P-ISSN : 2580-4316
https://doi.org/10.37817/ikraith-informatika.v10i2 E-ISSN : 2654-8054

Optimasi Multinomial Naive Bayes Untuk Klasifikasi Rating Tokopedia

[1]Muhammad Erlangga Gunawan, [2]Ali Alamsyah Kusumadinata, [3]Hillmy Aliy Andra Putra
[1,3]Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Djuanda, Bogor
[2]Program Studi Ilmu Komunikasi, Fakultas Ilmu Sosial dan Ilmu Politik, Universitas Djuanda, Bogor

E-mail: [1]i.2210161@unida.ac.id, [2]alialamsyahkusumadinata@gmail.com, [3]hilmy.aliy@unida.ac.id

ABSTRAK

Analisis sentimen pada ulasan *e-commerce* menghadapi tantangan berupa penggunaan bahasa informal (*slang*) dan ketidakseimbangan kelas yang dapat menurunkan kinerja model klasifikasi. Penelitian ini bertujuan mengoptimalkan algoritma Multinomial Naive Bayes untuk mengklasifikasikan ulasan Tokopedia ke dalam lima kelas rating ordinal. Pendekatan yang diterapkan mengintegrasikan normalisasi *slang* berbasis leksikon hibrida, representasi fitur TF-IDF n-gram, penyeimbangan data menggunakan Synthetic Minority Over-sampling Technique (SMOTE), dan optimasi *hyperparameter*. Evaluasi dilakukan menggunakan Stratified 10-Fold Cross Validation pada 37.146 ulasan valid. Hasil penelitian menunjukkan bahwa pendekatan yang diusulkan meningkatkan nilai rata-rata Macro F1-Score sebesar 102,42%, dari 0,1704 menjadi 0,3450, dengan akurasi sebesar 58%. Peningkatan tersebut menunjukkan bahwa kombinasi normalisasi *slang*, TF-IDF n-gram, SMOTE, dan optimasi *hyperparameter* mampu meningkatkan kemampuan model dalam mengenali kelas minoritas pada data ulasan yang tidak seimbang.

Kata kunci : ***Analisis Sentimen, Multinomial Naive Bayes, SMOTE, TF-IDF n-gram, Tokopedia***

ABSTRACT

Sentiment analysis of *e-commerce* reviews faces challenges due to the use of informal language (*slang*) and *class imbalance*, which can reduce the performance of classification models. This study aims to optimize the Multinomial Naive Bayes algorithm for classifying Tokopedia reviews into five ordinal rating classes. The proposed approach integrates lexicon-based hybrid *slang* normalization, TF-IDF n-gram feature representation, data balancing using the Synthetic Minority Over-sampling Technique (SMOTE), and *hyperparameter* optimization. The model was evaluated using Stratified 10-Fold Cross Validation on 37,146 valid reviews. The experimental results show that the proposed approach improved the average Macro F1-score by 102.42%, increasing from 0.1704 to 0.3450, while achieving an accuracy of 58%. These findings indicate that the combination of *slang normalization*, TF-IDF n-gram, SMOTE, and *hyperparameter* optimization effectively enhances the model's ability to recognize minority classes in imbalanced e-commerce review datasets.

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Keyword : ***Analisis Sentiment, Multinomial Naive Bayes, SMOTE, TF-IDF n-gram, Tokopedia***

1. PENDAHULUAN

Analisis sentimen pada ulasan *e-commerce* merupakan salah satu penerapan *Natural Language Processing* (NLP) yang digunakan untuk mengekstraksi opini pelanggan guna mendukung evaluasi produk dan pengambilan keputusan (Darwiesh et al., 2022). Pada platform seperti Tokopedia, ulasan pengguna umumnya mengandung bahasa informal (*slang*), singkatan, serta penulisan yang tidak konsisten sehingga dapat menurunkan kinerja model klasifikasi apabila tidak dilakukan prapemrosesan yang sesuai (Budiasa, 2021). Selain itu, distribusi kelas pada data ulasan *e-commerce* umumnya tidak seimbang, dengan dominasi ulasan positif yang menyebabkan model cenderung bias terhadap kelas mayoritas dan kurang mampu mengenali kelas minoritas (Munthe et al., 2025).

Berbagai penelitian telah menerapkan algoritma Multinomial Naive Bayes (MNB) pada analisis sentimen karena sederhana, efisien, dan memiliki performa yang kompetitif pada klasifikasi teks (Haris et al., 2025). Beberapa penelitian melaporkan tingkat akurasi antara 77% hingga 90% pada berbagai domain (Agustina et al., 2022; Faradian, 2024; Hanafi & R, 2024; Nisa & Kurniawan, 2024; Putri et al., 2023). Namun, sebagian besar penelitian masih memiliki beberapa keterbatasan, yaitu normalisasi *slang* yang belum adaptif, penggunaan TF-IDF yang masih didominasi unigram, serta optimasi *hyperparameter* yang masih terbatas sehingga konfigurasi model umumnya menggunakan nilai bawaan.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan mengoptimalkan algoritma MNB untuk klasifikasi ordinal ulasan Tokopedia berdasarkan rating 1–5 melalui integrasi normalisasi *slang* berbasis leksikon hibrida, representasi fitur TF-IDF n-gram, penyeimbangan data menggunakan Synthetic Minority Over-sampling Technique (SMOTE), dan optimasi *hyperparameter*. Pendekatan yang diusulkan diharapkan mampu meningkatkan kemampuan model dalam mengenali kelas minoritas serta memberikan alternatif strategi optimasi bagi analisis sentimen ulasan berbahasa Indonesia.

2. LANDASAN TEORI

Analisis Sentimen

Analisis sentimen (*sentiment analysis* atau *opinion mining*) merupakan salah satu bidang *Natural Language Processing* (NLP) yang bertujuan mengidentifikasi dan mengklasifikasikan opini atau polaritas sentimen dalam data teks terhadap suatu entitas, seperti produk atau layanan (Liu, 2012). Pada konteks *e-commerce*, analisis sentimen dapat digunakan untuk mengklasifikasikan ulasan berdasarkan tingkat rating bintang (1–5) yang termasuk dalam permasalahan klasifikasi ordinal (Kovács, 2025).

Analisis sentimen banyak diterapkan untuk mengekstraksi informasi dari ulasan pelanggan guna mendukung evaluasi produk dan pengambilan keputusan berbasis data. Penelitian ini menggunakan pendekatan *machine learning* dengan algoritma Multinomial Naive Bayes karena efisien, memiliki kompleksitas komputasi yang rendah, serta memberikan kinerja yang kompetitif pada tugas klasifikasi teks (Patil & Adhiya, 2023)

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dengan asumsi bahwa setiap fitur bersifat independen terhadap fitur lainnya pada kelas yang sama (Bishop, 2006). Algoritma ini banyak

digunakan pada klasifikasi teks karena mampu memodelkan distribusi frekuensi kata secara efisien serta memiliki kompleksitas komputasi yang rendah (Manning et al., 2008). Probabilitas posterior suatu dokumen terhadap kelas tertentu dihitung menggunakan Persamaan (1).

$$P(C|D) = \frac{P(D|C) \cdot P(C)}{P(D)} \quad (1)$$

Pada klasifikasi teks, bentuk tersebut disederhanakan menjadi Persamaan (2), sehingga proses klasifikasi dilakukan dengan memilih kelas yang memiliki probabilitas posterior terbesar.

$$P(C|D) \propto P(C) \cdot \prod_{i=1}^{n} P(f_i|C) \quad (2)$$

Untuk menghindari *zero-frequency problem*, probabilitas kemunculan setiap kata diestimasi menggunakan *additive smoothing* (Laplace/Lidstone smoothing) sebagaimana ditunjukkan pada Persamaan (3).

$$P(f_i \mid C) = \frac{Count(f_i, C) + \alpha}{TotalToken(C) + \alpha V} \quad (3)$$

MNB dipilih dalam penelitian ini karena sederhana, efisien, serta memberikan kinerja yang kompetitif pada berbagai tugas klasifikasi dokumen, khususnya ketika dikombinasikan dengan representasi fitur TF-IDF (Manning et al., 2008; McCallum & Nigam, 1998).

TF-IDF dan N-gram

Term Frequency–Inverse Document Frequency (TF-IDF) merupakan metode pembobotan fitur yang digunakan untuk merepresentasikan dokumen berdasarkan tingkat kepentingan setiap kata. Metode ini memberikan bobot tinggi pada kata yang sering muncul dalam suatu dokumen tetapi jarang ditemukan pada dokumen lain, sehingga mampu meningkatkan kemampuan diskriminasi fitur pada proses klasifikasi teks (Manning et al., 2008; Sparck Jones, 1972). Bobot TF-IDF dihitung menggunakan Persamaan (4).

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \quad (4)$$

Untuk menangkap hubungan antarkata, TF-IDF dikombinasikan dengan model n-gram, yaitu representasi fitur yang membentuk urutan kata secara berurutan dalam teks (Jurafsky & Martin, 2009). Pendekatan ini memungkinkan model mempertahankan konteks lokal yang tidak dapat direpresentasikan hanya menggunakan kata tunggal, seperti frasa *"tidak puas"* atau *"sangat baik"*.

Penelitian ini mengevaluasi tiga konfigurasi n-gram, yaitu unigram (*n*=1), bigram (*n*=2), dan trigram (*n*=3), untuk memperoleh representasi fitur yang lebih informatif dalam klasifikasi ulasan Tokopedia.

Normalisasi Slang

Normalisasi *slang* merupakan proses mengubah kata tidak baku menjadi bentuk baku untuk mengurangi variasi penulisan pada data teks (Jurafsky & Martin, 2009). Pada ulasan *e-commerce* berbahasa Indonesia, proses ini penting karena pengguna sering menggunakan kata *slang*, singkatan, maupun kesalahan penulisan (*typo*). Dalam penelitian ini, normalisasi dilakukan menggunakan kamus leksikon hibrida untuk mengonversi kata tidak baku menjadi bentuk baku sesuai kaidah bahasa Indonesia sehingga meningkatkan konsistensi fitur pada proses klasifikasi (Lin & Nuha, 2023).

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE merupakan metode *oversampling* yang digunakan untuk mengatasi ketidakseimbangan kelas dengan menghasilkan data sintetis pada kelas minoritas, bukan sekadar menduplikasi sampel yang sudah ada (Chawla et al., 2002). Data sintetis

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

dibentuk berdasarkan interpolasi antara suatu sampel minoritas dan *k-nearest neighbors*-nya sehingga distribusi kelas menjadi lebih seimbang.

Dalam penelitian ini, SMOTE diterapkan pada data pelatihan untuk meningkatkan representasi kelas minoritas dan mengurangi kecenderungan model mengklasifikasikan data ke kelas mayoritas. Pendekatan ini dipilih karena mampu meningkatkan kemampuan generalisasi model tanpa menghilangkan informasi pada kelas mayoritas (Chawla et al., 2002; He & Garcia, 2009).

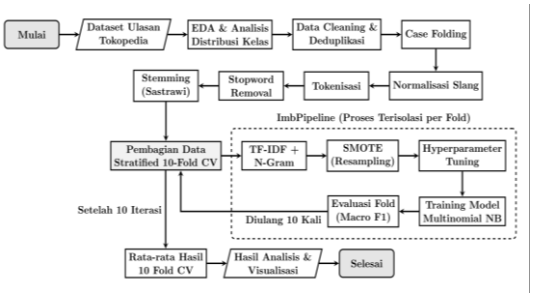
Hyperparameter Optimization

Hyperparameter optimization merupakan proses pencarian kombinasi parameter terbaik untuk meningkatkan kinerja model pembelajaran mesin. Pada penelitian ini, optimasi dilakukan menggunakan GridSearchCV dengan evaluasi Stratified 10-Fold Cross Validation untuk memperoleh konfigurasi parameter yang menghasilkan performa terbaik. Parameter yang dioptimalkan meliputi nilai alpha pada Multinomial Naive Bayes serta parameter TF-IDF, seperti n-gram_range dan max_features, sehingga diperoleh model yang lebih optimal untuk klasifikasi ulasan berbahasa Indonesia (Bergstra & Bengio, 2012; Pedregosa et al., 2011).

3. METODOLOGI

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimen komputasi mengevaluasi kinerja algoritma Multinomial Naive Bayes pada klasifikasi ordinal ulasan Tokopedia. Tahapan penelitian ini meliputi akuisisi data, prepemrosesan teks, pengembangan dan optimasi model, serta evaluasi kinerja menggunakan Accuracy, Precision, Recall, dan Macro F1-Score. Alur penelitian ditunjukkan pada Gambar 1.



Gambar 1 Tahapan penelitian

Tahap prapemrosesan mencakup *cleaning*, *case folding*, *tokenization*, normalisasi *slang*, *stopword removal*, dan *stemming*. Selanjutnya dilakukan serangkaian eksperimen untuk mengevaluasi konfigurasi TF-IDF n-gram, penerapan SMOTE, dan *hyperparameter tuning* pada Multinomial Naive Bayes. Evaluasi model menggunakan Stratified 10-Fold Cross-Validation dengan metrik utama Macro F1-Score.

Perangkat dan Lingkungan Penelitian

Eksperimen diimplementasikan menggunakan bahasa pemrograman Python 3.11. Proses pengembangan model dilakukan pada lingkungan Ubuntu 24.04 LTS menggunakan Visual Studio Code dan Miniconda, sedangkan eksplorasi data awal dilakukan dengan Kaggle Notebooks dan Google Colaboratory. Library utama yang digunakan meliputi Pandas, NumPy, NLTK, Sastrawi, Scikit-learn, Imbalanced-learn, dan Matplotlib untuk mendukung proses prapemrosesan, ekstraksi fitur, pemodelan, evaluasi, serta visualisasi hasil.

Dataset Penelitian

Penelitian ini menggunakan dataset sekunder *Tokopedia Product Reviews* yang dipublikasikan oleh Farhan (Farhan, 2019) di platform Kaggle. Dataset berisi teks ulasan pelanggan (*text*) dan *rating* produk sebagai variabel target. Setelah proses deduplikasi berbasarkan isi teks, jumlah data berkurang dari 40.607 menjadi 37.301 ulasan unik yang digunakan dalam penelitian.

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

Klasifikasi dilakukan terhadap lima kelas ordinal berdasarkan *rating* 1 hingga 5. Evaluasi model menggunakan Stratified 10-Fold Cross-Validation, sehingga seluruh data dimanfaatkan secara bergantian sebagai data latih dan data uji. Distribusi kelas setelah deduplikasi ditunjukkan pada Tabel 1.

Tabel 1. Distribusi dataset

Rating	*Jumlah*	*Persentase (%)*
1	531	1,42
2	380	1,02
3	1.762	4,72
4	6.995	18,75
5	27.633	74,08
Total	37.301	100,00

Distribusi kelas tidak seimbang menyebabkan dominasi kelas rating 5. Oleh karena itu, penelitian ini menerapkan SMOTE pada data latih di setiap *fold* untuk meningkatkan representasi kelas minoritas, sedangkan Macro F1-Score digunakan sebagai metrik evaluasi utama karena lebih sesuai untuk mengevaluasi performa model pada data tidak seimbang

Skenario Eksperimen

Eksperimen dirancang menggunakan pendekatan bertahap (*incremental ablation study*) untuk mengevaluasi kontribusi setiap strategi optimasi terhadap kinerja model Multinomial Naive Bayes. Setiap tahap mewarisi konfigurasi terbaik yang diperoleh pada tahap sebelumnya sehingga pengaruh masing-masing strategi optimsi dapat dievalusi secara bertahap. Konfigurasi setiap tahap eksperimen disajikan pada Tabel 2.

Tabel 2. Skenario eksperimen

Tahap	*Normal isasi*	*TF-IDF*	*SMOTE*	*Tuning*
1	×	unigram	×	*Default*
2	✓	unigram	×	*Default*
3	✓	n-gram	×	*Default*
4	✓	n-gram	✓	*Default*
5	✓	n-gram	✓	*Tuned*

Seluruh eksperimen dievaluasi menggunakan Stratified 10-Fold Cross Validation dengan konfigurasi *pipeline* yang sama agar hasil setiap tahap dapat dibandingkan secara adil. Performa setiap konfigurasi dianalisis menggunakan metrik Accuracy, Precision, Recall, dan Macro F1-Score. Macro F1-Score dipilih sebagai metrik utama karena memberikan evaluasi yang lebih representatif pada dataset dengan distribusi kelas tidak seimbang.

4. HASIL DAN PEMBAHASAN

Hasil Prapemrosesan Teks

Dataset *Tokopedia Product Reviews* yang semula berjumlah 40.607 ulasan dipraproses melalui tahapan deduplikasi, *cleaning*, *case folding*, normalisasi *slang*, *tokenization*, *stopword removal*, dan *stemming*. Setelah menghapus data duplikat dan 155 baris tidak valid, diperoleh 37.146 ulasan yang digunakan pada seluruh eksperimen. Ringkasan perubahan jumlah data disajikan pada Tabel 3.

Tabel 3. Perubahan jumlah data

Tahapan	*Jumlah Data*
Dataset awal	40.607
Setelah deduplikasi	37.301
Dataset akhir setelah prapemrosesan	37.146

Hasil tersebut menunjukkan bahwa proses pembersihan data hanya menghilangkan sekitar 0,42% dari data

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

hasil deduplikasi sehingga sebagian besar informasi pada dataset tetap dipertahankan. Dengan demikian, proses prapemrosesan berhasil meningkatkan kualitas data tanpa mengurangi representativitas dataset secara signifikan.

Untuk mengevaluasi pengaruh normalisasi bahasa informal, penelitian ini membandingkan empat konfigurasi prapemrosesan, yaitu tanpa normalisasi slang, kamus kustom, kamus publik, dan kamus hybrid. Evaluasi dilakukan berdasarkan ukuran *vocabulary* yang dihasilkan setelah prapemrosesan serta performa klasifikasi menggunakan Macro F1-Score. Hasil pengujian disajikan pada Tabel 4.

Tabel 4. Performa variasi kamus

Variasi	*Vocabulary*	*Macro F1-Score*
Tanpa normalisasi	4.640	0,2098
Kamus kustom	4.565	0,2098
Kamus publik	4.247	0,2105
Kamus hybrid	4.215	0,2113

Berdasarkan Tabel 4, penggunaan normalisasi slang berpengaruh terhadap jumlah kosakata unik (*vocabulary*) yang dihasilkan. Konfigurasi tanpa normalisasi menghasilkan ruang fitur terbesar karena variasi penulisan kata tidak baku tetap diperlakukan sebagai token yang berbeda. Sebaliknya, penggunaan kamus publik dan kamus hybrid mampu mengurangi jumlah kosakata unik secara lebih signifikan melalui penyatuan berbagai variasi penulisan ke dalam bentuk baku.

Kamus hybrid menghasilkan ukuran *vocabulary* paling kecil, yaitu 4.215 kata, sekaligus memberikan nilai Macro F1-Score tertinggi sebesar 0,2113. Hasil tersebut menunjukkan bahwa penggabungan kamus publik dengan kamus kustom lebih efektif dibandingkan penggunaan salah satu kamus secara terpisah. Kamus publik mampu menangani variasi bahasa informal yang umum digunakan, sedangkan kamus kustom melengkapi istilah-istilah yang lebih spesifik pada ulasan Tokopedia. Kombinasi keduanya menghasilkan representasi teks yang lebih konsisten sehingga mampu meningkatkan kualitas fitur yang digunakan oleh algoritma Multinomial Naive Bayes.

Peningkatan Macro F1-Score memang relatif kecil dibandingkan konfigurasi lainnya, namun hasil tersebut menunjukkan bahwa kualitas prapemrosesan memberikan kontribusi terhadap peningkatan performa model sebelum dilakukan optimasi pada tahap berikutnya. Oleh karena itu, konfigurasi kamus hybrid dipilih sebagai konfigurasi prapemrosesan terbaik dan digunakan pada seluruh eksperimen selanjutnya, termasuk pengujian variasi TF-IDF n-gram, penerapan SMOTE, dan *hyperparameter tuning*.

Hasil Pengembangan Model

Setelah diperoleh dataset hasil prapemrosesan menggunakan leksikon hybrid, dilakukan pembentukan fitur menggunakan TF-IDF serta pengembangan model klasifikasi berbasis Multinomial Naive Bayes. Seluruh proses ekstraksi fitur, penanganan ketidakseimbangan kelas menggunakan SMOTE, dan pelatihan model diintegrasikan ke dalam pipeline sehingga setiap tahapan hanya diterapkan pada data latih di setiap fold selama proses Stratified 10-Fold Cross Validation. Pendekatan ini bertujuan mencegah terjadinya data leakage dan menghasilkan evaluasi yang lebih objektif.

1) Pengaruh Konfigurasi TF-IDF n-gram

Eksperimen pertama mengevaluasi pengaruh konfigurasi TF-IDF n-gram terhadap performa klasifikasi. Tiga konfigurasi yang diuji meliputi unigram (1,1), unigram-bigram (1,2), dan unigram-bigram-trigram (1,3). Hasil evaluasi menunjukkan bahwa konfigurasi unigram-bigram memberikan performa

terbaik dengan nilai Macro F1-Score sebesar 0,3384, lebih tinggi dibandingkan unigram (0,3030) maupun unigram-bigram-trigram (0,3359). Selain itu, seluruh konfigurasi menghasilkan nilai standar deviasi yang rendah (0,0062–0,0102), yang menunjukkan bahwa performa model relatif konsisten pada setiap *fold cross-validation*.

Peningkatan performa pada konfigurasi unigram-bigram menunjukkan bahwa penambahan fitur bigram membantu model menangkap informasi kontekstual yang tidak dapat direpresentasikan oleh unigram saja, terutama pada frasa seperti *tidak sesuai* atau *kurang bagus*. Sebaliknya, penambahan trigram tidak memberikan peningkatan performa. Hal ini mengindikasikan bahwa perluasan ruang fitur hingga trigram menambah kompleksitas representasi tanpa memberikan informasi tambahan yang cukup untuk meningkatkan kemampuan klasifikasi.

2) Pengaruh Penanganan Ketidakseimbangan Kelas

Dataset penelitian memiliki distribusi kelas yang tidak seimbang, dengan kelas rating 5 mendominasi sebanyak 27.530 ulasan, sedangkan kelas rating 1, 2, 3, dan 4 masing-masing hanya terdiri atas 52**8**, 380, 1.746, dan 6.962 ulasan. Untuk mengatasi kondisi tersebut, penelitian menerapkan Synthetic Minority Over-sampling Technique (SMOTE) pada data latih di setiap fold *Stratified 10-Fold Cross Validation*.

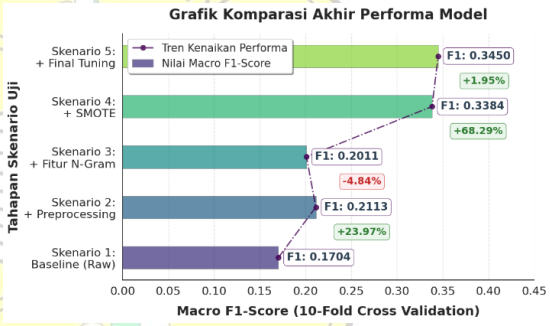
Penerapan SMOTE menghasilkan distribusi data latih yang seimbang, di mana seluruh kelas minoritas ditingkatkan hingga memiliki jumlah sampel yang setara dengan kelas mayoritas (**27.530** sampel). Namun, proses oversampling hanya dilakukan pada data latih dalam setiap fold, sedangkan data validasi tetap mempertahankan distribusi aslinya. Pendekatan ini mencegah terjadinya *data leakage* dan memastikan bahwa hasil evaluasi mencerminkan kemampuan generalisasi model terhadap data yang belum pernah dilihat.

3) Hasil Hyperparameter Tuning

Tahap selanjutnya adalah melakukan optimasi parameter menggunakan *GridSearchCV*. Hasil pencarian menunjukkan bahwa konfigurasi terbaik menggunakan normalisasi Hybrid Lexicon, TF-IDF unigram-bigram**,** alpha = 0,1**,** fit_prior = True, dan k_neighbors = 3 pada SMOTE. Konfigurasi tersebut menghasilkan Macro F1-Score sebesar 0,3450, yang menjadi performa terbaik pada penelitian ini.

4) Performa Akhir Model

Ringkasan perkembangan performa model pada setiap tahapan eksperimen disajikan pada Gambar 2.



Gambar 2. Performa akhir model

Berdasarkan Gambar 2, setiap tahapan memberikan kontribusi terhadap peningkatan performa model. Dibandingkan model baseline, model akhir meningkatkan nilai Macro F1-Score dari 0,1704 menjadi 0,3450 atau sebesar 102,42%**.** Hasil tersebut menunjukkan bahwa kombinasi prapemrosesan, representasi fitur TF-IDF, penanganan ketidakseimbangan kelas menggunakan SMOTE, dan optimasi *hyperparameter* mampu meningkatkan performa klasifikasi ordinal ulasan Tokopedia.

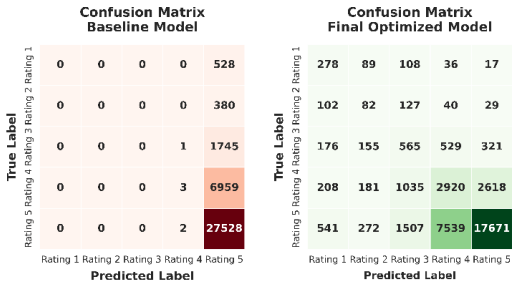
Analisis Model

1) Analisis Confusion Matrix

Gambar 3 menunjukkan confusion matrix dari model terbaik hasil integrasi

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

prapemrosesan, TF-IDF unigram-bigram, SMOTE, dan optimasi hyperparameter.



Gambar 3. Confusion matrix baseline dan model teroptimasi

Secara umum, sebagian besar prediksi yang benar berada pada diagonal utama confusion matrix, sedangkan kesalahan klasifikasi didominasi oleh kelas-kelas yang berdekatan, seperti antara rating 1 dan rating 2 atau antara rating 4 dan rating 5. Pola tersebut menunjukkan bahwa model mampu menangkap karakteristik ordinal pada ulasan, meskipun masih terdapat kemiripan linguistik antar-rating yang berdekatan.

Hasil ini mengindikasikan bahwa penerapan SMOTE di dalam pipeline berhasil mengurangi bias terhadap kelas mayoritas sehingga model mampu mengenali kelas minoritas dengan lebih baik dan menghasilkan kemampuan generalisasi yang lebih seimbang.

2) Analisis Feature Importance

Analisis terhadap fitur dengan bobot probabilitas tertinggi menunjukkan bahwa model berhasil menangkap karakteristik linguistik yang berbeda pada setiap kategori rating. Rating 1 dan 2 didominasi oleh kata dan frasa bernuansa keluhan, seperti tidak sesuai, barang tidak, kecewa, dan rusak. Pada Rating 3, fitur yang dominan berkaitan dengan proses transaksi, seperti barang, terima kasih, dan sesuai harga, yang mencerminkan ulasan dengan sentimen netral. Sementara itu, Rating 4 dan 5 didominasi oleh kata-kata positif, seperti bagus, cepat, mantap, dan frasa barang sesuai, yang merepresentasikan kepuasan pelanggan. Temuan ini menunjukkan bahwa representasi TF-IDF dengan konfigurasi unigram–bigram mampu menangkap konteks lokal sehingga membantu model membedakan karakteristik linguistik pada setiap tingkat rating.

Pembahasan

Hasil penelitian menunjukkan bahwa peningkatan performa model diperoleh melalui kombinasi normalisasi *slang*, representasi fitur TF-IDF n-gram, penanganan ketidakseimbangan kelas menggunakan SMOTE, serta optimasi *hyperparameter tuning*. Diantara seluruh tahapan, penerapan SMOTE memberikan kontribusi peningkatan terbesar, menunjukkan bahwa ketidakseimbangan merupakan faktor utama yang mempengaruhi kinerja klasifikasi ordinal pada dataset ulasan Tokopedia. Optimasi *hyperparameter* selanjutnya menghasilkan konfigurasi model yang lebih optimal sehingga meningkatkan kemampuan model dalam mengenali kelas minoritas. Temuan ini menunjukkan bahwa integrasi seluruh tahapan optimasi lebih efektif dibandingkan penerapan masing-masing teknik secara terpisah.

5. KESIMPULAN

Penelitian ini berhasil mengoptimalkan algoritma Multinomial Naive Bayes untuk klasifikasi ordinal ulasan Tokopedia melalui integrasi normalisasi slang, representasi fitur TF-IDF n-gram, SMOTE, dan *hyperparameter tuning*. Pendekatan yang diusulkan menghasilkan peningkatan Macro F1-Score dari 0,1704 menjadi 0,3450 atau sebesar 102,42% dibandingkan model baseline, serta meningkatkan kemampuan model dalam mengenali kelas minoritas pada data yang tidak seimbang.

DAFTAR PUSTAKA

Agustina, N., Citra, D. H., Purnama, W., Nisa, C., & Kurnia, A. R. (2022). Implementasi algoritma Naive Bayes

untuk analisis sentimen ulasan Shopee pada Google Play Store. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(1),47–54. https://doi.org/10.57152/malcom.v2i1.195

Kovács, B. (2025). *Five is the brightest star. But by how much? Testing the equidistance of star ratings in online reviews*. *Organizational Research Methods*, *28*(2), 269–295. https://doi.org/10.1177/10944281231223412

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305. https://doi.org/10.5555/2188385.2188395

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Budiasa, I. G. (2021). Slang language in Indonesian social media. *Lingual Journal of Language and Culture*, *11*(1), 30–30. https://doi.org/10.24843/LJLC.2021.v11.i01.p06

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Darwiesh, A., Alghamdi, Mohammed. I., El-Baz, A. H., & Elhoseny, M. (2022). Social media big data analysis: Towards enhancing competitiveness of firms in a post-pandemic world. *Journal of Healthcare Engineering*, *2022*, 6967158. https://doi.org/10.1155/2022/6967158

Faradian, H. (2024). Analisis sentimen terhadap penutupan tiktok shop menggunakan algoritma naïve bayes classifier pada media sosial X. *Scientica : Jurnal Ilmiah Sain dan Teknologi*, *2*(4), 150–163. http://repository.unas.ac.id/id/eprint/10417.

Farhan. (2019). *Tokopedia product Reviews* [Dataset]. Kaggle. https://doi.org/10.34740/kaggle/dsv/562904

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), 1263-1284.

Hanafi, M. R., & R, R. K. (2024). Analisis sentimen pada ulasan aplikasi Sirekap di Google Play menggunakan algoritma Naive Bayes: Sentiment analysis on Sirekap app reviews on Google Play using Naive Bayes algorithm. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1578–1586. https://doi.org/10.57152/malcom.v4i4.1693

Haris, A., Tukino, T., Hananto, A., & Nurapriani, F. (2025). Penerapan algoritma naive bayes untuk klasifikasi sentimen ulasan pengguna aplikasi PIDI umkm dengan pembobotan fitur TF-IDF. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(5), 8457–8462. https://doi.org/10.36040/jati.v9i5.15019

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Prentice Hall.

Lin, C.-H., & Nuha, U. (2023). Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy. *Journal of Big Data*, *10*(1), 88. https://doi.org/10.1186/s40537-023-00782-9

Liu, B. (2012). *Sentiment analysis and subjectivity* (2nd Edition). CRC Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In AAAI Workshop on Learning for Text Categorization (pp. 41–48). AAAI Press.

Munthe, S. R., Samsir, & Levianti, R. A. (2025). Hybrid framework addressing imbalanced data in Indonesian e-commerce multimodal sentiment analysis. *Teknika*, *14*(3), 363–370. https://doi.org/10.34148/teknika.v14i3.1332

Nisa, H. L., & Kurniawan, M. H. S. (2024). Analisis sentimen terhadap komentar aplikasi allstats BPS dengan klasifikasi Naïve Bayes: Analisis sentimen terhadap komentar aplikasi allstats bps dengan klasifikasi Naïve Bayes. *Emerging Statistics and Data Science Journal*, *2*(3), 315–329. https://doi.org/10.20885/esds.vol2.iss.3.art24

Patil, S. A., & Adhiya, K. P. (2023). Literature survey of lexicon based method and machine learning methods of sentiment analysis of student's feedback. *Mathematical Statistician and Engineering Applications*, *72*(1), 2257–2261.

Pedregosa, F., et al., (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Putri, K. S., Setiawan, I. R., & Pambudi, A. (2023). Analisis sentimen terhadap brand skincare lokal menggunakan Naïve Bayes classifier. *Technologia: Jurnal Ilmiah*, *14*(3), 227–232.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal Of Documentation*, *28*(1), 11–21. Https://Doi.Org/10.1108/Eb026526