi ketidakseimbangan data dalam klasifikasi dengan menghasilkan data sintetis tambahan pada kelas minoritas. Hal ini membantu meningkatkan distribusi data dan akan meningkatkan akurasi serta performa model klasifikasi, terutama dalam kasus di mana satu kelas data jauh lebih sedikit dibandingkan dengan kelas lainnya [5]. Sementara itu, *Mutual Information* dan *Chi-Square* digunakan sebagai metode seleksi fitur yang menganalisis relevansi setiap fitur dengan label sentimen, dengan menghitung hubungan antara variabel fitur (kata-kata) dan kelas sentimen. *Mutual Information*

dan *Chi-Square* efektif dalam memilih fitur yang paling signifikan, sehingga dapat meningkatkan akurasi dan efisiensi dalam pengolahan data besar pada analisis sentimen [1]. *Mutual Information* dan Information Gain merupakan dua metode seleksi fitur yang digunakan dalam pengolahan teks. *Mutual Information* dan Information gain mempunyai konsep seleksi yang berbeda. *Mutual Information* mengukur tingkat keterkaitan antara suatu fitur (seperti kata atau frasa) dan kelas target, yaitu seberapa banyak informasi yang diberikan fitur tersebut tentang kelas target [6]. Sebaliknya, Information Gain mengukur pengurangan ketidakpastian (entropi) dalam memprediksi kelas target setelah data dibagi berdasarkan fitur tersebut. Dengan demikian, *Mutual Information* fokus pada hubungan antara fitur dan kelas target, sementara Information Gain menekankan pada pengurangan ketidakpastian dalam pembagian data [7].

Dengan menerapkan *Random Forest*, *Mutual Information*, *Chi-Square* dan *SMOTE*, penelitian ini bertujuan meningkatkan akurasi klasifikasi sentimen ulasan pariwisata di Sumenep, mempercepat analisis data, dan mendukung pengembangan pariwisata serta strategi promosi yang lebih tepat.

1.2 Perumusan Masalah

1.2.1 Permasalahan

Dengan meningkatnya popularitas destinasi pariwisata di Sumenep, jumlah ulasan yang diterima meningkat pesat, sehingga analisis manual menjadi tidak efisien dan rentan kesalahan. Penelitian ini menggunakan algoritma *Random Forest* untuk menganalisis sentimen ulasan dengan mengatasi ketidakseimbangan data melalui oversampling *SMOTE*. Selain itu, seleksi fitur dilakukan menggunakan metode *Mutual Information* dan *Chi-Square* sebagai pembanding untuk menyaring fitur yang paling relevan. Kombinasi metode ini diharapkan mampu meningkatkan efisiensi dan akurasi analisis sentimen, mendukung pengambilan keputusan dalam pengembangan pariwisata Sumenep.

1.2.2 Metode Usulan

Penambahan seleksi fitur *Mutual Information* dan *Chi-Square* dengan *Oversampling SMOTE* pada metode *Random Forest* untuk klasifikasi ulasan objek wisata Sumenep pada *google maps*.

1.2.3 Pertanyaan Penelitian

1. Berapa akurasi metode *Random Forest* dengan *Mutual Information* dan *Chi-Square* serta Oversampling *SMOTE* dalam mengklasifikasikan sentimen ulasan pariwisata di Sumenep berdasarkan volume ulasan yang besar?
2. Bagaimana perbandingan akurasi antara metode *Random Forest* dengan dan tanpa penggunaan *Mutual Information* dan *Chi-Square* dalam mengklasifikasikan sentimen ulasan pariwisata di Sumenep?
3. Seberapa besar perubahan akurasi metode *Random Forest* dalam analisis sentimen ulasan pariwisata di Sumenep setelah diterapkan Oversampling *SMOTE* dengan seleksi fitur *Mutual Information* dan *Chi-Square*?

1.3 Tujuan Dan Manfaat

1.3.1 Tujuan Penelitian

Untuk meningkatkan pemahaman tentang metode analisis sentimen ulasan pariwisata di Sumenep menggunakan metode *Random Forest* dengan *Mutual Information* dan *Chi-Square*, serta mengukur kinerja metode ini dalam menghadapi volume besar data ulasan dengan distribusi data yang beragam. Selain itu, penelitian ini juga bertujuan untuk mengevaluasi pengaruh seleksi fitur serta oversampling *SMOTE* terhadap akurasi dan efisiensi analisis sentimen.

1.3.2 Manfaat Penelitian

Manfaat yang diharapkan dari klasifikasi sentimen pada penelitian ini adalah sebagai berikut:

1. Penelitian ini dapat meningkatkan kualitas analisis sentimen ulasan pariwisata di Sumenep. Dengan mengevaluasi efektivitas metode *Random Forest* dengan *Mutual Information* dan oversampling *SMOTE*, diharapkan hasil analisis sentimen ulasan akan menjadi lebih akurat.
2. Penelitian ini akan menyumbangkan pengetahuan ilmiah tentang penggunaan metode *Random Forest* dengan *Mutual Information* dan *Chi-Square* dalam analisis sentimen.

1.4 Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut:

1. Penelitian berfokus pada enam destinasi wisata di Sumenep, yaitu Pantai Sembilan, Museum Kraton, Bukit Tinggi, Pantai Lombang, Goa Soekarno, dan Gili Labak.

2. Data yang digunakan pada penelitian ini diambil dari ulasan *google maps*.
3. Klasifikasi yang dilakukan dibagi menjadi 3 kategori yaitu: positif, netral dan negatif
4. Data diambil dari 5 tahun lalu dari 2019-2024.
5. Bahasa pemrograman yang dipakai adalah python.
6. Sistem yang akan dirancang berbasis Web.
7. Metode yang digunakan *Random Forest* dengan seleksi fitur *Mutual Information* dan *Chi-Square*

1.5 Sistematika Penyusunan

Sistematika penyusunan skripsi ini terdiri dari beberapa bab, yaitu sebagai berikut:

1. Bab I Pendahuluan

Bab ini berisi latar belakang, perumusan masalah yang meliputi permasalahan, metode usulan, dan pertanyaan penelitian, kemudian selanjutnya terdapat tujuan dan manfaat penelitian, batasan-batasan penelitian dan sistematika penulisan.

2. Bab II Kajian Pustaka

Bab ini berisi penjelasan mengenai teori-teori pendukung dari berbagai sumber yang berhubungan dengan penelitian ini.

3. Bab III Metode Usulan

Bab ini berisi penjelasan tentang data penelitian, arsitektur sistem, dan perkiraan jadwal pengerjaan.

4. Bab IV Hasil dan Pembahasan

Bab ini berisi hasil dari eksperimen yang dilakukan, analisis data, serta interpretasi hasil yang diperoleh. Pembahasan mencakup performa model yang digunakan, perbandingan dengan metode lain jika ada, serta faktor-faktor yang mempengaruhi hasil analisis sentimen.

5. Bab V Kesimpulan dan Saran

Bab ini berisi kesimpulan dari penelitian yang telah dilakukan, menjawab pertanyaan penelitian yang telah dirumuskan sebelumnya. Selain itu, diberikan saran untuk pengembangan lebih lanjut agar penelitian ini dapat ditingkatkan atau diterapkan dalam skala yang lebih luas.