

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kemajuan teknologi digital dalam satu dekade terakhir telah membawa perubahan signifikan pada pola transaksi masyarakat Indonesia. Diantara berbagai platform perdagangan digital yang berkembang di Asia Tenggara, Shopee menempati sebagai salah satu marketplace dengan basis pengguna terbesar di Indonesia. Platform ini mencatat transaksi dalam jumlah sangat besar setiap harinya, khususnya pada kategori produk elektronik yang menjadi salah satu segmen dengan permintaan tertinggi. Tingginya volume transaksi tersebut menghasilkan data ulasan dalam jumlah besar yang berisi opini, penilaian, dan pengalaman pembeli terhadap berbagai produk yang telah dibeli.

Data ulasan yang ditinggalkan pembeli setelah bertransaksi memiliki nilai yang strategis tinggi, baik bagi konsumen yang membutuhkan referensi sebelum berbelanja maupun bagi pelaku usaha yang ingin memahami tingkat kepuasan pelanggan terhadap produk mereka. Namun, pertumbuhan volume review yang terus bertambah setiap harinya menyebabkan proses analisis secara manual menjadi tidak efisien dan tidak praktis untuk dilakukan. Diperlukan sebuah pendekatan otomatis yang mampu mengklasifikasikan ulasan dalam jumlah besar secara cepat dan akurat.

Untuk mengatasi permasalahan tersebut, diperlukan pendekatan komputasional yang mampu memproses teks atau perasaan yang terkandung dalam suatu teks secara otomatis. Dengan menerapkan analisis sentimen, ulasan produk yang berjumlah besar dapat diklasifikasikan secara otomatis ke dalam polaritas Positif atau Negatif, sehingga memungkinkan, kemudahan bagi berbagai pihak dalam memahami kecenderungan opini konsumen secara efisien dan terstruktur.

Dari berbagai pendekatan machine learning yang tersedia untuk klasifikasi teks, Multinomial Naïve Bayes dipilih dalam penelitian ini karena penerapannya yang relatif sederhana, proses komputasi yang cepat, serta kemampuannya yang terbukti efektif dalam menangani representasi teks berupa frekuensi kemunculan kata. Pendekatan probabilistik yang mendasari algoritma ini mengasumsikan bahwa setiap fitur kata berkontribusi secara independen terhadap prediksi kelas, sehingga mampu mengklasifikasi dokumen teks ke dalam kelas sentimen yang sesuai secara efisien. Pemilihan varian multinomial pada penelitian ini didasarkan pada kesesuaiannya dengan karakteristik data teks yang direpresentasikan menggunakan nilai frekuensi kemunculan token.

Tantangan yang muncul dalam klasifikasi teks berbasis TF-IDF adalah banyaknya dimensi data yang terjadi saat seluruh kosakata dijadikan fitur. Dalam penelitian ini, proses vektorisasi TF-IDF menghasilkan total 2.678 fitur yang merepresentasikan seluruh kata dan frasa unik dalam dataset. Dimensi data yang tinggi berpotensi memperlambat proses komputasi dan menurunkan performa model akibat keberadaan fitur-fitur yang tidak relevan. Oleh karena itu, penerapan seleksi fitur Chi-Square guna menyaring 500 fitur dengan informatif dan relevan terhadap klasifikasi sentimen dari total 2.678 fitur yang tersedia.

Chi-Square merupakan metode seleksi fitur yang mengukur hubungan statistik antara setiap fitur dengan kelas label sentimen. Penerapan Chi-Square dalam penelitian ini berhasil mereduksi dimensi data sebesar 81,33% dari 2.678 menjadi 500 fitur, sekaligus meningkatkan performa model Naïve Bayes. Hal ini dibuktikan dengan peningkatan akurasi dari 90,04% menjadi 91,10%, peningkatan recall dari 90,04% menjadi 91,10%, serta peningkatan F1-Score dari 87,40% menjadi 89,59% setelah penerapan seleksi fitur Chi-Square.

Penelitian ini membangun sistem klasifikasi sentimen ulasan produk elektronik pada platform Shopee menggunakan algoritma Naïve Bayes yang dioptimasi dengan metode seleksi fitur Chi-Square. Data ulasan dikumpulkan secara otomatis melalui proses web scraping menggunakan bahasa pemrograman Python dengan library Selenium dan Undetected ChromeDriver, kemudian diproses melalui serangkaian tahapan preprocessing teks berbahasa Indonesia menggunakan library Sastrawi. Hasil penelitian menghasilkan model terbaik dengan tingkat akurasi tertinggi sebesar 91,10% dengan nilai cross-validation sebesar $89,21\% \pm 0,76\%$, yang membuktikan bahwa kombinasi Naïve Bayes dan Chi-Square efektif dalam mengklasifikasikan sentimen ulasan produk elektronik berbahasa Indonesia di platform e-commerce Shopee.

1.2 Rumusan Masalah

Berdasarkan permasalahan yang telah dikemukakan pada bagian latar belakang, terdapat beberapa pokok pertanyaan penelitian yang akan dijawab melalui studi ini, yaitu:

1. Bagaimana proses prapemrosesan data (text preprocessing) pada ulasan produk elektronik di Shopee agar dapat diklasifikasi ke dalam sentimen positif dan negatif?
2. Bagaimana penerapan model Naïve Bayes Classifier dalam mengkategorikan data teks ulasan produk elektronik di platform Shopee?
3. Seberapa besar peningkatan akurasi komputasi yang dihasilkan oleh algoritma Naïve Bayes setelah dioptimasi menggunakan seleksi fitur Chi-Square?
4. Seberapa optimal tingkat akurasi model klasifikasi sentimen yang dibangun?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Penelitian ini memiliki tujuan utama, di antaranya:

1. Mengimplementasikan tahapan prapemrosesan teks (text preprocessing) pada data ulasan produk di platform Shopee yang mencakup tahapan case folding, cleansing, tokenizing, penghapusan stopword, dan stemming dengan memanfaatkan library Sastrawi yang dirancang khusus untuk teks berbahasa Indonesia.
2. Menerapkan pendekatan Multinomial Naïve Bayes dalam memetakan polaritas opini konsumen menjadi dua kelas sentimen pada dataset ulasan produk elektronik yang dikumpulkan dari Shopee.
3. Menerapkan metode seleksi fitur Chi-Square yang berhasil mereduksi dimensi data sebesar 81,33% dari 2.678 menjadi 500 fitur sehingga mengoptimalkan kinerja komputasi model Naïve Bayes dalam proses klasifikasi sentimen.
4. Mengevaluasi tingkat akurasi dan performa akhir dari sistem klasifikasi sentimen yang dibangun dengan pengukuran kinerja menggunakan metrik akurasi, presisi, recall, dan F1-Score, di mana model dengan performa terbaik mencapai nilai akurasi sebesar 91,10% dengan nilai cross-validation sebesar $89,21\% \pm 0,76\%$.

1.3.2 Manfaat

Adapun keuntungan yang diperoleh dari penelitian ini sebutkan sebagai berikut:

1. Memperkaya khazanah keilmuan dibidang text mining dan machine learning, khususnya dilingkungan Teknik Informatika UNPRI, sebagai bahan referensi akademik terkait optimasi Naïve Bayes menggunakan Chi-Square pada teks ulasan berbahasa Indonesia
2. Membantu penjual di marketplace Shopee dalam mengevaluasi opini pelanggan secara otomatis terkait produk elektronik yang ditawarkan, guna memahami pandangan konsumen dan meningkatkan kualitas produk serta layanan yang diberikan.
3. Memberikan kemudahan dan efisiensi waktu bagi calon pembeli dalam mendeteksi kecenderungan sentimen suatu produk elektronik (positif atau negatif) secara singkat dan akurat sebelum memutuskan untuk melakukan transaksi di platform Shopee.
4. Menyediakan fondasi empiris yang dapat dimanfaatkan oleh peneliti berikutnya dalam mengembangkan sistem serupa, baik yang menggunakan metodologi yang sama maupun variasi algoritma lain di ranah akademik dan industri.

1.4 Keterbatasan Masalah

Dalam upaya menjaga relevansi penelitian, batasan-batasan yang diterapkan dalam studi ini meliputi:

1. Data yang digunakan adalah ulasan produk elektronik yang mencakup kategori televisi, handphone, laptop, dan headset di platform e-commerce Shopee Indonesia, yang dihimpun secara otomatis menggunakan web scraping berbasis Python yang menghasilkan 2.118 rekaman data dari 6 file CSV.
2. Klasifikasi sentimen hanya dikategorikan ke dalam 2 kelas yaitu sentimen Positif (rating 4–5 bintang) dan sentimen Negatif (rating 1–2 bintang). Ulasan dengan rating 3 bintang tidak diikutsertakan dalam dataset karena diasumsikan bersifat netral.
3. Algoritma klasifikasi yang diterapkan adalah Multinomial Naïve Bayes dengan parameter Laplace smoothing ($\alpha = 1,0$).
4. Metode pemilihan fitur yang digunakan adalah Chi-Square dengan jumlah 500 fitur terpilih dari total 2.678 fitur hasil vektorisasi TF-IDF.
5. Prosedur analitik dijalankan menggunakan metodologi pemrosesan teks yang mencakup normalisasi kasus, pemurnian data, tokenisasi, penghapusan kata kunci (penyaringan), dan stemming, menggunakan perpustakaan Sastrawi yang dirancang khusus untuk bahasa Indonesia.