ANTIVIRUS: Jurnal Ilmiah Teknik Informatika (p – ISSN: 1978 – 5232; e – ISSN: 2527 – 337X)
Vol. 20 No. 1 Mei 2026, pp. 98 – 109

NAÏVE BAYES SENTIMENT ACCURACY WITH CHI-SQUARE AND INFORMATION GAIN

Diterima Redaksi: 27 November 2025; Revisi Akhir: 26 Mei 2026; Diterbitkan Online: 31 Mei 2026

Muhamad Fahrurozi[1)], Dian Ade Kurnia[2)], Yudhistira Arie Wijaya[3)], Puji Pramudya Marta[4)], Khaerul Anam[5)]
[1, 5)] Teknik Informatika, STMIK IKMI Cirebon
[2)] Manajemen Informatika, STMIK IKMI Cirebon
[3, 4)] Sistem Informasi, STMIK IKMI Cirebon
[1, 2, 3, 4, 5)] Jl Perjuangan No. 10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat 45135, Indonesia
e-mail: muhammadfahrurozi148@gmail.com[1)], dianade2012@gmail.com[2)], yudhistira010471@gmail.com[3)], prammarta88@gmail.com[4)], jodiust9@gmail.com[5)]

Abstrak : *Penelitian ini bertujuan untuk mengevaluasi pengaruh teknik seleksi fitur Chi-Square dan Information Gain terhadap kinerja klasifikasi sentimen pada ulasan aplikasi Gojek menggunakan algoritma Naïve Bayes. Permasalahan utama dalam penelitian ini adalah tingginya dimensi fitur pada data teks berbahasa Indonesia yang berpotensi memengaruhi performa model. Untuk mengatasi hal tersebut, dilakukan tahapan praproses teks yang meliputi pelabelan sentimen, normalisasi, serta ekstraksi fitur menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) dengan jumlah fitur awal sebanyak 5000. Dataset terdiri dari 225.043 ulasan aplikasi Gojek, dan setelah proses pembersihan diperoleh 220.860 ulasan valid yang digunakan dalam eksperimen. Hasil penelitian menunjukkan bahwa metode Chi-Square menghasilkan akurasi terbaik sebesar 0,877864 pada 3000 fitur, sedangkan Information Gain mencapai akurasi 0,877751 pada 4000 fitur. Kedua metode tersebut menghasilkan performa yang sedikit lebih rendah dibandingkan model tanpa seleksi fitur yang mencapai akurasi 0,878045. Temuan ini menunjukkan bahwa reduksi fitur tidak selalu meningkatkan kinerja Naïve Bayes, karena model ini bekerja lebih optimal dengan distribusi fitur yang luas. Oleh karena itu, seleksi fitur perlu diterapkan secara hati-hati, khususnya pada data ulasan berbahasa Indonesia yang bersifat informal dan memiliki variasi ekspresi yang tinggi.*

Kata kunci—Chi-Square, Information Gain, klasifikasi sentimen, Naive Bayes, seleksi fitur.

Abstract **:** *This study aims to evaluate the effect of Chi-Square and Information Gain feature selection techniques on the performance of sentiment classification of Gojek application reviews using the Naïve Bayes algorithm. The main problem addressed in this study is the high dimensionality of features in Indonesian-language text data, which may affect model performance. To address this issue, a text preprocessing pipeline was applied, including sentiment labeling, normalization, and feature extraction using the Term Frequency–Inverse Document Frequency (TF-IDF) method with an initial feature size of 5,000. The dataset consists of 225,043 Gojek application reviews, and after data cleaning, 220,860 valid reviews were used for the experiments. The results show that the Chi-Square method achieved the best accuracy of 0.877864 with 3,000 features, while Information Gain reached an accuracy of 0.877751 with 4,000 features. Both methods performed slightly lower than the model without feature selection, which achieved an accuracy of 0.878045. These findings indicate that feature reduction does not always improve the performance of Naïve Bayes, as the algorithm performs more optimally with a broader feature distribution. Therefore, feature selection should be applied carefully, especially for Indonesian-language review data that are informal and highly varied in expression.*

Keywords—Chi-Square, Information Gain, sentiment classification, Naive Bayes, feature selection.

Penulis Korespondensi:
Nama Penulis, Muhamad Fahrurozi
Program Studi Penulis, Teknik Informatika
Institusi Penulis, STMIK IKMI Cirebon
Email: muhammadfahrurozi148@gmail.com
Orchid ID: http://orcid.org/0009-0007-1320-7771

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

I. PENDAHULUAN

PERTUMBUHAN pesat ekosistem aplikasi digital di Indonesia telah mendorong meningkatnya volume ulasan pengguna dari tahun ke tahun. Ulasan tersebut merepresentasikan persepsi, evaluasi, serta pengalaman faktual pengguna terhadap layanan yang mereka gunakan, sehingga menjadi sumber informasi strategis bagi penyedia layanan maupun calon pengguna dalam menilai kualitas aplikasi [1]. Namun demikian, jumlah ulasan yang dapat mencapai puluhan ribu hingga jutaan menyebabkan analisis manual tidak lagi layak dilakukan secara konsisten maupun efisien.

Dalam kondisi tersebut, analisis sentimen menjadi pendekatan komputasional yang relevan untuk mengidentifikasi dan mengelompokkan opini pengguna ke dalam kategori seperti positif, negatif, atau netral. Kebutuhan terhadap metode ini semakin meningkat seiring bertambahnya data tidak terstruktur dalam bentuk teks pada berbagai platform digital [2]. Analisis sentimen tidak hanya berfungsi untuk menangkap kecenderungan opini publik, tetapi juga berperan strategis dalam pengambilan keputusan terkait peningkatan kualitas layanan aplikasi.

Naïve Bayes merupakan salah satu algoritma yang paling banyak digunakan dalam tugas analisis sentimen karena kesederhanaan, efisiensi komputasi, dan performanya yang stabil pada berbagai studi klasifikasi teks [3]. Meskipun demikian, tantangan signifikan muncul pada data teks berdimensi tinggi, di mana sejumlah besar fitur berupa token tidak memberikan kontribusi yang berarti terhadap klasifikasi. Kondisi ini berpotensi mengurangi performa model apabila tidak dilakukan pengendalian terhadap redundansi atau noise pada fitur.

Untuk mengatasi permasalahan tersebut, teknik seleksi fitur diperlukan untuk menyaring fitur yang paling relevan dari keseluruhan korpus. Chi-Square dan Information Gain merupakan dua metode yang paling banyak digunakan dalam konteks ini. Chi-Square mengevaluasi tingkat keterkaitan antara fitur dan kelas sentimen, sedangkan Information Gain mengukur kontribusi fitur dalam mengurangi ketidakpastian model [4], [5]. Penerapan seleksi fitur terbukti penting dalam meningkatkan efisiensi dan performa model.

Penelitian sebelumnya menunjukkan bahwa penggunaan seleksi fitur dapat menghasilkan peningkatan akurasi model secara signifikan, bahkan melebihi 10% dalam beberapa studi analisis sentimen [6]. Temuan tersebut menegaskan bahwa seleksi fitur merupakan komponen kunci dalam pipeline pemrosesan teks, terutama pada dataset berukuran besar seperti ulasan aplikasi.

Walaupun berbagai studi telah mengonfirmasi efektivitas Chi-Square dan Information Gain, kajian komparatif terhadap kedua teknik tersebut pada ulasan aplikasi berbahasa Indonesia masih terbatas. Keterbatasan ini menunjukkan adanya ruang penelitian yang penting untuk dieksplorasi, khususnya dalam konteks aplikasi lokal seperti Gojek yang memiliki basis pengguna luas dan volume ulasan yang besar. Oleh karena itu, penelitian ini dilakukan untuk memberikan pemahaman empiris mengenai dampak penerapan kedua metode seleksi fitur tersebut terhadap performa model klasifikasi sentimen pada data ulasan aplikasi berbahasa Indonesia.

II. TINJAUAN PUSTAKA

A. Text Mining

Text mining merupakan pendekatan komputasional yang digunakan untuk mengekstraksi informasi dan pengetahuan dari data teks tidak terstruktur. Proses ini mencakup tahapan praproses, ekstraksi fitur, seleksi fitur, dan klasifikasi, sehingga memungkinkan transformasi teks mentah menjadi representasi numerik yang dapat dianalisis oleh algoritma pembelajaran mesin [7]. Melalui proses ini, sistem mampu mengidentifikasi pola linguistik dalam data berukuran besar yang sebelumnya sulit dianalisis secara manual.

Tahapan utama dalam text mining dimulai dengan pembersihan teks untuk menghilangkan elemen non-informatif seperti karakter khusus dan tag HTML. Proses dilanjutkan dengan tokenisasi, normalisasi (case folding, stemming, atau lemmatization), serta konversi teks menjadi representasi numerik menggunakan metode seperti bag-of-words atau TF-IDF[8], [9]. Kualitas representasi fitur ini dapat ditingkatkan melalui seleksi fitur seperti Chi-Square dan Information Gain yang terbukti meningkatkan efektivitas model klasifikasi, termasuk Naïve Bayes [10], [11]. Teknik text mining banyak diterapkan dalam berbagai tugas, mulai dari pengelompokan topik hingga analisis sentimen.

Text mining sangat relevan untuk analisis ulasan aplikasi dan media sosial karena kemampuannya

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

mengolah data tidak terstruktur dalam skala besar secara efisien. Dengan meningkatnya volume komentar pengguna, teknik ini menjadi alat strategis dalam menghasilkan insight yang mendukung pengambilan keputusan berbasis data [12], [13].

B. Analisis Sentimen

Analisis sentimen merupakan cabang text mining yang berfokus pada identifikasi kecenderungan opini positif, negatif, atau netral berdasarkan informasi yang terkandung dalam teks [14]. Teknik ini digunakan untuk memahami persepsi dan respons pengguna terhadap suatu entitas sehingga menjadi instrumen penting dalam mengevaluasi kualitas layanan digital [12]. Perkembangannya semakin pesat seiring meningkatnya ketersediaan data ulasan pada platform digital.

Analisis sentimen dapat dilakukan pada beberapa tingkat: tingkat dokumen, yang menilai polaritas keseluruhan teks; tingkat kalimat, yang menganalisis polaritas tiap kalimat; dan tingkat aspek, yang mengevaluasi sentimen terhadap komponen tertentu dari suatu entitas [15]. Penelitian ini menggunakan pendekatan tingkat dokumen, di mana setiap ulasan dianggap sebagai satu unit analisis.

Dalam konteks ulasan aplikasi digital, penelitian-penelitian terdahulu menunjukkan bahwa analisis sentimen efektif membantu penyedia layanan memahami persepsi publik, mengidentifikasi isu utama, dan memantau kualitas layanan secara berkelanjutan [16], [17]. Namun, sedikit penelitian yang mengkaji penggunaan seleksi fitur seperti Chi-Square dan Information Gain pada ulasan aplikasi berbahasa Indonesia, sehingga aspek ini masih menjadi ruang penelitian yang terbuka.

C. TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) merupakan salah satu metode representasi fitur yang paling banyak digunakan dalam klasifikasi teks. Metode ini memberikan bobot lebih besar pada kata yang penting dalam dokumen tertentu namun jarang muncul pada korpus secara keseluruhan, sehingga menghasilkan representasi yang diskriminatif [18], [19]. Dalam analisis sentimen, TF-IDF efektif membantu menonjolkan kata kunci yang relevan untuk membedakan polaritas dokumen.

Representasi n-gram (khususnya unigram dan bigram) sering dipadukan dengan TF-IDF untuk menangkap konteks lokal antar kata [20]. Secara matematis, TF-IDF dihitung berdasarkan Rumus 2.1–2.4, yang mencakup frekuensi kemunculan kata dalam dokumen dan persebaran kata dalam korpus. Kombinasi TF-IDF dan n-gram dilaporkan mampu meningkatkan akurasi klasifikasi secara signifikan pada berbagai domain teks [20], [21].

D. N-gram

N-gram adalah representasi sekuensial dari n kata yang muncul secara berurutan dalam teks. Pendekatan ini memungkinkan model menangkap hubungan lokal antar kata yang tidak dapat diperoleh melalui bag-of-words [22], [23]. Dalam analisis ulasan aplikasi, n-gram sangat bermanfaat untuk menangkap frasa penting seperti negasi (“tidak puas”) atau intensifier (“sangat lambat”).

Unigram, bigram, dan trigram merupakan bentuk n-gram yang paling umum. Namun penggunaan n-gram berukuran besar dapat menyebabkan sparsity, sehingga penelitian ini memilih kombinasi unigram–bigram (n = 1,2) sebagai kompromi yang optimal.

E. Seleksi Fitur

Seleksi fitur merupakan tahap penting dalam pemrosesan teks yang bertujuan mereduksi jumlah fitur tanpa kehilangan informasi penting, sehingga model dapat bekerja lebih efisien dan akurat [24], [25]. Dengan volume fitur yang dapat mencapai puluhan ribu, seleksi fitur membantu mengurangi noise dan meningkatkan performa model seperti Naïve Bayes.

Penelitian ini membandingkan dua metode seleksi fitur yang populer, yaitu Chi-Square dan Information Gain, untuk menilai efektivitas masing-masing metode dalam klasifikasi sentimen ulasan aplikasi Gojek.

F. Chi-Square (χ^2)

Chi-Square mengevaluasi tingkat dependensi antara kata dan kelas sentimen berdasarkan tabel kontingensi. Fitur dengan nilai χ^2 tinggi dianggap relevan, sedangkan fitur dengan nilai rendah dapat dieliminasi [26].

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

G. *Information Gain (IG)*

Information Gain mengukur seberapa besar suatu fitur mengurangi ketidakpastian (entropy) dalam proses klasifikasi [27]. Metode ini banyak digunakan dalam analisis sentimen karena kontribusinya terhadap peningkatan akurasi dan f1-score [21].

H. *Klasifikasi Teks*

Klasifikasi teks mengacu pada proses mengelompokkan dokumen ke dalam kategori tertentu berdasarkan pola linguistik yang dipelajari dari data berlabel [28]. Tantangan utama dalam klasifikasi teks meliputi tingginya dimensi fitur, ambiguitas bahasa, serta variasi gaya penulisan, sehingga diperlukan algoritma yang efisien dan robust seperti Naïve Bayes [29].

I. *Naïve Bayes*

Naïve Bayes, khususnya MultinomialNB, adalah algoritma probabilistik yang banyak digunakan pada klasifikasi teks karena kesederhanaan dan efisiensinya [3]. Meskipun mengasumsikan independensi antar fitur, algoritma ini tetap menghasilkan performa yang kompetitif [30]. Formula dasar Naïve Bayes digunakan untuk menghitung probabilitas posterior, dengan kelas dipilih berdasarkan probabilitas tertinggi. Penelitian terdahulu menunjukkan bahwa Naïve Bayes dapat mencapai akurasi tinggi terutama ketika dipadukan dengan seleksi fitur [31], [32].

J. *Studi Terdahulu dan Research Gap*

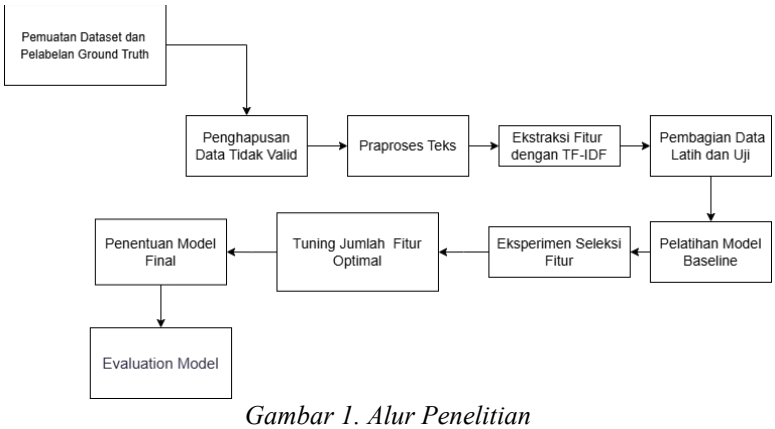
Literatur menunjukkan bahwa Chi-Square dapat meningkatkan akurasi model secara signifikan [4], Information Gain meningkatkan stabilitas model [21], dan penggunaan TF-IDF dengan n-gram menghasilkan performa optimal [20]. Namun penelitian terkait ulasan aplikasi Indonesia masih terbatas, terutama dalam hal:

1. Minimnya studi komparatif antara Chi-Square dan Information Gain.
2. Kurangnya penelitian yang menggunakan dataset besar seperti ulasan Gojek.
3. Belum adanya bukti empiris mengenai pengaruh kedua teknik seleksi fitur terhadap Naïve Bayes dalam konteks bahasa Indonesia.

Penelitian ini hadir untuk mengisi kesenjangan tersebut melalui analisis komprehensif terhadap dua teknik seleksi fitur pada dataset ulasan aplikasi Gojek berbahasa Indonesia.

III. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif eksperimental untuk mengevaluasi pengaruh teknik seleksi fitur Chi-Square dan Information Gain terhadap akurasi klasifikasi sentimen berbasis Naïve Bayes. Objek penelitian berupa dataset ulasan aplikasi Gojek yang diambil dari platform Kaggle melalui sumber resmi *Gojek App Reviews Bahasa Indonesia*, yang menyediakan data ulasan pengguna dengan skor penilaian sebagai dasar pelabelan sentimen. Seluruh eksperimen dilakukan secara terkontrol dan komparatif, di mana kedua teknik seleksi fitur diuji pada dataset, parameter model, dan konfigurasi praproses yang sama untuk memastikan konsistensi hasil. Dengan menerapkan prinsip kontrol variabel dan replikasi, penelitian ini mampu mengisolasi kontribusi masing-masing teknik terhadap performa model, sehingga temuan yang diperoleh memiliki validitas internal yang kuat serta dapat direplikasi pada penelitian selanjutnya. Adapun alur penelitian yang dilakukan pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

Tabel 1. Penjelasan Alur Penelitian

Tahapan	Aktivitas	Deskripsi Aktivitas
Dataset Loading & Ground Truth Labeling	Memuat dataset dan memberikan label sentimen	Dataset GojekAppReview1.csv dimuat dan dilakukan pelabelan otomatis berdasarkan skor ulasan (1–2 = negatif, 3 = netral, 4–5 = positif). Proses ini memastikan setiap ulasan memiliki label sentimen yang konsisten sebagai dasar pemodelan.
Data Cleaning	Menghapus data tidak valid	Baris yang tidak memiliki teks, nilai skor kosong, atau konten tidak relevan dihapus. Tahap ini dilakukan untuk memastikan kualitas data dan mencegah gangguan (noise) pada proses praproses dan pemodelan.
Text Preprocessing	Membersihkan dan menormalkan teks	Proses mencakup cleaning, case folding, tokenisasi, stopword removal, dan stemming. Praproses dilakukan untuk menyeragamkan struktur teks serta mengurangi variasi kata yang tidak diperlukan sehingga fitur menjadi lebih representatif.
TF-IDF Feature Extraction	Mengubah teks menjadi fitur numerik	TF-IDF dengan n-gram (1,2) dan max_features = 5000 digunakan untuk membentuk matriks fitur. Metode ini memberikan bobot kata berdasarkan tingkat kepentingannya dalam korpus, sehingga fitur lebih informatif untuk klasifikasi sentimen.
Train–Test Split	Membagi data menjadi data latih dan data uji	Data dibagi menggunakan stratified split (80:20) untuk menjaga distribusi label yang seimbang pada kedua subset. Tahap ini memastikan evaluasi performa model lebih akurat dan tidak bias.
Baseline Model Training	Melatih model awal tanpa seleksi fitur	Model Multinomial Naïve Bayes dilatih menggunakan fitur TF-IDF 5000 dimensi. Model baseline ini digunakan sebagai titik pembanding terhadap model dengan seleksi fitur.
Feature Selection Experimentation	Menerapkan Chi-Square dan Information Gain	Fitur dievaluasi menggunakan Chi-Square dan Information Gain secara terpisah. Proses ini menghasilkan ranking fitur berdasarkan tingkat relevansinya terhadap kelas sentimen.
Feature Tuning (K Optimization)	Menguji berbagai jumlah fitur (K)	Percobaan dilakukan pada K = {100, 500, 1000, 2000, 3000, 4000, 5000}. Tahap ini menentukan jumlah fitur optimal yang memberikan akurasi tertinggi pada model.
Final Model Selection	Memilih model terbaik	Model dengan performa terbaik dipilih berdasarkan perbandingan akurasi, precision, recall, dan F1-score dari setiap skenario (baseline, Chi-Square, Information Gain).
Model Evaluation	Mengukur kinerja model	Evaluasi dilakukan pada tiga konfigurasi model menggunakan metrik klasifikasi standar. Hasil evaluasi dibahas untuk mengetahui pengaruh seleksi fitur terhadap performa model.

Penelitian ini dilakukan dengan Kebutuhan Hardware dan Software spesifikasi sebagai berikut.

1) *Perangkat Keras (Hardware)*
   a) Prosesor: Intel Core i5 atau setara
   b) RAM: 8 GB
   c) Penyimpanan: 512 GB SSD
   d) Sistem Operasi: Windows 10 64-bit
2) *Perangkat Lunak (Software)*
   a) Bahasa Pemrograman: Python 3.x
   b) Library: Scikit-learn, Pandas, NumPy, Matplotlib
   c) Tools: Jupyter Notebook
   d) Dataset: Ulasan aplikasi Gojek dari Kaggle

Spesifikasi perangkat keras dan perangkat lunak digunakan untuk memastikan replikasi eksperimen dapat dilakukan secara konsisten.

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

IV. HASIL DAN PEMBAHASAN

1. *Hasil Pemuatan Dataset*

Untuk memberikan gambaran awal mengenai karakteristik dataset yang digunakan dalam penelitian, dilakukan pemuatan data ulasan aplikasi Gojek dari platform Kaggle. Ringkasan informasi mengenai jumlah data dan distribusi label sentimen disajikan pada Tabel 2 berikut.

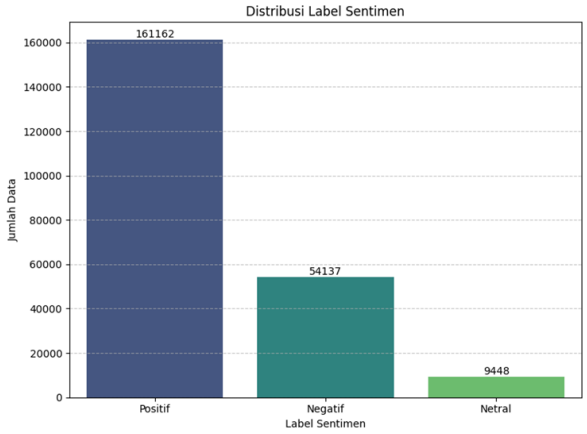
Tabel 2. Hasil Pemuatan Dataset

Parameter	Jumlah
Total Dataset	225.043
Positif	161.162
Negatif	54.137
Netral	9.448
Name	count
Datatype	int64

Dataset penelitian terdiri dari 225.043 ulasan aplikasi Gojek yang diperoleh melalui platform Kaggle. Ulasan dipetakan ke dalam tiga kategori sentimen berdasarkan skor rating, yaitu positif, negatif, dan netral. Tabel 2 menunjukkan bahwa distribusi kelas tidak seimbang, dengan sentimen positif mendominasi (161.162 atau 71,6%), diikuti negatif (54.137 atau 24,0%), dan netral (9.448 atau 4,2%). Ketidakseimbangan tersebut menandakan potensi bias model terhadap kelas mayoritas dan perlu diperhatikan dalam proses evaluasi. Struktur dataset yang bersih dan tidak mengandung nilai hilang mendukung konsistensi tahap praproses dan pemodelan.

2. *Hasil Pelabelan Ground Truth*

Setelah proses pemetaan skor rating menjadi tiga kategori sentimen, dilakukan visualisasi untuk mengamati distribusi kelas yang terbentuk dari pelabelan awal. Gambar 2 berikut menampilkan hasil pelabelan ground truth berdasarkan rating ulasan pengguna.



Gambar 2. Hasil Pelabelan Ground Truth

Visualisasi pada Gambar 2 mengonfirmasi ketimpangan distribusi label, dengan kelas positif sebagai mayoritas absolut. Dominasi ulasan positif mengindikasikan tingkat kepuasan pengguna yang tinggi, sedangkan proporsi ulasan negatif menunjukkan keberadaan masalah layanan tertentu yang perlu diperhatikan. Kelas netral yang sangat kecil (4,18%) menjadi tantangan penting dalam pemodelan karena jumlah data yang terbatas dapat menurunkan kemampuan model dalam mengenali pola linguistik kelas tersebut. Ketidakseimbangan ini meningkatkan baseline accuracy secara artifisial, sehingga interpretasi hasil model harus mempertimbangkan metrik precision, recall, dan F1-score per kelas.

Untuk mengukur tingkat ketidakseimbangan, digunakan *Imbalance Ratio (IR)*. Nilai IR yang tinggi pada kelas netral menunjukkan bahwa kelas ini merupakan minoritas ekstrem, sehingga berpotensi diabaikan oleh model. Pendekatan evaluasi yang lebih komprehensif menjadi penting untuk memastikan bahwa performa model tidak hanya ditentukan oleh kemampuan memprediksi kelas mayoritas.

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

3. Hasil Penghapusan Data Tidak Valid

Untuk meningkatkan kualitas data, dilakukan proses identifikasi dan penghapusan baris yang tidak relevan atau tidak memenuhi standar kelayakan analisis. Contoh baris yang dihapus serta alasan penghapusannya disajikan pada Tabel 3 dan Tabel 4, sedangkan Tabel 5 menunjukkan data yang menjadi kosong setelah proses praproses.

Tabel 3. Contoh Data Tidak Valid

No	userName	Content	score	at	appVersion
738	Nurlaili	S.Pd.I	Kenapa setelah update saya tidak bisa mengkone...	1	27/01/2024 06:28
1241	6_ Abdul Aziz	SHI	Saya hanya pakai gojek	5	31/01/2024 12:14
1426	Elas Junika	A. Md	Kenapa belakangan ini sering terjadi orderan g...	3	30/01/2024 10:44
2590	YUDI ASRA	ST	Sampai saai ini masih memuaskan pertahankandan...	5	12/01/2024 10:10
3736	SERVIS AC TERDEKAT AC	MESIN CUCI (PEMASANGAN AC BARU	SEKEN)	Parah ini hasil mapsnyabisa bisanya titik yg d...	1

Tabel 4. Distribusi Alasan Penghapusan

No	Content	score	Sentiment label mapped
738	S.Pd.I	Kenapa setelah update saya tidak bisa mengkone...	None
1241	SHI	Saya hanya pakai gojek	None
1426	A. Md	Kenapa belakangan ini sering terjadi orderan g...	None
2590	ST	Sampai saai ini masih memuaskan pertahankandan...	None
3736	MESIN CUCI (PEMASANGAN AC BARU	SEKEN)	None

Tabel 5. Jumlah baris yang kontennya menjadi kosong setelah praproses.

No	content	clean_stemmed_content
41	Baik	
48	Baik	
58	Baik	
59	Sangat baik	
152	Baik	

Proses pembersihan data menghilangkan baris yang tidak memiliki skor, tidak memiliki teks, atau menghasilkan teks kosong setelah praproses. Dari total data awal, 4.176 baris dihapus, menyisakan 220.860 ulasan yang layak digunakan. Tabel 3 dan Tabel 4 menunjukkan contoh baris yang dihapus, termasuk ulasan tanpa skor dan teks yang tidak memiliki makna semantik. Beberapa teks yang sangat pendek seperti “Baik” juga terhapus akibat stopword removal, sebagaimana terlihat pada Tabel 5, karena tidak menghasilkan token setelah seluruh tahap praproses diterapkan. Eliminasi data tidak valid ini meningkatkan keandalan representasi fitur dan mengurangi risiko noise pada model.

4. Hasil Praproses Teks

Tahap praproses teks dilakukan untuk menghasilkan representasi teks yang bersih, terstruktur, dan siap digunakan dalam proses ekstraksi fitur. Contoh transformasi teks sebelum dan sesudah praproses ditampilkan pada Tabel 6 berikut.

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

Tabel 6. Hasil Praproses Teks

No	Percobaan	Original Text	Processed Text
1	1	Bagus	bagus
2	2	Ok bgt cepat	ok bgt cepat
3	3	Edit bintang 1 ke bintang 5 Seneng deh akhirnya sekarang TopUp GoPay udah bisa di Indomart Minimal jumlah Top Up nya jg kecil 10 rbu aja dan biaya admin nya 2 rbu Bagus pertahankan ya Promo Voucher Discount GoFood nya TOLONG DIPERBANYAK Thanks	edit bintang bintang neng deh topup gopay udah indomart minimal top up nya jg rbu aja biaya admin nya rbu bagus tahan ya promo voucher discount gofood nya tolong banyak thanks
4	4	Ok	ok
5	5	Gojek emang top	gojek emang top

Tahap praproses berhasil menormalkan teks menjadi format yang konsisten dan siap diekstraksi ke dalam fitur numerik. Contoh pada Tabel 6 menunjukkan bahwa ulasan panjang mengalami reduksi signifikan melalui proses cleaning, tokenisasi, stopword removal, dan stemming. Sebaliknya, ulasan dengan struktur sederhana tetap dipertahankan tanpa kehilangan konteks. Transformasi ini memastikan bahwa noise linguistik seperti variasi ejaan, huruf kapital, dan simbol tidak memengaruhi kualitas fitur yang dihasilkan.

5. Hasil Ekstraksi Fitur dengan TF-IDF

Setelah teks melewati proses praproses, dilakukan ekstraksi fitur menggunakan metode TF-IDF untuk mengidentifikasi kata-kata yang memiliki peran penting dalam membedakan kategori sentimen. Tabel 7 berikut menyajikan daftar fitur dengan bobot tertinggi.

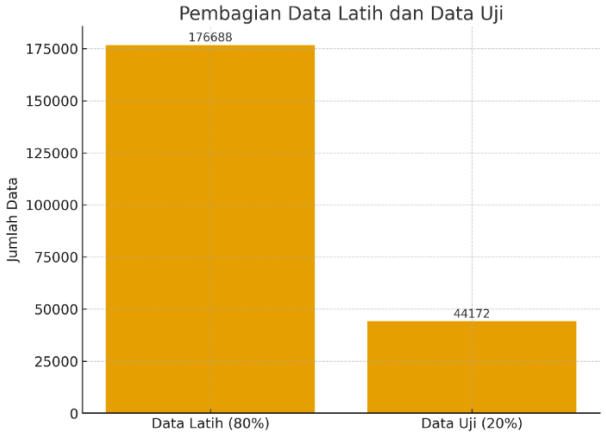
Tabel 7. Hasil Ekstrasin Fitur TF-IDF

Feature	Importance
Mantap	12095.4
ok	11542.4
Bantu	11214.8
Bagus	10728.5
Good	8430.49

Ekstraksi fitur menggunakan TF-IDF dengan *max_features* = **5000** dan *ngram_range* = (1, 2). Tabel 7 memperlihatkan fitur dengan bobot tertinggi, seperti “mantap”, “ok”, dan “bantu”, yang menunjukkan peran kata-kata tersebut sebagai indikator kuat dalam membedakan sentimen. Keberadaan kata berbahasa Inggris seperti “good” menunjukkan variasi bahasa dalam ulasan pengguna. Distribusi bobot TF-IDF yang jelas membantu model memahami kata-kata yang paling representatif di setiap kategori sentimen.

6. Hasil Pembagian Data Latih dan Uji

Untuk menguji performa model, dataset dibagi menjadi data latih dan data uji menggunakan stratified split. Proporsi pembagian data dapat dilihat pada Gambar 3 berikut.



Gambar 3. Data Train dan Test

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

Melalui *stratified split* dengan rasio 80:20, sistem menghasilkan 176.688 data latih dan 44.172 data uji disajikan pada Gambar 3. Pendekatan stratifikasi mempertahankan proporsi label pada kedua subset, sehingga evaluasi model dapat mencerminkan kemampuan generalisasi secara lebih stabil.

7. Hasil Pelatihan Model Baseline

Sebagai langkah evaluasi awal, dilakukan pelatihan model baseline menggunakan algoritma Multinomial Naïve Bayes. Hasil akurasi dan metrik evaluasi per kelas ditampilkan pada Tabel 8 dan Tabel 9 berikut.

Tabel 8 Hasil Pelatihan Model Baseline

Metode	Akurasi
Baseline (No FS)	0.878045

Tabel 9 Matriks Akurasi Model Baseline

Kelas	Precision	Recall	F1-Score	Support
Negatif	0.73	0.89	0.80	10.800
Netral	0.33	0.00	0.01	1.866
Positif	0.94	0.92	0.93	31.506
Accuracy			0.88	44.172
Macro Avg	0.67	0.61	0.58	44.172
Weighted Avg	0.86	0.88	0.86	44.172

Model baseline menggunakan Multinomial Naïve Bayes dengan 5000 fitur TF-IDF. Seperti terlihat pada Tabel 8, baseline mencapai akurasi 0,8780, nilai tertinggi dari seluruh eksperimen. Model menunjukkan performa kuat pada kelas positif dan negatif, namun gagal mengidentifikasi kelas netral (recall = 0,00) terlihat pada Tabel 9, mencerminkan dampak class imbalance terhadap model probabilistik.

8. Hasil Eksperimen Seleksi Fitur

Untuk menilai dampak seleksi fitur terhadap performa model, dilakukan eksperimen menggunakan Chi-Square dan Information Gain. Hasil perbandingan nilai akurasi masing-masing metode disajikan pada Tabel 10 berikut.

Tabel 10. Hasil Eksperimen Seleksi Fitur

Metode	Akurasi
Baseline (No FS)	0.878045
Chi-Square (K=1000)	0.873517
Information Gain (K=1000)	0.870959

Eksperimen awal pada K=1000 menunjukkan bahwa Chi-Square menghasilkan akurasi 0,8735, sementara Information Gain menghasilkan 0,8710 terlihat pada Tabel 10. Kedua nilai sedikit lebih rendah dibandingkan baseline, tetapi menunjukkan bahwa sebagian besar fitur tidak relevan telah berhasil dipangkas. Temuan ini mengindikasikan bahwa pada corpus besar seperti ulasan Gojek, pengurangan fitur secara agresif tidak selalu menghasilkan peningkatan performa.

9. Hasil Tuning Jumlah Fitur Optimal

Eksperimen lanjutan dilakukan untuk menentukan jumlah fitur optimal yang memberikan kinerja terbaik. Ringkasan performa model pada berbagai nilai K ditunjukkan pada Tabel 11.

Tabel 11. Hasil Tuning Jumlah Fitur Optimal

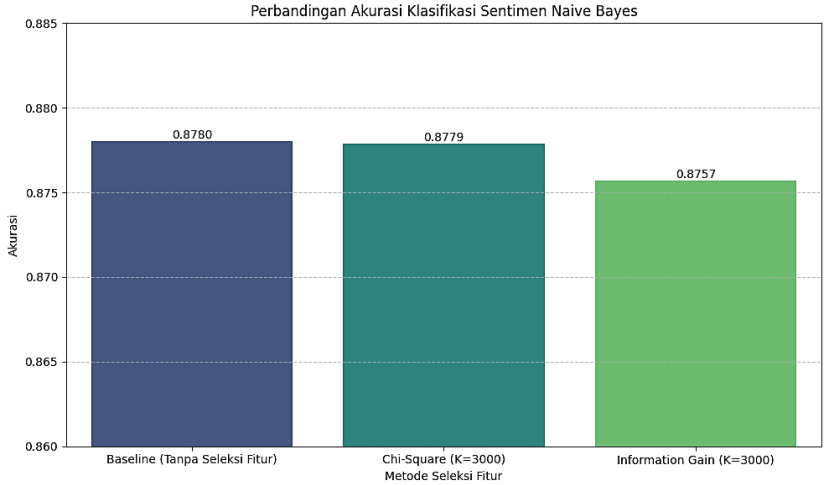
K	Chi-Square	Information Gain
100	0.832858	0.811442
500	0.869487	0.863579
1000	0.873517	0.870959
2000	0.877253	0.87431
3000	0.877864	0.87569
4000	0.877411	0.877751
Baseline (5000)	0.878045	0.878045

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

Pengujian K pada nilai 100 hingga 4000 menunjukkan pola peningkatan kinerja seiring bertambahnya jumlah fitur Terlihat pada Tabel 11. Chi-Square mencapai akurasi tertinggi pada K=3000 (0,877864), sementara Information Gain stabil pada K=4000 (0,877751). Walaupun baseline tetap yang tertinggi, model dengan K=3000 menunjukkan efisiensi komputasi yang meningkat dengan penurunan dimensi sebesar 40%.

10. Hasil Penentuan Model Final

Berdasarkan evaluasi akurasi dan efisiensi komputasi, konfigurasi model terbaik dipilih sebagai model final. Visualisasi perbandingan performa model dapat dilihat pada Gambar 4 berikut.



Gambar 4. Penentuan Model Final

Model final dipilih berdasarkan keseimbangan antara akurasi dan efisiensi. Gambar 4 menunjukkan bahwa Naïve Bayes + Chi-Square K=3000 (0,877864) menjadi konfigurasi terbaik karena memberikan performa mendekati baseline dengan jumlah fitur yang jauh lebih sedikit.

11. Hasil Evaluasi Model

Model terbaik kemudian dievaluasi menggunakan classification report untuk menganalisis performa per kelas secara lebih detail. Hasil evaluasi tersebut disajikan pada Tabel 12 berikut.

Tabel 12. Classification Report Terbaik(K=3000)

Kelas	Precision	Recall	F1-Score	Support
Negatif	0.73	0.89	0.80	10.800
Netral	0.33	0.00	0.01	1.866
Positif	0.94	0.92	0.93	31.506
Accuracy			0.88	44.172
Macro Avg	0.67	0.61	0.58	44.172
Weighted Avg	0.86	0.88	0.86	44.172

Evaluasi model final Pada Tabel 12 menunjukkan performa solid pada kelas positif dan negatif, namun kelas netral tidak terdeteksi dengan baik. Akurasi keseluruhan tetap tinggi pada 0,88, namun performa per kelas mengungkapkan perlunya strategi yang lebih efektif untuk menangani ketidakseimbangan kelas.

12. Pembahasan

Hasil penelitian menunjukkan bahwa reduksi fitur menggunakan Chi-Square dan Information Gain tidak memberikan peningkatan performa signifikan dibandingkan baseline. Naïve Bayes justru bekerja lebih optimal tanpa reduksi fitur, sejalan dengan literatur yang menyatakan bahwa model probabilistik membutuhkan cakupan distribusi kata yang lebih luas. Hal ini menjelaskan mengapa baseline dengan 5000 fitur mencapai akurasi tertinggi (0,8780).

Chi-Square menunjukkan hasil sedikit lebih baik dibandingkan Information Gain, khususnya pada jumlah fitur di atas 2000, namun selisih antar metode relatif kecil. Temuan ini mengonfirmasi bahwa efektivitas seleksi fitur sangat berkaitan dengan karakteristik dataset yang besar dan variatif. Penelitian

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

ini memperkuat pemahaman bahwa pada data ulasan berbahasa Indonesia yang kaya konteks dan informal, pemangkasan fitur terlalu agresif dapat menghilangkan informasi penting yang justru berkontribusi pada klasifikasi sentimen.

V. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, model Naïve Bayes tanpa seleksi fitur menghasilkan akurasi tertinggi sebesar 0,878045. Penerapan Chi-Square dan Information Gain tidak mampu meningkatkan akurasi, namun Chi-Square mampu mengurangi jumlah fitur hingga 40% dengan performa yang tetap mendekati baseline.

Hal ini menunjukkan bahwa Naïve Bayes bekerja lebih optimal dengan distribusi fitur yang luas, sehingga reduksi fitur perlu dilakukan secara selektif. Ketidakseimbangan data juga menjadi faktor utama rendahnya performa pada kelas netral.

Penelitian selanjutnya disarankan untuk menerapkan teknik penanganan ketidakseimbangan data serta menggunakan algoritma lain untuk meningkatkan performa pada kelas minoritas.

DAFTAR RUJUKAN

[1] H. Duong, "User Perception Analysis in Digital Application Ecosystems," *J. Digit. Interact. Stud.*, vol. 12, no. 1, pp. 45–58, 2024, doi: 10.1016/j.jdis.2024.01.004.

[2] L. Zhang, "Advances in Sentiment Analysis for Large-Scale Text Data," *Int. J. Comput. Linguist.*, vol. 9, no. 2, pp. 77–93, 2020, doi: 10.1186/s40104-020-00452-7.

[3] A. McCallum, "A Comparison of Event Models for Naive Bayes Text Classification," in *AAAI-98 Workshop on Learning for Text Categorization*, 1998, pp. 41–48.

[4] Y. Liu, M. Chen, and J. Xu, "Improving text classification using feature selection with Chi-Square and Information Gain," *Appl. Intell.*, vol. 53, pp. 14212–14228, 2023, doi: 10.1007/s10489-023-04576-0.

[5] W. Tan, "Evaluating Information Gain for Modern Text Mining Pipelines," *Data Sci. Rev.*, vol. 14, no. 3, pp. 201–215, 2023, doi: 10.1145/3591123.

[6] Q. Wen, "Improving Sentiment Classification Accuracy with Feature Filtering," *J. Intell. Syst.*, vol. 16, no. 2, pp. 89–102, 2024, doi: 10.1515/jisys-2023-0047.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[8] R. Churchill and L. Singh, "textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data ," 2021. doi: 10.5220/0010559000600070 .

[9] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers ," 2024. doi: 10.1016/j.is.2023.102342 .

[10] I. M. Mubaroq and E. B. Setiawan, "The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine ," 2020. doi: 10.34818/INDOJC.2020.5.2.499 .

[11] D. C. Sami'un, A. Sugiharto, and F. Jie, "Chi-Square feature selection for improving sentiment analysis of news data privacy treats ," 2024. doi: (tidak ada DOI pada versi PDF, tetapi artikel tersedia di JATIT) .

[12] M. T. F. Maulana, B. I. Nugroho, and E. U. S. Utami, "Analisis Sentimen Ulasan Aplikasi TikTok di Google Playstore Menggunakan Algoritma Naïve Bayes," *RIGGS J. Artif. Intell. Digit. Bus.*, vol. 4, no. 3, 2025, doi: 10.31004/riggs.v4i3.2962.

[13] S. Ulyaa, A. Ridwan, and W. C. Wahyudi, "Text Mining Sentimen Analisis Pengguna Aplikasi Marketplace Tokopedia Berdasarkan Rating dan Komentar pada Google Play Store," *J. Bisnis Digit. dan Sist. Inf.*, pp. 33–40, 2022, [Online]. Available: https://ejr.umku.ac.id/index.php/BIDISFO/article/download/1799/1071

[14] R. Zulfiqri, B. N. Sari, and T. N. Padilah, "Analisis Sentimen Ulasan Pengguna Aplikasi Media Sosial Instagram pada Google Play Store Menggunakan Naïve Bayes Classifier," *J. Inform. dan Tek. Elektro Terap.*, vol. 12, no. 3, 2024, doi: 10.23960/jitet.v12i3.4995.

[15] W. Zhang, X. Li, Y. Deng, L. Bing, and W.-K. Lam, "A Survey on Aspect-Based Sentiment

DOI: https://doi.org/10.35457/antivirus.v20i1.5370

Analysis: Tasks, Methods, and Challenges ," 2022. doi: 10.1109/TKDE.2022.3230975 .
[16] R. Helma, "Sentiment Analysis of Indonesian Mobile Application Reviews," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 101–118, 2025, doi: 10.51595/ijds.2025.4.2.101.
[17] M. Zakaria, "Sentiment Analysis of User Feedback in Digital Applications," *J. Inf. Syst. Technol.*, vol. 12, no. 1, pp. 55–68, 2024, doi: 10.5829/jist.2024.12.1.55.
[18] S. Khoerunnisa, D. F. Shiddieq, and D. Nurhayati, "Penerapan Algoritma Naive Bayes dengan Teknik TF-IDF dan Cross Validation untuk Analisis Sentimen Terhadap Starlink," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 2, 2025, doi: 10.57152/malcom.v5i2.1852.
[19] A. D. M. Putri, N. Sulistianingsih, and R. Rismayati, "Pengaruh Teknik Representasi Teks Bag-of-Words dan TF-IDF terhadap Akurasi Klasifikasi Sentimen Teks Multi-Domain," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 7, no. 4, 2025, doi: 10.35746/jtim.v7i4.756.
[20] W. Tan, "Evaluating N-gram and TF-IDF Combinations for Text Classification," *Data Sci. Rev.*, vol. 14, no. 3, pp. 201–215, 2023, doi: 10.1145/3591123.
[21] Q. Wen, "Improving Feature Selection for Sentiment Classification Using Information Gain," *J. Intell. Syst.*, vol. 16, no. 2, pp. 89–102, 2024, doi: 10.1515/jisys-2023-0047.
[22] S. A. Anjani and A. Fauzan, "Implementasi n-Gram dalam Analisis Sentimen Masyarakat DIY terhadap PSBB Jawa-Bali," *Statistika*, vol. 21, no. 2, 2021, doi: 10.29313/statistika.v21i2.294.
[23] A. Solikhatun and E. Sugiharti, "Application of the Naïve Bayes Classifier Algorithm using N-Gram and Information Gain to Improve the Accuracy of Restaurant Review Sentiment Analysis," *J. Adv. Inf. Syst. Technol.*, vol. 2, no. 2, pp. 11–20, 2020, doi: 10.15294/jaist.v2i2.44303.
[24] N. Fadhilah, Y. Prasetyo, and A. Wibowo, "Peningkatan Akurasi Klasifikasi Teks melalui Seleksi Fitur Information Gain dan Chi-Square," *J. Sist. Inf.*, vol. 19, no. 2, pp. 101–110, 2023, doi: 10.24089/jsi.v19i2.4567.
[25] A. Ramadhan and R. A. Asmara, "Penerapan Metode Seleksi Fitur Chi-Square pada Klasifikasi Sentimen Menggunakan Naïve Bayes," *J. Teknol. Inf. dan Komput.*, vol. 10, no. 1, pp. 55–63, 2022, doi: 10.33365/jtik.v10i1.2431.
[26] Y. Liu, "Chi-Square Based Feature Selection for Text Classification," *J. Mach. Learn. Appl.*, vol. 5, no. 4, pp. 120–137, 2023, doi: 10.32604/jmla.2023.024111.
[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann, 1993.
[28] H. Allam, M. Yaghoobi, and H. X. Nguyen, "Text Classification: How Machine Learning Is Used to Categorize Text Data ," 2025. doi: 10.3390/info16020130 .
[29] K. Taha, "A Comprehensive Survey of Modern Text Classification Techniques ," 2024.
[30] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/
[31] D. Nurhayati, "Sentiment Analysis of Indonesian App Reviews using Naive Bayes and SVM," *J. Inform. Indones.*, vol. 5, no. 2, pp. 75–86, 2019, doi: 10.25008/jii.2019.5.2.75.
[32] R. Singh, "Enhanced Naive Bayes with Feature Selection for Sentiment Analysis," *Int. J. Data Sci.*, vol. 7, no. 1, pp. 44–58, 2020, doi: 10.5120/ijds.2020.71.44.

DOI: https://doi.org/10.35457/antivirus.v20i1.5370