JUSTIFY: Jurnal Sistem Informasi Ibrahimy
Vol. 5, No. 1, Juli 2026, hal. 1-10
E-ISSN: 2962-3642 | P-ISSN: 2962-3871
DOI: 10.35316/justify.v4i2.9154

Seleksi Fitur *Information Gain* pada Klasifikasi Ulasan *Traveloka* di *Google Play Store* Menggunakan *Naïve Bayes*

Putri Noer Kumalasari[1], Moh Dasuki[2], Miftahur Rohman[3]

[1,3] Teknik Informatika, Teknik, Universitas Muhammadiya Jember, Indonesia
[2] Teknik Informatika, Teknik, Universitas Muhammadiyah Jember, Indonesia

Info Artikel

Riwayat Artikel:

Diterima : **28-Mei-2026**
Direvisi : **19-Juni-2026**
Disetujui : **6-Juli-2026**

Kata Kunci:

analisis sentimen,
Naïve Bayes,
web scraping,
Traveloka,
Information Gain,
Bag of Word,

ABSTRAK

Penelitian ini membandingkan analisis sentimen pengguna aplikasi Traveloka menggunakan algoritma Multinomial Naïve Bayes pada dua skenario, yaitu dengan seleksi fitur Information Gain dan tanpa seleksi fitur. Data penelitian diperoleh melalui teknik web scraping terhadap Google Play Store menggunakan google-play-scraper, dengan total sebanyak 500 komentar berbahasa Indonesia. Proses pelabelan sentimen dilakukan menggunakan dua pendekatan, yaitu pelabelan manual oleh pakar bahasa dan pelabelan otomatis oleh peneliti. Data yang telah dikumpulkan selanjutnya melalui tahapan preprocessing yang meliputi cleaning data, tokenizing, case folding, stopword removal, dan stemming. Pada skenario pertama, seleksi fitur dilakukan menggunakan metode Information Gain untuk mengidentifikasi kata-kata yang paling berpengaruh terhadap klasifikasi sentimen, sedangkan pada skenario kedua digunakan pendekatan Bag of Words tanpa seleksi fitur. Evaluasi model dilakukan menggunakan metode 5-Fold Cross Validation dengan metrik akurasi, precision, recall, f1-score, dan confusion matrix, dengan membandingkan hasil prediksi model terhadap label pakar bahasa. Hasil penelitian menunjukkan bahwa model Naïve Bayes dengan seleksi fitur Information Gain memperoleh rata-rata akurasi sebesar 87,4%, sedangkan model tanpa seleksi fitur menghasilkan akurasi sebesar 86,0%. Hasil ini menunjukkan bahwa penerapan seleksi fitur Information Gain mampu meningkatkan kinerja dan efisiensi model dalam analisis sentimen ulasan pengguna.

***Keywords*:**

sentiment analysis,
Naïve Bayes,
web scraping,
Traveloka,
Information Gain,
Bag of Word,

ABSTRACT

This study compares sentiment analysis of Traveloka application user reviews using the Multinomial Naïve Bayes algorithm under two scenarios: with Information Gain feature selection and without feature selection. The research data were collected through web scraping from the Google Play Store using the google-play-scraper, resulting in a total of 500 Indonesian-language reviews. Sentiment labeling was conducted using two approaches: manual labeling by a language expert and automatic labeling by the researcher. The collected data then underwent preprocessing stages including data cleaning, tokenizing, case folding, stopword removal, and stemming. In the first scenario, feature selection was applied using the Information Gain method to identify the most influential words for sentiment classification, while in the second scenario a Bag of Words approach without feature selection was employed. Model evaluation was performed using 5-Fold Cross Validation with accuracy, precision, recall, f1-score, and confusion matrix as evaluation metrics, by comparing the model predictions with expert-labeled data. The results show that the Naïve Bayes model with Information Gain feature selection achieved an average accuracy of 87.4%, while the model without feature selection achieved an accuracy of 86.0%. These findings indicate that the application of Information Gain feature selection improves both the performance and efficiency of the sentiment classification model.

Penulis Korespondensi:
Putri Noer Kumalasari,
Teknik Informatika,
Universitas Muhammadiyah Jember
Email: Putrinoerkumalasarii@gmail.com

1. PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi digital yang berlangsung sangat cepat telah mengubah cara masyarakat mengakses dan memanfaatkan berbagai layanan, khususnya pada sektor transportasi dan pariwisata. Ketersediaan internet yang semakin luas mendorong munculnya berbagai platform berbasis aplikasi yang menawarkan layanan serba cepat, praktis, dan terintegrasi. Salah satu platform yang menonjol di kawasan Asia Tenggara adalah Traveloka, yang menyediakan layanan pemesanan tiket transportasi, akomodasi, hingga berbagai produk pendukung perjalanan dalam satu aplikasi terpadu.

Berdasarkan data dari *Databoks* Katadata, *Traveloka* tercatat sebagai situs perjalanan dengan trafik tertinggi di Indonesia, dengan jumlah kunjungan mencapai 7,2 juta pada Maret 2022 dan meningkat secara signifikan menjadi 22,3 juta pada April 2024 [1]. Selain itu, aplikasi Traveloka telah diunduh lebih dari 50 juta kali di *Google Play Store*, yang menunjukkan tingkat adopsi pengguna yang tinggi serta kepercayaan konsumen terhadap layanan yang ditawarkan. Pertumbuhan jumlah pengguna ini secara langsung berdampak pada meningkatnya volume ulasan dan umpan balik yang diberikan oleh pengguna aplikasi.

Ulasan pengguna tidak hanya mencerminkan tingkat kepuasan, tetapi juga mengandung kritik, saran, serta pengalaman pengguna yang berharga. Informasi tersebut dapat dimanfaatkan sebagai dasar evaluasi kualitas layanan dan pengambilan keputusan strategis bagi perusahaan. Namun, besarnya jumlah ulasan menyebabkan proses analisis manual menjadi tidak efektif dan tidak efisien, sehingga diperlukan pendekatan berbasis otomatisasi.

1.2. Permasalahan Dan Tantangan

Seiring meningkatnya jumlah ulasan yang masuk setiap hari, perusahaan dihadapkan pada tantangan untuk memahami persepsi pengguna secara cepat dan akurat. Analisis sentimen merupakan salah satu pendekatan dalam *text mining* yang digunakan untuk mengidentifikasi opini, emosi, dan kecenderungan sikap pengguna terhadap suatu produk atau layanan [2]. Melalui analisis sentimen, ulasan teks dapat diklasifikasikan ke dalam kategori positif, netral, atau negatif.

Salah satu algoritma yang banyak digunakan dalam analisis sentimen adalah *Naïve Bayes Classifier* (NBC). Algoritma ini dikenal memiliki keunggulan dalam kesederhanaan, kecepatan komputasi, serta kemampuannya menangani data berukuran besar dan mengandung *noise* [3]. Meskipun demikian, tantangan utama dalam klasifikasi teks menggunakan NBC adalah tingginya jumlah fitur yang dihasilkan dari proses ekstraksi kata, seperti tokenisasi dan pembobotan berbasis *Bag of Words*. Jumlah fitur yang terlalu besar dapat meningkatkan kompleksitas model, memperbesar risiko *overfitting*, serta menurunkan efisiensi komputasi.

1.3. Tinjauan Penelitian Terkait

Berbagai penelitian sebelumnya telah membuktikan efektivitas algoritma *Naïve Bayes* dalam analisis sentimen ulasan aplikasi. Rita dan Pambudi menunjukkan bahwa penggunaan seleksi fitur *Information Gain* mampu meningkatkan performa klasifikasi pada ulasan aplikasi PeduliLindungi dengan nilai *F1-score* yang tinggi [4]. Penelitian lain oleh Hasibuan dan Heriyanto membuktikan bahwa *Multinomial Naïve Bayes* efektif dalam mengklasifikasikan sentimen ulasan *Amazon Shopping di Google Play Store* [5]. Hasil serupa juga ditunjukkan pada penelitian ulasan aplikasi *Maxim dan BRIMO*, yang memperlihatkan performa NBC yang konsisten dalam klasifikasi sentimen [6], [7].

Meskipun demikian, sebagian penelitian masih berfokus pada akurasi keseluruhan tanpa membahas secara mendalam dampak seleksi fitur terhadap keseimbangan performa antar kelas sentimen, khususnya pada kelas minoritas seperti netral dan negatif.

1.4. Pendekatan Yang Diusulkan Dan Nilai Inovasi

Untuk mengatasi permasalahan tingginya dimensi fitur, penelitian ini mengusulkan penggunaan seleksi fitur *Information Gain* (IG) sebagai tahap preprocessing sebelum proses klasifikasi menggunakan *Naïve Bayes Classifier*. *Information Gain* berfungsi untuk mengukur tingkat kontribusi suatu fitur terhadap pengurangan ketidakpastian dalam proses klasifikasi, sehingga hanya fitur-fitur yang paling relevan yang digunakan dalam pemodelan [8].

Pendekatan ini diharapkan mampu mengurangi kompleksitas model, meningkatkan efisiensi komputasi, serta memperbaiki kinerja klasifikasi, khususnya pada data ulasan yang mengandung variasi bahasa, slang, dan struktur kalimat tidak baku seperti pada *Google Play Store*.

1.5. Kontribusi Dan Kebaruan Penelitian

Kontribusi utama dari penelitian ini terletak pada analisis komparatif kinerja *Naïve Bayes Classifier* dengan dan tanpa seleksi fitur *Information Gain* pada dataset ulasan Traveloka berbahasa Indonesia. Penelitian ini tidak hanya mengevaluasi akurasi, tetapi juga membandingkan metrik presisi, *recall*, dan *F1-score* pada setiap kelas sentimen. Dengan demikian, penelitian ini memberikan wawasan yang lebih komprehensif mengenai pentingnya seleksi fitur dalam menghasilkan model analisis sentimen yang seimbang, efisien, dan relevan dengan kebutuhan industri digital. Hasil penelitian diharapkan dapat dimanfaatkan sebagai referensi akademik sekaligus sebagai dasar pengambilan keputusan strategis dalam peningkatan kualitas layanan berbasis umpan balik pengguna.

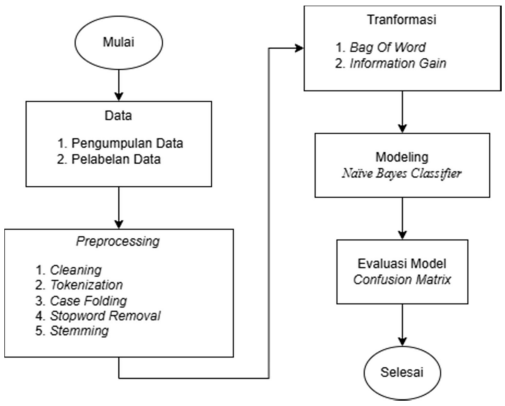
2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini bertujuan untuk menjelaskan secara sistematis tahapan-tahapan yang dilakukan dalam menyelesaikan permasalahan penelitian, mulai dari perancangan konsep hingga evaluasi hasil. Secara umum, metode penelitian ini terdiri atas perancangan konsep, pengumpulan data, produksi dan pemrosesan data, pengujian dan evaluasi, serta penyajian hasil penelitian. Alur penelitian disusun secara terstruktur agar setiap tahapan dapat dipertanggungjawabkan secara ilmiah

2.1 Desain Dan Kronologis Penelitian

Desain penelitian ini menggunakan pendekatan eksperimental kuantitatif, dengan fokus pada analisis sentimen ulasan pengguna aplikasi *Traveloka* menggunakan metode *Naïve Bayes Classifier* (NBC). Alur penelitian dimulai dari pengumpulan data ulasan, dilanjutkan dengan preprocessing data teks, seleksi fitur, pemodelan klasifikasi, hingga evaluasi performa model.

Tahapan penelitian divisualisasikan dalam Gambar 2.1 Tahapan Penelitian, yang menunjukkan alur kerja penelitian secara keseluruhan, mulai dari pengumpulan data hingga evaluasi model



Gambar 2.1. Tahapan Penelitian

2.2 Pengumpulan Dan Perolehan Data

Data yang digunakan dalam penelitian ini berupa ulasan pengguna aplikasi *Traveloka* yang diperoleh dari *Google Play Store*. Pengumpulan data dilakukan menggunakan metode *web scraping*, sehingga data yang diperoleh bersifat aktual dan merepresentasikan pengalaman pengguna secara langsung.

Data yang terkumpul kemudian diberi label sentimen (positif, negatif, dan netral). Proses pelabelan dilakukan oleh pakar bahasa Indonesia, sehingga keakuratan klasifikasi sentimen dapat dipertanggungjawabkan secara linguistik dan ilmiah.

2.3 Prosedur Penelitian (Algoritma Dan Tahapan)

Prosedur penelitian dilaksanakan melalui beberapa tahapan utama sebagai berikut:

2.3.1 *Preprocessing* Data Teks

Tahapan *preprocessing* bertujuan untuk membersihkan dan menyiapkan data agar siap digunakan dalam proses klasifikasi. Tahapan ini meliputi beberapa proses, yang masing-masing didukung oleh contoh data yang disajikan dalam bentuk tabel, antara lain:

1. *Cleaning*

   Pada tahap ini dilakukan pembersihan teks dari elemen tidak relevan seperti tanda baca, angka, simbol, emoji, dan karakter khusus untuk meningkatkan kualitas data sebelum analisis sentimen.

2. *Tokenization*

   Pada tahap ini, teks dipecah menjadi unit-unit kecil (*token*) berupa kata atau tanda baca untuk memudahkan pemrosesan dan analisis lanjutan.

3. *Case Folding*

   Tahap *case folding* mengubah seluruh teks menjadi huruf kecil untuk menjaga konsistensi kata dan memudahkan analisis.

4. *Stopword Removal*

   Pada tahap ini dilakukan penghapusan *stopwords*, yaitu kata-kata umum yang tidak memiliki kontribusi signifikan dalam analisis teks, untuk meningkatkan efisiensi model.

5. *Stemming*

   Tahap *stemming* mengubah kata berimbuhan menjadi bentuk dasar untuk menyederhanakan analisis teks.

2.3.2 Representasi Fitur *(Bag Of Words)*

Setelah *preprocessing*, data direpresentasikan menggunakan metode *Bag of Words (BoW)*, di mana setiap ulasan diubah menjadi vektor frekuensi kata. Hasil representasi ini ditampilkan dalam:

Tabel 2.1 Data Ulasan Setelah *Preprocessing*

No	Nama Pengguna	Ulasan
1	Andi Pratama	"aplikasi ini bagus banget"
2	Bunga Sari	"tidak bisa login parah"
3	Citra Lestari	"pesan tiket 2x tapi gagal"
4	Dedi Kurnia	"pesan tiket via traveloka super cepat"
5	Eka Putri	"cs lambat banget tidak respon"

Tabel 2.2 Pemecahan Kalimat Menjadi Kata

No	Nama Pengguna	Ulasan
1	Andi Pratama	“aplikasi”, “ini”, “bagus”, “banget”
2	Bunga Sari	“tidak”, “bisa”, “login”, “parah”
3	Citra Lestari	“pesan”, “tiket”, “2x”, “tapi”, “gagal”
4	Dedi Kurnia	“pesan”, “tiket”, “via”, “traveloka”, “super”, “cepat”
5	Eka Putri	“cs”, “lambat”, “banget”, “tidak”, “respon”

Tabel 2.3 Frekuensi Kata

Kata	No Ulasan					Frekuensi
	1	2	3	4	5	
Aplikasi	1	0	0	0	0	1
Ini	1	0	0	0	0	1
Bagus	1	0	0	0	0	1
Banget	1	0	0	0	0	2
Tidak	0	1	0	0	1	2
Bisa	0	1	0	0	0	1
Login	0	1	0	0	0	1
Parah	0	1	0	0	0	1
Pesan	0	0	1	1	0	2
Tiket	0	0	1	1	0	2
2x	0	0	1	0	0	1
Tapi	0	0	1	0	0	1
Gagal	0	0	1	0	0	1
Via	0	0	0	1	0	1
Traveloka	0	0	0	1	0	1
Super	0	0	0	1	0	1
Cepat	0	0	0	1	0	1

cs	0	0	0	0	1	1
Lambat	0	0	0	0	1	1
Respon	0	0	0	0	1	1

2.3.3 Seleksi Fitur *Information Gain*

Information Gain (IG) adalah metode dalam seleksi fitur yang digunakan dalam analisis sentimen dan pembelajaran mesin untuk menentukan seberapa banyak suatu fitur (kata atau atribut) dapat mengurangi ketidakpastian dalam klasifikasi [19]. Information Gain adalah sebuah metode seleksifitur yang mengukur sejauh mana kehadiran atau ketidakhadiran sebuah kata berpengaruh dalam membuat keputusan klasifikasi yang akurat pada kelas data tertentu. Tujuannya adalah untuk mengurangi dimensi data dan meningkatkan akurasi klasifikasi [4].

Dikutip dari [1] *Information Gain* (IG) dengan menggunakan rumus berikut:

$$IG\ (A) = H\ (D) - H\ (D|\ A) \qquad (1)$$

Keterangan:

- *IG* (*A*) = *Information Gain* dari fitur A (kata atau atribut dalam ulasan).
- *H* (*D*) = Entropi sebelum pemisahan berdasarkan fitur A (entropi keseluruhan dataset).
- *H* (*D*|*A*) = Entropi setelah pemisahan berdasarkan fitur A (entropi kondisional).

Dikutip dari [14] dan [17] beberapa hal yang harus diperhitungkan untuk mengetahui nilai dari information gain pada sebuah fitur sebagai berikut:
Adapun rumus untuk perhitungan probabilitas prior:

$$P\ (C_i) = \frac{Jumlah\ Data\ Dalam\ Kelas\ C_i}{Total\ Jumlah\ Data} \qquad (2)$$

Probabilitas fitur digunakan untuk mengukur **s**eberapa sering sebuah fitur (kata) muncul dalam keseluruhan data atau dalam data dengan kelas tertentu:

$$P\ X_i = \frac{Jumlah\ data\ yang\ mengandung\ fitur\ X_i}{Total\ Jumlah\ Data} \qquad (3)$$

$$P\ (X_i|C_i) = \frac{Jumlah\ data\ dalam\ kelas\ C_i\ yang\ mengandung\ fitur\ X_i}{Total\ jumlah\ data\ dalam\ kelas\ C_i} \qquad (4)$$

2.3.3.1 Menghitung Entropi

Entropi adalah ukuran ketidakteraturan atau ketidakpastian dalam data.
Rumus entropi adalah:

$$H\ (D) = -\ \sum_{i=1}^{n} P\ (c_i) \log_2 P\ (\mathrm{c}_i) \qquad (5)$$

Keterangan:

- *H* (*D*) = Entropi dataset *D*.
- *P* (Ci) = Probabilitas suatu kategori sentimen Ci (positif, netral, negatif).
- n = Jumlah kategori dalam dataset.

Entropi bernilai 0 jika semua data termasuk dalam satu kelas maka tidak ada kebingungan atau tidak ada ketidakpastian. Itulah sebabnya entropi bernilai 0. Entropi bernilai maksimum jika data terbagi rata ke dalam semua kelas.

2.3.3.2 Menghitung Entropi Kondisional

Setelah memisahkan data berdasarkan fitur tertentu kita menghitung entropi baru:

$$H\ (D|X_i) = \sum_{v \in A} P\ (X_i) \times H(X_i|C_i) \qquad (6)$$

Keterangan:

- X_i= Nilai dari fitur .
- *P* (X_i) = Probabilitas kejadian X_i dalam dataset.
- *H* ($X_i|C_i$)= Entropi subset data setelah dipisahkan oleh X_i.

Semakin kecil entropi setelah pemisahan, semakin tinggi *Information Gain*, yang berarti fitur tersebut sangat berguna dalam klasifikasi sentimen.

2.4 Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) adalah metode klasifikasi yang berdasar pada teorema *Bayes*. *Naïve Bayes Classifier* (NBC) bersifat sederhana, walaupun demikian klasifikasi ini memberikan hasil yang baik dan juga memberikan kecepatan dalam memproses data dalam kuantitas yang besar (Syafitri Kustanto et al., 2022). *Naïve Bayes Classifier* (NBC) mengasumsikan bahwa ada atau tidaknya sebuah fitur dalam sebuah kelas adalah independen (tidak bergantung), artinya sebuah fitur dalam sebuah kelas tidak punya keterkaitan dengan keberadaan fitur lain dari data yang sama (Jurafsky, 2019).
Rumus umum untuk pemodelan *naïve bayes*:

$$P\,(c|x_1,x_2,\dots\dots,x_n,) = P(c) \times \prod_{i=1}^{n} P\,(x_i|C_i) \qquad (7)$$

Keterangan:

- C_i adalah label kelas
- X_i adalah fitur ke-i
- n adalah jumlah fitur
- $P(C_i)$ adalah probabilitas prior
- $P(x_i|C_1)$ adalah probabilitas fitur x_i muncul dalam kelas c

Menurut [17] Probabilitas kondisional digunakan untuk menghitung peluang munculnya setiap kata pada masing-masing kelas. Untuk menghindari nilai $P(X_i|Kelas)$ = 0 dilakukan laplace smoothing yaitu penambahan suatu nilai, misalkan parameter α, pada setiap perhitungan banyaknya kemunculan kata. Pendekatan paling sederhana adalah menambahkan satu nilai (α = 1) ke setiap jumlah peristiwa yang diamati termasuk kemungkinan hitungan nol. Adapun rumus dari *laplace smoothing* sebagai berikut:

$$P(X_i|C_i) = \frac{Count\left(x_{i,C_i}\right)+1}{Count(C_i)+k} \qquad (8)$$

Keterangan:

- *Count* $(X_{i,}\ C_i)$ adalah jumlah kemunculan nilai fitur Xi dalam kelas.
- *Count* (C_i) adalah total data dalam kelas.
- k adalah jumlah nilai yang kemungkinan dari fitur (misalnya, untuk fitur biner seperti "fitur", k=2 karena nilainya 0 atau 1). Sedangkan k adalah jumlah kata unik dalam kosa kata untuk pendekatan *bag of word* jika tanpa adanya seleksi fitur.
- +1 adalah penambahan satu untuk menghindari probabilitas nol.

Dari hasil perhitungan manual yang dilakukan oleh [16] didalam artikelnya dikatakan bahwa dari hasil perhitungan probabilitas tersebut, dapat disimpulkan bahwa dokumen Text474 termasuk dalam kelas negatif dikarenakan P(Negatif|Text474) lebih besar daripada P(Positif|Text474) dan P(Netral|Text474). Artinya kelas yang memiliki nilai tertinggi dianggap sebagai hasil prediksinya.

2.5 Cara Pengujian Dan Evaluasi Model

Pengujian model dilakukan menggunakan metode *K-Fold Cross Validation* dengan jumlah lipatan sebanyak 5 *fold* (K=5). Setiap fold digunakan secara bergantian sebagai data latih dan data uji untuk memperoleh hasil evaluasi yang lebih objektif dan stabil.

Evaluasi performa model dilakukan dengan menggunakan *confusion matrix*, serta pengukuran metrik:

1. Akurasi
2. Presisi
3. *Recall*
4. *F1-Score*

Hasil evaluasi ditampilkan dalam table dibawah ini:

Tabel 2. 4 *Confusion Matrix*

		Prediksi		
		Kelas A	**Kelas B**	**Kelas C**
Aktual	**Kelas A**	AA	AB	AC
	Kelas B	BA	BB	BC
	Kelas C	CA	CB	CC

Keterangan

- AA : Nilai positif yang diperkirakan positif
- AB : Nilai positif yang diperkirakan negatif

- AC : Nilai positif yang diperkirakan netral
- BA : Nilai negatif yang diperkirakan positif
- BB : Nilai negatif yang diperkirakan negatif
- BC : Nilai negatif yang diperkirakan netral
- CA : Nilai netral yang diperkirakan positif
- CB : Nilai netral yang diperkirakan negatif
- CC : Nilai netral yang diperkirakan netral

Akurasi, presisi, *recall* dan *F-1 score* dapat dihitung menggunakan informasi yang diberikan dalam *confusion matrix*. Akurasi merupakan seberapa akurat model dapat mengklasifikasi dengan benar, nilai ini dapat dihitung dengan membagi jumlah prediksi yang benar dengan jumlah total sample. *Presisi* menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model, nilai *presisi* yang besar menunjukkan rendahnya nilai *false negative*. Menggabungkan nilai *presisi* dan *recall* akan menghasilkan matrix tunggal yang dikenal sebagi *F1-score*, yang merupakan rata-rata harmonik tertimbang dari *presisi* dan *recall* [4].

Tabel 2. 5 Formula *Confusion Matrix*

Metrix	Formula
Akurasi	$\frac{AA + BB + CC}{AA + AB + AC + BA + BB + BC + CA + CB + CC}$
Recall	$\frac{AA}{AA + AB + AC}$
Presisi	$\frac{AA}{AA + BA + CA}$
F1-Score	$2 \times \frac{Presisi \times Recall}{Presisi + Recall}$

3. HASIL DAN ANALISIS

3.1 PENGUMPULAN DAN PELABELAN DATA

Data penelitian diperoleh melalui teknik *web scraping* terhadap ulasan pengguna aplikasi Traveloka di Google Play Store menggunakan *library* Google Play Scraper. Sebanyak 500 ulasan berbahasa Indonesia berhasil dikumpulkan dan disimpan dalam format CSV, Excel, dan JSON.
Pelabelan sentimen dilakukan dengan dua pendekatan, yaitu pelabelan manual oleh pakar bahasa dan pelabelan otomatis oleh peneliti menggunakan model berbasis *transformer*. Label dari peneliti digunakan sebagai data latih, sedangkan label pakar dijadikan sebagai acuan evaluasi untuk mengukur tingkat kesesuaian hasil klasifikasi sistem dengan penilaian ahli.

3.2 Preprocessing Data

Tahapan *preprocessing* meliputi cleaning data, tokenizing, case folding, stopword removal, dan stemming. Proses ini bertujuan untuk membersihkan teks dari unsur yang tidak relevan serta menyeragamkan bentuk kata agar siap digunakan dalam pemodelan. Library SpaCy, NLTK, dan Sastrawi digunakan untuk mendukung proses pemrosesan bahasa Indonesia. Hasil preprocessing menghasilkan data teks yang lebih terstruktur dan representatif terhadap isi ulasan pengguna.

3.3 Seleksi Fitur Information Gain

Seleksi fitur dilakukan menggunakan metode Information Gain (IG) untuk mengidentifikasi kata-kata yang paling berpengaruh terhadap klasifikasi sentimen. Perhitungan IG menunjukkan bahwa kata-kata seperti *buruk, jelek,* dan *promo* memiliki nilai IG tertinggi, sehingga dianggap paling informatif dalam membedakan kelas sentimen. Hasil seleksi fitur ini digunakan untuk membangun model klasifikasi yang lebih efisien dengan mengurangi dimensi data dan meminimalkan fitur yang kurang relevan.

3.4 Pemodelan Naive Bayes

Penelitian ini menggunakan algoritma Multinomial Naïve Bayes dengan skema 5-Fold Cross Validation. Dua skenario pemodelan dilakukan, yaitu:

1. Naive Bayes dengan seleksi fitur Information Gain
2. Naive Bayes tanpa seleksi fitur

Evaluasi dilakukan menggunakan akurasi, confusion matrix, precision, recall, dan f1-score, dengan membandingkan hasil prediksi model terhadap label pakar bahasa.

3.5 Evaluasi dan Analisis Hasil

Hasil pengujian menunjukkan bahwa model Naive Bayes dengan seleksi fitur Information Gain memperoleh rata-rata akurasi sebesar 87,4%, sedangkan model tanpa seleksi fitur menghasilkan akurasi rata-rata sebesar 86,0%.
Peningkatan akurasi menunjukkan bahwa penerapan seleksi fitur Information Gain mampu meningkatkan performa model dengan menghilangkan kata-kata yang kurang relevan serta memperkuat kontribusi fitur yang informatif. Analisis confusion matrix juga memperlihatkan bahwa kesalahan klasifikasi paling banyak terjadi pada kelas netral, yang cenderung memiliki karakteristik kata yang tumpang tindih dengan kelas positif dan negatif.

3.6 Pembahasan

Berdasarkan hasil pengujian, dapat disimpulkan bahwa penggunaan seleksi fitur Information Gain pada klasifikasi sentimen ulasan Traveloka terbukti meningkatkan kinerja model Naive Bayes. Selain menghasilkan akurasi yang lebih tinggi, model juga menjadi lebih efisien dalam pemrosesan fitur. Hasil ini menunjukkan bahwa kombinasi preprocessing yang tepat, seleksi fitur, dan algoritma klasifikasi sederhana seperti Naive Bayes masih sangat efektif untuk analisis sentimen berbasis teks ulasan pengguna.

4. KESIMPULAN

1. Penerapan Seleksi Fitur *Information Gain* dalam Mengidentifikasi Fitur yang Paling Relevan.

   Seleksi fitur Information Gain (IG) memilih kata-kata dengan kontribusi tertinggi terhadap klasifikasi sentimen sehingga hanya fitur yang relevan digunakan dalam pemodelan. Meskipun mampu mengurangi kompleksitas data dan meningkatkan fokus model, hasil evaluasi menunjukkan bahwa fitur yang terpilih lebih banyak merepresentasikan kelas mayoritas, yaitu positif. Akibatnya, model menjadi sangat baik dalam mengenali ulasan positif tetapi kurang optimal dalam mendeteksi sentimen netral dan negatif. Hal ini mengindikasikan bahwa meskipun IG efektif secara statistik, fitur yang dipilih belum sepenuhnya mampu mewakili variasi kata dari semua kelas secara seimbang.

2. Pengaruh Seleksi Fitur *Information Gain* terhadap Akurasi, Presisi, Recall, dan F1-Score.

   Berdasarkan hasil pengujian, penerapan seleksi fitur Information Gain terbukti memberikan peningkatan performa pada model Multinomial Naïve Bayes. Hal ini terlihat dari kenaikan rata-rata akurasi dari 0.8600 menjadi 0.8740 setelah seleksi fitur diterapkan. Selain itu, nilai presisi, recall, dan f1-score pada model dengan seleksi fitur juga menunjukkan hasil yang lebih stabil, terutama pada kelas positif dan negatif. Proses seleksi fitur membantu model mengurangi kata-kata yang tidak relevan sehingga prediksi menjadi lebih tepat dan konsisten. Dengan demikian, Information Gain berpengaruh positif dalam meningkatkan kualitas klasifikasi sentimen dibandingkan model tanpa seleksi fitur.

3. Perbandingan Akurasi Pemodelan dengan dan tanpa Seleksi *Fitur Information Gain.*

   Perbandingan akurasi antara dua pendekatan menunjukkan bahwa dengan adanya penambahan menggunakan pembagian data menggunakan K-Fold Cross Validation kedunya memiliki akurasi sebagai berikut.

Tabel 3. 1 Hasil Akurasi Dengan Seleksi Fitur

Dengan Seleksi Fitur IG		
Fold	Akurasi	Prediksi Benar
1	89 %	89 dari 100 data
2	87%	87 dari 100 data
3	84%	84 dari 100 data
4	91%	91 dari 100 data
5	86%	86 dari 100 data

Tabel 3. 2 Hasil Akurasi Tanpa Seleksi Fitur

Tanpa Seleksi Fitur IG		
Fold	Akurasi	Prediksi Benar
1	89 %	89 dari 100 data
2	86%	86 dari 100 data
3	83%	83 dari 100 data
4	88%	88 dari 100 data
5	84%	84 dari 100 data

REFERENSI

[1] Cindy caterine yolanda, syafriandi syafriandi, yenni kurniawati, & dina fitria. (2024). sentiment analysis of dana application reviews on google play store using naïve bayes classifier algorithm based on information gain. unp journal of statistics and data science

[2] Endang Wahyu Pamungkas, D. G. P. P. (2016). An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia. IEEE, 1–23.

[3] Firmahsyah, f., & gantini, t. (2016). penerapan metode content-based filtering pada sistem rekomendasi kegiatan ekstrakulikuler (studi kasus di sekolah abc). jurnal teknik informatika dan sistem informasi, 2(3). https://doi.org/10.28932/jutisi.v2i3.548

[4] Hasibuan, e., & heriyanto, e. a. (n.d.-b). analisis sentimen pada ulasan aplikasi amazon shopping di google play store menggunakan naive bayes classifier. jts, 1(3).

[5] Https://www.traveloka.com/id-id/about-us (diakses : 21 September 2025)

[6] Hudaya, c. s., fakhrurroja, h., & alamsyah, a. (2019). analisis persepsi konsumen terhadap brand go-jek pada media sosial twitter menggunakan metode sentiment analysis dan topic modelling. jurnal mitra manajemen, 3(6), 664–673. https://doi.org/10.52160/ejmm.v3i6.244

[7] Insan, m. k., hayati, u., & nurdiawan, o. (2023). analisis sentimen aplikasi brimo pada ulasan pengguna di. jurnal mahasiswa teknik informatika, 7(1), 478–483.

[8] Maulida surbakti, n., talia, a., br perangin-angin, c., olivia nainggolan, d., devi friskauly, n., & ruth br tumorang, s. (2024). penggunaan bahasa pemrograman python dalam pembelajaran kalkulus fungsi dua variabel. kebumian dan angkasa, 2(3), 98–107.

[9] Normawati, d., & prayogi, s. a. (2021). implementasi naïve bayes classifier dan confusion matrix pada analisis sentimen berbasis teks pada twitter. jurnal sains komputer & informatika (j-sakti, 5(2), 697–711.

[10] Novita, d., & helena, f. (2021). analisis kepuasan pengguna aplikasi traveloka menggunakan metode technology acceptance model (tam) dan end-user computing satisfaction (eucs). jurnal teknologi sistem informasi, 2(1), 22–37. https://doi.org/10.35957/jtsi.v2i1.846

[11] Prabowo, w. a., & wiguna, c. (2021). sistem informasi umkm bengkel berbasis web menggunakan metode scrum. jurnal media informatika budidarma, 5(1), 149. https://doi.org/10.30865/mib.v5i1.2604

[12] Pratama, s. f., andrean, r., & nugroho, a. (2019). analisis sentimen twitter debat calon presiden indonesia menggunakan metode fined-grained sentiment analysis. jointecs (journal of information technology and computer science), 4(2), 39. https://doi.org/10.31328/jointecs.v4i2.1004

[13] Rita, s., & pambudi, a. (2023). penggunaan support vector machine untuk analisis sentimen ulasan aplikasi truecaller dan getcontact using support vector machine for sentiment analysis of truecaller and getcontact app reviews (vol. 20, issue 2).

[14] Salsabila, s. n., sari, b. n., & mayasari, r. (2023). klasifikasi ulasan pengguna aplikasi discord menggunakan metode information gain dan naïve bayes classifier. infotech journal, 9(2), 383–392. https://doi.org/10.31949/infotech.v9i2.6277

[15] Saputra, a., & noor hasan, f. (2023). analisis sentimen terhadap aplikasi coffee meets bagel dengan algoritma naïve bayes classifier. sibatik journal: jurnal ilmiah bidang sosial, ekonomi, budaya, teknologi, dan pendidikan, 2(2), 465–474. https://doi.org/10.54443/sibatik.v2i2.579

[16] Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online JD.ID Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. Jurnal SIMETRIS, 10(2), 681–686.

[17] Syafitri kustanto, n., gusriani, n., & firdaniza. (2022). analisis sentimen dengan metode klasifikasi naïve bayes dan seleksi fitur information gain (studi kasus: ulasan aplikasi pedulilindungi). in search, 21(02), 134–144.

[18] Yanuar, r. a. a. (2024). jurnal teknik informatika, vol. 16, no. 2, april 2024. 16(2), 1–7.

[19] Yuni, k., & santi, i. g. (2024). analisis sentimen ulasan traveloka menggunakan metode naïve bayes classifier dan information gain. 3(november), 63–70.