

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi dan komunikasi di era digital telah mendisrupsi berbagai sektor, termasuk industri kuliner. Salah satu manifestasi nyata dari disrupsi ini adalah pergeseran perilaku konsumen yang kini sangat bergantung pada informasi daring, khususnya ulasan pelanggan pada platform *Google Maps*. Fitur ulasan pada platform ini tidak hanya berfungsi sebagai penunjuk arah, tetapi memainkan peran krusial dalam merepresentasikan reputasi sebuah restoran dan secara langsung memengaruhi keputusan calon konsumen [1].

Rumah Makan Bebek Sinjay, sebagai salah satu destinasi wisata kuliner primadona yang masif dikunjungi, menerima ribuan ulasan teks setiap harinya. Ulasan-ulasan tersebut memuat berbagai opini yang merepresentasikan tingkat kepuasan pelanggan, baik berupa sentimen positif maupun sentimen negatif. Namun, volume data tekstual yang terus bertambah secara eksponensial ini membuat analisis sentimen secara manual menjadi sangat tidak efisien, memakan waktu, dan rentan terhadap bias subjektivitas pengamat [2].

Untuk memecahkan kendala tersebut, dalam ranah *Natural Language Processing* (NLP), analisis sentimen hadir sebagai solusi komputasional untuk mengekstraksi dan mengklasifikasikan polaritas opini publik secara otomatis. Pendekatan algoritma *Machine Learning* seperti *K-Nearest Neighbor* (K-NN) sering direkomendasikan untuk klasifikasi teks karena ketangguhannya dalam ruang dimensi fitur yang tinggi [3], terutama jika dipadukan dengan skema pembobotan kata *Term Frequency-Inverse Document Frequency* (TF-IDF) yang berfungsi memberikan bobot pada kata-kata penting dalam dokumen [4].

Meskipun demikian, penerapan algoritma klasifikasi pada dataset ulasan alamiah sering kali dihadapkan pada masalah fundamental, yaitu ketidakseimbangan kelas (*class imbalance*). Pada kasus ulasan restoran, distribusi opini pelanggan umumnya sangat asimetris, di mana ulasan bernada positif jauh mendominasi dibandingkan ulasan negatif [5], [6]. Kondisi ketidakseimbangan

yang ekstrem ini dapat menyebabkan model algoritma mengalami *majority class bias* (bias kelas mayoritas), sehingga sistem kehilangan sensitivitas dan gagal mengenali karakteristik ulasan dari kelas minoritas. Selain masalah ketidakseimbangan data, analisis sentimen pada domain kuliner memiliki tingkat kompleksitas bahasa yang tinggi, seperti penggunaan bahasa gaul (*slang*), istilah lokal, hingga kalimat implisit yang menyulitkan proses pelabelan data awal.

Untuk menjembatani celah penelitian (*research gap*) tersebut, penelitian ini mengusulkan sebuah pendekatan sistem hibrida. Sistem ini mengintegrasikan perangkat lunak *Orange Data Mining* dengan injeksi *script* kustom berbasis bahasa pemrograman Python pada tahap awal. Sebelum diklasifikasikan menggunakan K-NN, dataset ulasan diekstraksi dan dilabeli secara semi-otomatis menggunakan pendekatan *Lexicon-Based* dengan menyematkan *Custom Dictionary* (kamus kustom) [7]. Kamus ini secara spesifik dirancang untuk menangkap leksikon opini pada domain kuliner, guna menghasilkan data latih yang lebih seimbang, bersih, dan representatif.

Berdasarkan latar belakang permasalahan tersebut, penelitian ini bertujuan untuk membangun model analisis sentimen yang tangguh pada ulasan pelanggan Rumah Makan Bebek Sinjay. Penelitian ini secara spesifik akan mengevaluasi integrasi antara ekstraksi fitur TF-IDF dan algoritma K-NN pada dataset hasil penyeimbangan *Lexicon-Based*. Diharapkan, melalui pengujian *hyperparameter* nilai ketetanggaan (k), model yang dibangun mampu memberikan batasan keputusan (*decision boundary*) yang optimal dalam memproses teks berbahasa Indonesia sekaligus memitigasi bias prediksi.

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan arsitektur sistem analisis sentimen yang memadukan *software Orange Data Mining* dengan injeksi *script* kustom Python untuk tahapan *stemming* dan pelabelan *Lexicon-Based* berbahasa Indonesia?

2. Bagaimana penerapan metode ekstraksi fitur TF-IDF dan normalisasi vektor *L2 (Euclidean)* dalam merepresentasikan data teks ulasan ke dalam ruang vektor?
3. Bagaimana performa tingkat akurasi algoritma *Distance-weighted K-Nearest Neighbor* (K-NN) dalam mengklasifikasikan sentimen ulasan pelanggan Bebek Sinjay, serta bagaimana pengaruh variasi nilai k terhadap penanganan fenomena paradoks akurasi pada data yang tidak seimbang?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan tidak menyimpang dari tujuan utama, maka ditetapkan beberapa batasan masalah sebagai berikut:

1. Dataset yang digunakan bersumber dari data sekunder penelitian sebelumnya, yang berisi 1.587 dokumen mentah ulasan pelanggan Rumah Makan Bebek Sinjay berbahasa Indonesia dari platform *Google Maps*.
2. Kelas sentimen difokuskan pada klasifikasi biner, yaitu kelas Positif dan kelas Negatif, yang dilabeli secara komputasional menggunakan *Lexicon-Based Custom Dictionary*.
3. Pra-pemrosesan spesifik (*stemming* dan leksikon) dibangun menggunakan *Python Script*, sedangkan *preprocessing* lanjutan, ekstraksi fitur (TF-IDF), dan klasifikasi dibangun menggunakan perangkat lunak *Orange Data Mining*.
4. Algoritma klasifikasi yang dievaluasi adalah *K-Nearest Neighbor* (K-NN) dengan metrik jarak *Euclidean* dan pembobotan berbasis jarak (*Weight: By distance*), diuji pada nilai $k = 3, 5, 7, 9$.

1.4 Tujuan dan Manfaat

Berikut adalah tujuan dan manfaat dari penelitian ini berdasarkan rumusan masalah yang telah dijelaskan.

1.4.1 Tujuan

Tujuan yang ingin dicapai melalui penelitian ini adalah:

1. Mengembangkan arsitektur sistem analisis sentimen yang inovatif dengan mengintegrasikan *Orange Data Mining* dan *script* kustom Python guna memproses kompleksitas teks berbahasa Indonesia.
2. Menerapkan ekstraksi fitur TF-IDF teroptimasi untuk mengubah teks ulasan pelanggan Bebek Sinjay menjadi format matriks numerik yang siap dikomputasi oleh model pembelajaran mesin.
3. Mengukur dan mengevaluasi kinerja algoritma K-NN menggunakan metode *10-Fold Cross Validation* serta matriks konfusi (*Confusion Matrix*) guna menemukan nilai k yang optimal sebagai batas keputusan (*decision boundary*) dalam menghindari bias kelas mayoritas.

1.4.2 Manfaat

Manfaat yang diharapkan dari hasil penelitian ini meliputi dua aspek, yaitu:

1. Manfaat Teoritis: Memberikan kontribusi literatur pada bidang *Text Mining* dan *Natural Language Processing* (NLP), khususnya dalam membuktikan bahwa injeksi *script* kustom mampu memecahkan batasan bahasa *software* bawaan. Penelitian ini juga memberikan wawasan terkait mitigasi *accuracy paradox* (*over-smoothing*) pada dataset yang tidak seimbang (*imbalanced data*).
2. Manfaat Praktis: Menghasilkan sebuah sistem pemodelan cerdas yang dapat dimanfaatkan oleh pihak manajemen Rumah Makan Bebek Sinjay sebagai alat bantu analitis praktis untuk mengevaluasi umpan balik dan kepuasan konsumen secara presisi.

1.5 Sistematika Laporan

Untuk memberikan gambaran yang jelas dan menyeluruh mengenai penyusunan skripsi ini, maka penulisan laporan dibagi menjadi 5 (lima) bab dengan sistematika sebagai berikut:

BAB I PENDAHULUAN

Bab ini menguraikan fondasi dasar penelitian yang meliputi latar belakang masalah mengenai pentingnya analisis sentimen pada ulasan Bebek Sinjay,

perumusan masalah, batasan ruang lingkup penelitian, tujuan yang ingin dicapai, manfaat penelitian secara teoritis dan praktis, serta sistematika penulisan laporan.

BAB II KAJIAN PUSTAKA

Bab ini memaparkan landasan teori dan kajian pustaka yang menjadi dasar pemecahan masalah. Konsep yang dibahas meliputi definisi Analisis Sentimen, *Text Mining*, pra-pemrosesan teks (*stemming*, tokenisasi), pendekatan *Lexicon-Based*, ekstraksi fitur *Term Frequency-Inverse Document Frequency* (TF-IDF), algoritma klasifikasi *K-Nearest Neighbor* (K-NN) dengan metrik *Euclidean Distance*, pengujian *Cross Validation*, dan *Confusion Matrix*. Selain itu, bab ini juga memuat ulasan mengenai penelitian-penelitian terdahulu yang relevan sebagai perbandingan (*state of the art*).

BAB III METODOLOGI PENELITIAN

Bab ini menjelaskan tahapan dan arsitektur sistem yang dirancang untuk menyelesaikan masalah penelitian. Tahapan tersebut mencakup proses pengumpulan dataset dari *Google Maps*, injeksi *script* kustom Python untuk normalisasi morfologis (*stemming*) dan pelabelan sentimen (*Lexicon-Based*), pra-pemrosesan teks lanjutan menggunakan perangkat lunak *Orange Data Mining*, transformasi data menggunakan TF-IDF, serta skenario pelatihan dan pengujian model K-NN.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menguraikan hasil implementasi dari arsitektur sistem yang telah dirancang pada bab sebelumnya. Pembahasan difokuskan pada hasil komputasi dari setiap tahapan pemrosesan teks, evaluasi kinerja algoritma K-NN menggunakan variasi parameter ketetanggaan (k), serta analisis mendalam terhadap matriks konfusi (*Confusion Matrix*) untuk membuktikan batasan keputusan optimal dalam menangani fenomena paradoks akurasi pada dataset ulasan yang tidak seimbang.

BAB V KESIMPULAN DAN SARAN

Bab ini merupakan bagian penutup yang berisi rangkuman kesimpulan untuk menjawab rumusan masalah yang telah ditetapkan. Selain itu, bab ini juga memberikan saran-saran konstruktif dan rekomendasi bagi pengembangan maupun perbaikan sistem analisis sentimen pada penelitian-penelitian selanjutnya.