Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 619~629
e-ISSN: 2548-6861

Optimizing Feature Extraction for Naïve Bayes Sentiment Analysis

Achmad [1]*, Fikri Budiman [2]*
* Teknik Informatika, Universitas Dian Nuswantoro
111202214062@mhs.dinus.ac.id [1], fikri.budiman@dsn.dinus.ac.id [2]

Article Info

Article history:

Received 2025-12-16
Revised 2025-12-27
Accepted 2026-01-08

Keyword:

Naïve Bayes,
Sentiment Analysis,
Feature Extraction,
Optimization.

ABSTRACT

The rapid growth of e-commerce platforms such as Tokopedia has generated a large volume of user reviews containing diverse opinions about products and services. These reviews reflect consumer perceptions and provide valuable insights for business decision-making. This study aims to enhance sentiment analysis performance by optimizing the Naïve Bayes algorithm through a comparison of two feature extraction techniques, namely Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). The dataset consists of 5,400 Tokopedia product reviews obtained from the Kaggle platform, which are categorized into positive and negative sentiments. The research process includes text preprocessing consisting of text cleaning, case folding, tokenization, stopword removal, and stemming, feature extraction using Bag of Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF), handling data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), and model training using the Naïve Bayes. The dataset is divided into 80% training data and 20% testing data, and model performance is evaluated using accuracy, precision, recall, and F1-score. The results show that BoW achieved the highest accuracy of 93%, while TF-IDF reached 83%, indicating that BoW provides more effective feature representation and more stable performance for Naïve Bayes-based sentiment analysis on this dataset.

I. PENDAHULUAN

Tokopedia adalah salah satu situs e-commerce terbesar di Indonesia, yang telah berkembang secara signifikan. Perkembangan Tokopedia merupakan indikasi perubahan penting dalam lingkungan e-commerce di Indonesia, yang didorong oleh kerja sama strategis, inovasi teknologi, dan perluasan akses internet nasional [1]. Tokopedia, yang juga termasuk sebagai platform e-commerce terkemuka di Indonesia, menghasilkan data ulasan pengguna dalam jumlah besar yang menawarkan wawasan berharga mengenai kepuasan pelanggan melalui analisis sentimen [2]. Kehadiran fitur ulasan konsumen menjadi salah satu aspek penting karena berfungsi sebagai indikator kualitas produk sekaligus bahan pertimbangan bagi calon pembeli dalam mengambil keputusan.

Sentiment analysis, yang juga dikenal sebagai opinion mining, merupakan teknik pemrosesan bahasa alami yang bertujuan untuk mengidentifikasi, mengekstrak, dan mengkategorikan opini yang diekspresikan dalam teks ke dalam sentimen positif, negatif dan netral [3]. Beberapa penelitian juga menegaskan bahwa tugas utama sentiment analysis adalah menentukan polaritas opini untuk mengidentifikasi sikap pengguna terhadap suatu entitas [4][5]. Pada konteks ulasan e-commerce, analysis ini berperan penting untuk mengetahui kepuasan pelanggan serta reputasi produk. Penerapan sentiment analysis terbukti mampu mendukung proses evaluasi kualitas layanan, memberikan masukan untuk pengembangan produk, serta menjadi dasar dalam pengambilan keputusan strategis berbasis data, sehingga menyebabkan sentiment analysis berkembang dengan pesat dan banyak digunakan [6][7][8].

Untuk dapat melakukan klasifikasi secara akurat, sentiment analysis memerlukan ekstraksi fitur (*feature extraction*). Ekstraksi fitur merupakan suatu tahap untuk memproses sebuah kata yang menjelaskan suatu sentiment

yang terdapat pada suatu dataset untuk mengekstraksi atribut sehingga mendapatkan nilai unik, penting, dan tidak duplikasi [9]. Ekstraksi fitur yang kurang tepat dapat menurunkan kinerja algoritma [10]. Oleh karena itu, dua metode populer yang banyak digunakan adalah Bag of Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF). BoW menghitung frekuensi kemunculan kata pada seluruh dokumen yang di proses [11][12]. Sedangkan TF-IDF adalah perhitungan atau pembobotan kata melalui teknik tokenisasi, stopwords, dan stemming, dan frekuensi munculnya kata dalam dokumen yang diberikan menunjukkan pentingnya kata itu di dalam sebuah dokumen [13][14].

Beberapa penelitian mengenai sentiment analysis dengan membandingkan metode ekstraksi fitur telah dilakukan sebelumnya. Salah satu penelitian menunjukkan bahwa pemilihan metode ekstraksi fitur berpengaruh signifikan terhadap performa algoritma Naïve Bayes, dengan TF-IDF terbukti lebih unggul dibandingkan Word2Vec dalam beberapa kasus [15]. Penelitian lain secara khusus menganalisis sentimen pelanggan Tokopedia menggunakan Naïve Bayes dan berhasil mengklasifikasikan komentar pelanggan ke dalam dua kategori utama yaitu positif dan negatif [16]. Selanjutnya, ada juga penelitian yang membandingkan TF-IDF, Bag of Words (BoW), dan FastText pada data Twitter terkait kenaikan harga BBM menggunakan algoritma SVM, di mana hasilnya menunjukkan bahwa BoW memberikan performa terbaik [17]. Penelitian lainnya membandingkan TF-IDF dan BoW dalam analisis sentimen ulasan diet kopi Americano di Twitter menggunakan Naïve Bayes, dengan hasil bahwa TF-IDF memberikan akurasi lebih tinggi dibandingkan BoW [18]. Selain itu, analisis sentimen pada review pengguna Tokopedia mengenai produk kesehatan dengan metode Naïve Bayes juga menghasilkan akurasi sebesar 88%, yang membuktikan efektivitas algoritma ini dalam mengklasifikasikan opini ke dalam kategori positif dan negatif [19].

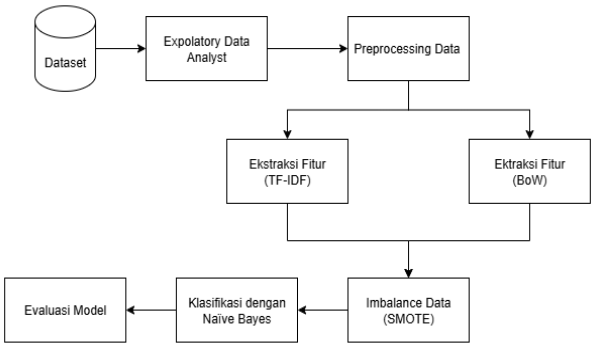
Namun, penerapan Naïve Bayes dalam analisis sentimen tidak terlepas dari sejumlah keterbatasan. Asumsi independensi fitur seringkali tidak realistis, representasi BoW dapat menimbulkan masalah sparsity, dan penggunaan TF-IDF kadang justru menurunkan performa pada algoritma probabilistik seperti Naïve Bayes. Kondisi ini membuat pentingnya dilakukan perbandingan metode ekstraksi fitur untuk mengetahui kombinasi yang paling sesuai dalam meningkatkan akurasi klasifikasi sentimen [20].

Meskipun berbagai penelitian telah dilakukan, terdapat research gap karena belum banyak studi yang secara eksplisit membandingkan performa Naïve Bayes menggunakan BoW dan TF-IDF pada data ulasan Tokopedia. Kebanyakan penelitian hanya menguji salah satu metode pembobotan atau pada domain aplikasi selain e-commerce.

Berdasarkan hal tersebut, penelitian ini bertujuan untuk membandingkan performa algoritma Naïve Bayes menggunakan dua metode ekstraksi fitur, yaitu BoW dan TF-IDF, dalam klasifikasi sentimen pada ulasan pengguna Tokopedia. Penelitian ini diharapkan dapat memberikan kontribusi dalam menentukan metode representasi fitur yang paling sesuai untuk meningkatkan akurasi klasifikasi pada domain e-commerce. Penelitian ini diharapkan dapat memberikan gambaran mengenai metode representasi fitur yang lebih sesuai untuk meningkatkan akurasi klasifikasi sentimen pada domain e-commerce, dengan evaluasi performa menggunakan confusion matrix, khususnya pada data ulasan Tokopedia yang belum banyak dilakukan pada penelitian sebelumnya

II. METODE

Cara Untuk mencapai tujuan penelitian yang telah dirumuskan sebelumnya, diperlukan suatu tahapan metodologi yang sistematis. Metode penelitian ini mencakup proses pengolahan data ulasan Tokopedia mulai dari tahap pengumpulan data, eksplorasi, preprocessing, ekstraksi fitur, hingga evaluasi model. Alur lengkap dari tahapan penelitian ditunjukkan pada Gambar 1 dalam bentuk flow sistem, yang menggambarkan proses mulai dari dataset hingga evaluasi performa model menggunakan algoritma Naïve Bayes dengan ekstraksi fitur BoW dan TF-IDF.



Gambar 1. Tahapan Penelitian

Dari flow sistem diatas dapat dilihat bahwa penelitian ini terdiri dari beberapa langkah utama dalam proses sentiment analysis. Agar lebih mudah dipahami, pada bagian berikut akan dijelaskan secara lebih rinci mengenai setiap tahapan yang ada, mulai dari pengolahan dataset, eksplorasi data, preprocessing, ekstraksi fitur, penanganan ketidakseimbangan data, pembangunan model, hingga evaluasi performa model.

A. Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari platform Kaggle dan terdiri dari 5.400 entri ulasan konsumen terhadap produk pada e-commerce Tokopedia dengan 11 kolom, yaitu *Category*, *Product Name*, *Location*, *Price*, *Overall Rating*, *Number Sold*, *Total Review*, *Customer Rating*, *Customer Review*, *Sentiment*, dan *Emotion*. Kolom *Category* menunjukkan kategori produk, *Product Name* berisi nama produk, *Location* merepresentasikan lokasi penjual, *Price* menunjukkan harga produk, *Overall Rating* menggambarkan rata-rata penilaian produk, *Number Sold*

menunjukkan jumlah produk terjual, dan *Total Review* menyatakan jumlah ulasan yang diterima produk. Kolom *Customer Rating* berisi nilai numerik penilaian konsumen yang digunakan sebagai dasar pelabelan sentimen dengan pengelompokan rating 1–2 sebagai sentimen negatif, rating 3 sebagai netral, dan rating 4–5 sebagai sentimen positif, dengan distribusi kelas yang tidak seimbang. Sementara itu, kolom *Customer Review* memuat teks ulasan konsumen dengan karakteristik bahasa yang beragam, sedangkan kolom *Sentiment* dan *Emotion* menyediakan label sentimen dan emosi yang menyertai setiap ulasan.

B. Exploratory Data Analysis (EDA)

Pada tahap Exploratory Data Analysis (EDA) dilakukan analisis untuk memahami karakteristik awal dataset sebelum masuk ke tahap preprocessing dan pemodelan. Analisis ini meliputi pemeriksaan struktur data untuk melihat tipe data dan jumlah nilai non-null pada setiap kolom. Selanjutnya dilakukan pemeriksaan deskripsi statistik yang memberikan ringkasan distribusi data, nilai unik, serta statistik dasar lainnya. Selain itu, juga dilakukan pengecekan nilai kosong untuk memastikan kelengkapan data, serta perhitungan jumlah nilai unik pada setiap kolom untuk memahami variasi data yang tersedia.

Distribusi Customer Rating dianalisis untuk mengetahui persebaran ulasan berdasarkan skor numerik (1–5), kemudian divisualisasikan dalam bentuk diagram batang agar lebih mudah dipahami. Hasil analisis menunjukkan adanya distribusi kelas sentimen yang tidak seimbang, di mana ulasan positif cenderung lebih dominan dibandingkan dengan ulasan negatif serta terdapat sejumlah data duplikat yang berpotensi mempengaruhi kualitas hasil penelitian. Untuk mengatasi masalah data duplikat yang terjadi dalam proses EDA, maka penghapusan data duplikat akan dilakukan pada tahap preprocessing teks selanjutnya.

C. Preprocessing Data

Tahap preprocessing data merupakan langkah awal yang sangat penting dalam analisis sentimen, terutama ketika teks bersumber dari e-commerce atau media sosial yang sering kali tidak mengikuti kaidah tata bahasa dan ejaan yang baku. Preprocessing diperlukan untuk membersihkan serta menormalisasi teks sebelum dianalisis, karena data tekstual umumnya dipenuhi dengan singkatan, emoticon, emoji, kalimat terpotong, dan bahasa gaul. Proses ini bertujuan untuk menghasilkan teks yang lebih terstruktur dan konsisten agar dapat diolah secara optimal oleh algoritma machine learning [21].

Dalam penelitian ini, tahap preprocessing diawali dengan pembersihan teks (cleaning), yang mencakup penghapusan kata tidak baku seperti “yg”, kata penghubung umum seperti “di”, “ke”, “dari”, dan “dan”, serta karakter non-alfabet yang tidak memiliki makna kontekstual. Seluruh teks kemudian diubah menjadi huruf kecil (lowercase) yang berguna untuk menjaga konsistensi penulisan. Setelah itu, dilakukan tokenisasi (tokenization) untuk memecah setiap kalimat ulasan menjadi satuan kata (token) agar lebih mudah dianalisis. Tahap berikutnya ialah normalisasi teks, yaitu proses penyamaan bentuk penulisan kata seperti mengubah pengulangan huruf (“bangetttt”) menjadi bentuk yang baku (“banget”) serta menyesuaikan penulisan kata tidak baku agar sesuai dengan kaidah bahasa Indonesia. Kemudian dilakukan penghapusan kata umum (stopword removal) dengan memanfaatkan daftar stopwords bahasa Indonesia untuk menghilangkan kata-kata yang tidak memiliki makna penting terhadap sentimen, misalnya “yang”, “untuk”, atau “dengan”. Tahap terakhir adalah stemming, yaitu proses mengembalikan setiap kata ke bentuk dasarnya menggunakan pustaka Sastrawi, sehingga kata-kata seperti “membeli”, “dibeli”, dan “pembelian” akan diubah menjadi “beli”. Melalui serangkaian langkah tersebut, teks ulasan menjadi lebih bersih, seragam, dan siap digunakan pada tahap analisis berikutnya.

D. Ekstraksi Fitur

Tahap ekstraksi fitur bertujuan untuk mengubah data teks menjadi bentuk representasi numerik agar dapat diproses oleh model machine learning. Dalam pengolahan data berbasis teks, kumpulan kata atau kalimat perlu diubah terlebih dahulu menjadi angka sehingga algoritma dapat mengenali pola dan melakukan pembelajaran secara optimal. Pada penelitian ini digunakan dua metode ekstraksi fitur, yaitu Term Frequency–Inverse Document Frequency (TF-IDF) dan Bag of Words (BoW) .

1) *TF-IDF*: Metode pertama pada tahap ekstraksi fitur adalah pembobotan kata menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Pembobotan ini berfungsi untuk mengubah data berbentuk teks menjadi representasi numerik. Dalam bidang machine learning maupun deep learning, data numerik diperlukan agar model dapat bekerja secara optimal. Oleh karena itu, pada penelitian yang berkaitan dengan analisis sentimen, data berupa kumpulan kata atau kalimat perlu diubah terlebih dahulu menjadi bentuk angka melalui proses pembobotan kata. TF-IDF merupakan salah satu teknik yang digunakan dalam pengolahan teks untuk memberikan bobot pada kata-kata dalam sebuah dokumen. Tujuan utama dari metode ini adalah untuk menentukan kata-kata yang paling berpengaruh atau memiliki tingkat kepentingan tinggi dalam suatu dokumen maupun kumpulan dokumen [22]. Term Frequency (TF) sendiri menggambarkan seberapa sering suatu kata muncul di dalam dokumen, dan nilainya dapat dihitung menggunakan rumus pada persamaan (1).

$$Tf = \frac{frekuensi\ kata\ t\ pada\ dokumen\ d}{jumlah\ kata\ pada\ dokumen\ d} \quad (1)$$

Nilai Term Frequency (TF) menggambarkan seberapa sering suatu kata muncul dalam dokumen, tetapi nilai ini belum dapat menunjukkan seberapa penting kata tersebut terhadap isi keseluruhan dokumen. Kata-kata umum

seperti kata sambung atau kata hubung sering kali memiliki nilai TF yang tinggi, padahal tidak memiliki makna penting dalam konteks dokumen. Untuk mengatasi hal tersebut, digunakan teknik tambahan yaitu Inverse Document Frequency (IDF), yang memberikan bobot lebih besar pada kata-kata yang jarang muncul di seluruh kumpulan dokumen. Pendekatan ini membantu sistem untuk membedakan kata yang benar-benar bermakna terhadap konteks dokumen dari kata yang hanya sering muncul tanpa memberikan informasi penting. Dengan demikian, hasil analisis menjadi lebih akurat karena setiap kata memiliki bobot yang sesuai dengan kontribusinya terhadap isi dokumen. Kata yang muncul pada banyak dokumen akan memiliki nilai IDF rendah, sedangkan kata yang jarang muncul akan memiliki nilai IDF tinggi. Setelah nilai TF dan IDF diperoleh, keduanya dikalikan untuk mendapatkan bobot akhir TF-IDF, di mana kata dengan nilai TF-IDF yang tinggi dianggap lebih penting karena berkontribusi besar dalam menentukan topik utama dokumen atau kumpulan dokumen. Nilai IDF dapat dihitung menggunakan rumus seperti pada persamaan (2).

$$IDF = log\frac{N}{n} \quad (2)$$

dengan N adalah jumlah total dokumen dalam korpus, dan n merupakan jumlah dokumen yang mengandung kata tersebut.

2) *BoW*: Metode kedua pada tahap ekstraksi fitur adalah pembobotan kata menggunakan metode Bag of Words (BoW). Metode Bag of Words (BoW) merupakan pendekatan klasik dan sederhana yang digunakan untuk mengubah data teks menjadi bentuk representasi numerik agar dapat diolah oleh model machine learning. Metode ini sering digunakan untuk mengidentifikasi pola dalam teks dengan mengukur frekuensi kemunculan kata.

Konsep dasar dari BoW adalah merepresentasikan setiap dokumen sebagai vektor kata berdasarkan jumlah kemunculan kata dalam dokumen tersebut, tanpa memperhatikan urutan kata. Setiap dokumen diubah menjadi vektor berdimensi V, di mana V merupakan ukuran kosakata (vocabulary size) dari seluruh kumpulan dokumen (corpus) [23]. Nilai pada setiap dimensi menunjukkan seberapa sering suatu kata muncul dalam dokumen tertentu. Dengan demikian, setiap dokumen memiliki representasi numerik yang menggambarkan distribusi kata di dalamnya. Misalkan $V = \{d_1, d_2, \ldots, d_n\}$ adalah himpunan kosakata yang berisi kata-kata unik di seluruh korpus dengan jumlah kata unik sebanyak n. Misalkan f_{di} merupakan frekuensi kemunculan kata d_i pada dokumen D. Maka, representasi Bag of Words dari dokumen D dapat dinyatakan seperti pada persamaan (3).

$$BoW(D) =. (f_{d1}, f_{d2}, \ldots, f_{dn}) \quad (3)$$

Dengan demikian, setiap dokumen dapat dinyatakan sebagai vektor numerik berukuran n, yang merepresentasikan frekuensi kemunculan setiap kata dalam kosakata terhadap dokumen tersebut. Representasi ini kemudian digunakan sebagai masukan (input features) bagi model machine learning pada tahap selanjutnya. Namun, meskipun proses ekstraksi fitur telah menghasilkan representasi numerik yang baik, performa model dapat tetap terpengaruh apabila data pelatihan memiliki distribusi kelas yang tidak seimbang (imbalanced data).

E. Imbalance Data (SMOTE)

Dalam machine learning, salah satu permasalahan yang sering muncul adalah ketidakseimbangan jumlah data antar kelas atau yang dikenal dengan istilah imbalanced data. Kondisi ini terjadi ketika salah satu kelas memiliki jumlah data yang jauh lebih sedikit dibandingkan dengan kelas lainnya, sehingga algoritma cenderung lebih fokus pada kelas dengan jumlah data terbanyak [24]. Akibatnya, model lebih mudah mengenali pola dari kelas mayoritas namun kesulitan dalam mengklasifikasikan kelas minoritas, yang berdampak pada menurunnya performa model secara keseluruhan. Permasalahan ini sering kali menyebabkan model bersifat bias terhadap kelas mayoritas karena proses pembelajaran tidak memperoleh representasi yang cukup dari kelas minoritas.

Salah satu pendekatan yang umum digunakan untuk mengatasi permasalahan tersebut adalah teknik resampling, yaitu proses penyesuaian proporsi jumlah data antar kelas agar distribusinya menjadi lebih seimbang [25]. Pendekatan ini dapat dilakukan dengan dua cara, yaitu undersampling yang mengurangi jumlah data pada kelas mayoritas dan oversampling yang menambah jumlah data pada kelas minoritas. Meskipun undersampling dapat memperkecil ketimpangan, metode ini berpotensi menghilangkan informasi penting dari data mayoritas. Oleh karena itu, metode oversampling yang lebih cerdas seperti Synthetic Minority Oversampling Technique (SMOTE) sering digunakan untuk mengatasi keterbatasan tersebut.

Metode SMOTE bekerja dengan cara membuat data sintetis baru dari kelas minoritas, bukan dengan menggandakan data yang sudah ada [26]. Algoritma ini melakukan interpolasi linier antara titik data minoritas dengan beberapa tetangga terdekatnya berdasarkan jarak Euclidean, sehingga menghasilkan data baru yang berada di antara dua titik data kelas minoritas yang sudah ada. Secara matematis, proses pembentukan data sintetis pada SMOTE dapat dijelaskan pada persamaan (4).

$$X_{new} = X_i + (\widehat{X_k} - X_i) \times \delta \quad (4)$$

di mana X_{new} adalah data sintetis baru, X_i adalah data dari kelas minoritas, $\widehat{X_k}$ merupakan data dari k tetangga terdekat yang memiliki jarak terdekat dengan X_i , dan δ adalah bilangan acak antara 0 dan 1. Persamaan tersebut menunjukkan bahwa sampel baru X_{new} dibentuk di antara dua titik data minoritas yang ada, sehingga meningkatkan

kepadatan distribusi kelas minoritas tanpa menimbulkan duplikasi data.

Secara keseluruhan, penggunaan SMOTE merupakan strategi yang efektif dalam menangani ketidakseimbangan data karena mampu meningkatkan representasi kelas minoritas tanpa mengorbankan data mayoritas. Dengan menambahkan data sintetis berdasarkan karakteristik data yang ada, SMOTE membantu model pembelajaran mesin dalam mempelajari pola dari kelas minoritas secara lebih baik. Pendekatan ini banyak digunakan dalam berbagai bidang klasifikasi untuk meningkatkan akurasi prediksi pada data minoritas.

F. Klasifikasi dengan Naïve Bayes

Algoritma Naïve Bayes merupakan salah satu metode klasifikasi yang banyak digunakan dalam bidang text mining karena memiliki kemampuan tinggi dalam mengolah data berukuran besar secara cepat dan efisien. Algoritma ini bekerja berdasarkan Teorema Bayes, yaitu teorema yang menjelaskan hubungan antara dua probabilitas bersyarat [27]. Prinsip utama dari algoritma ini adalah menghitung peluang suatu kelas berdasarkan bukti atau data yang tersedia, dengan mengasumsikan bahwa setiap fitur atau kata dalam data bersifat independen satu sama lain dalam mempengaruhi hasil klasifikasi. Asumsi independensi inilah yang menjadi dasar penamaan naive (sederhana). Pada algoritma ini dasar teori dari algoritma ini dinyatakan menggunakan Teorema Bayes seperti pada persamaan (5).

$$P(A|B) = \frac{P(B|A \times P(A)}{P(B)} \quad (5)$$

Dengan demikian, persamaan tersebut menyatakan bahwa probabilitas terjadinya suatu kejadian *A*apabila diketahui bahwa kejadian *B* telah terjadi, ditentukan oleh hasil perkalian antara probabilitas *B* terjadi jika *A* benar ($P(B|A)$ dengan probabilitas awal dari $A(P(A))$, kemudian dibagi dengan probabilitas total dari $B(P(B))$. Dengan kata lain, persamaan ini digunakan untuk menghitung seberapa besar kemungkinan suatu hipotesis benar berdasarkan bukti yang tersedia.

Dalam konteks penelitian ini, kejadian *A* diartikan sebagai kelas sentimen yang terdiri atas dua kategori, yaitu positif dan negatif, sedangkan kejadian *B* merepresentasikan ulasan pelanggan Tokopedia atau sekumpulan fitur berupa kata yang terdapat dalam teks ulasan tersebut. Nilai $P(A|B)$menunjukkan probabilitas bahwa sebuah ulasan termasuk ke dalam kelas sentimen tertentu apabila isi ulasan telah diketahui. Sementara itu, $P(B|A)$ menggambarkan probabilitas kemunculan kata atau pola tertentu dalam ulasan apabila diketahui kelas sentimennya, misalnya kata "bagus", "cepat", atau "memuaskan" pada kelas positif, serta kata "lama" atau "rusak" pada kelas negatif. Adapun $P(A)$ merupakan probabilitas awal bahwa sebuah ulasan tergolong dalam salah satu kelas sentimen sebelum memperhitungkan isi teks, sedangkan $P(B)$ adalah probabilitas keseluruhan dari data ulasan yang digunakan dalam proses klasifikasi.

Berdasarkan prinsip tersebut, algoritma Naïve Bayes dalam penelitian ini digunakan untuk menghitung probabilitas setiap kelas sentimen berdasarkan bukti yang terdapat pada teks ulasan pelanggan Tokopedia. Sistem kemudian akan memilih kelas dengan probabilitas tertinggi sebagai hasil klasifikasi akhir.

G. Evaluasi Model

Confusion matrix Confusion matrix merupakan salah satu metode evaluasi yang banyak digunakan dalam mengukur performa model klasifikasi. Pada penelitian ini, confusion matrix dimanfaatkan untuk menganalisis sejauh mana tingkat keakuratan dan ketepatan prediksi dari model Naïve Bayes yang dikombinasikan dengan metode Bag of Words (BoW) serta Naïve Bayes yang dikombinasikan dengan metode Term Frequency–Inverse Document Frequency (TF-IDF). Melalui confusion matrix, dapat diketahui kemampuan masing-masing model dalam membedakan kelas positif dan negatif berdasarkan hasil prediksi terhadap data uji.

Evaluasi performa model dilakukan dengan menghitung beberapa metrik penting, yaitu accuracy, precision, recall, dan F1-score. Nilai accuracy menunjukkan proporsi keseluruhan prediksi yang benar dibandingkan dengan jumlah total data uji sebagaimana ditunjukkan pada persamaan (6). Sementara itu, precision menggambarkan proporsi data yang benar-benar positif dari seluruh data yang diprediksi positif, seperti yang dijelaskan pada persamaan (7). Recall digunakan untuk mengukur sejauh mana model mampu mengenali atau menangkap data positif yang sebenarnya, sebagaimana terdapat pada persamaan (8). Adapun F1-score merupakan rata-rata harmonis antara precision dan recall, yang memberikan keseimbangan antara kedua metrik tersebut seperti ditunjukkan pada persamaan (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; x \; 100\% \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \; x \; 100\% \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \; x \; 100\% \quad (8)$$

$$F1 - score = 2 \; x \; \frac{Precision \; x \; Recall}{Precision + Recall} \; x \; 100\% \quad (9)$$

dengan TP adalah True Positive, TN adalah True Negative, FP adalah False Positive, dan FN adalah False Negative.

Classification report merupakan bentuk ringkasan yang digunakan untuk menilai dan menggambarkan kinerja suatu model klasifikasi, baik dalam kasus klasifikasi biner maupun multi-kelas [28]. Laporan ini mengevaluasi performa model menggunakan metrik accuracy, precision, recall, dan F1-score. Seluruh metrik dihitung berdasarkan *confusion matrix* yang memuat nilai True Positive, True Negative, False Positive, dan False Negative.

III. HASIL DAN PEMBAHASAN

Pada bab ini akan disajikan hasil eksperimen terkait perbandingan dua metode ekstraksi fitur, yaitu Bag of Words (BoW) dan TF-IDF, dalam penerapan algoritma Naïve Bayes untuk klasifikasi sentimen. Pembahasan mencakup tahapan pengujian, penggunaan metrik evaluasi untuk menilai kinerja model, dan menganalisis perbedaan performa pada masing-masing representasi fitur.

A. *Dataset*

Pada penelitian ini, dataset yang digunakan berupa ulasan produk Tokopedia yang telah dijelaskan pada bab sebelumnya. Dalam proses pemodelan sentimen, penelitian ini memanfaatkan dua atribut utama, yaitu *Customer Review* dan *Customer Rating*. Kolom *Customer Review* akan digunakan sebagai data teks yang selanjutnya akan diproses melalui tahap *preprocessing* dan ekstraksi fitur, sedangkan kolom *Customer Rating* dimanfaatkan sebagai dasar dalam proses pelabelan sentimen.

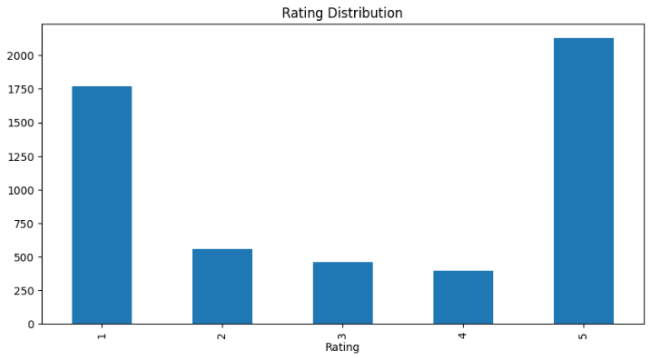
Pelabelan sentimen dilakukan dengan mengonversi nilai *Customer Rating* ke dalam dua kelas sentimen, yaitu positif dan negatif, dimana rating 4–5 dikategorikan sebagai sentimen positif, sedangkan rating 1–3 sebagai sentimen negatif. Pemilihan kedua kolom ini didasarkan pada keterkaitannya secara langsung dalam merepresentasikan opini pelanggan dan evaluasi kepuasan yang diberikan, sehingga dinilai paling relevan untuk tujuan klasifikasi sentimen. Untuk memberikan gambaran yang lebih jelas mengenai struktur data yang digunakan pada tahap selanjutnya, potongan dataset yang memuat kedua kolom tersebut ditampilkan pada Gambar 2.

	Customer Rating	Customer Review
0	5	Alhamdulillah berfungsi dengan baik. Packaging...
1	5	barang bagus dan respon cepat, harga bersaing ...
2	5	barang bagus, berfungsi dengan baik, seler ram...
3	5	bagus sesuai harapan penjual nya juga ramah. t...
4	5	Barang Bagus, pengemasan Aman, dapat Berfungsi...

Gambar 2. Struktur Dataset yang Digunakan

B. *Exploratory Data Analysis (EDA)*

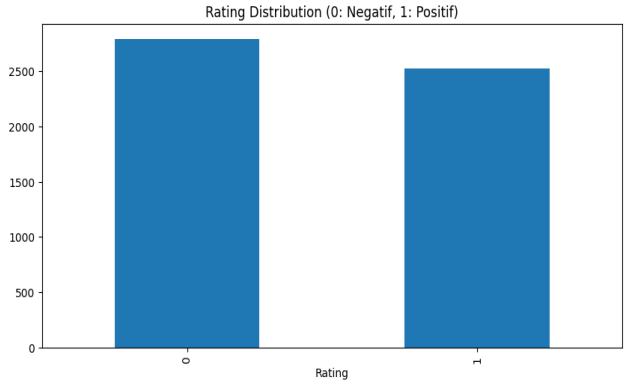
Pada tahap Exploratory Data Analysis (EDA), dilakukan pemeriksaan awal terhadap dataset untuk memahami karakteristik data sebelum memasuki tahap *preprocessing* dan pemodelan. Analisis diawali dengan mengevaluasi struktur data serta memastikan bahwa kolom yang digunakan dalam penelitian, yaitu *Customer Review* dan *Customer Rating*, memiliki format yang sesuai untuk proses pengolahan lebih lanjut. Selain itu, dilakukan pengecekan terhadap keberadaan data duplikat dan nilai kosong. Berdasarkan hasil analisis, ditemukan sebanyak 88 baris data duplikat, sedangkan seluruh kolom pada dataset tidak memiliki nilai kosong (*null*). Keberadaan data duplikat tersebut penting untuk diperhatikan karena dapat memengaruhi performa model dan menghasilkan bias pada proses pelatihan. Oleh karena itu, baris data duplikat dihapus agar dataset menjadi lebih bersih dan representatif, sehingga analisis dan pemodelan yang dilakukan dapat memberikan hasil yang lebih akurat. Selain pemeriksaan kualitas data, tahap EDA juga mengungkap adanya ketidakseimbangan jumlah data antar kelas sentimen yang berpotensi memengaruhi kinerja model klasifikasi apabila tidak ditangani dengan tepat. Kondisi ini dapat menyebabkan model cenderung bias terhadap kelas dengan jumlah data yang lebih dominan dan kurang optimal dalam mengenali kelas minoritas.



Gambar 3. Distribusi Rating

Tahap EDA juga mencakup analisis distribusi rating yang divisualisasikan pada Gambar 3. Grafik tersebut menunjukkan bahwa dataset memiliki lima kategori rating dengan distribusi yang tidak merata. Rating 5 menjadi kategori dengan jumlah terbanyak, yaitu sekitar 2.129 ulasan, diikuti oleh rating 1 dengan jumlah sekitar 1.772 ulasan. Sementara itu, rating 2 memiliki sekitar 557 ulasan, rating 3 sekitar 460 ulasan, dan rating 4 merupakan kategori dengan jumlah paling sedikit, yaitu sekitar 394 ulasan. Pola ini memperlihatkan adanya class imbalance, di mana sebagian besar ulasan cenderung sangat positif atau sangat negatif.

Informasi persebaran rating ini dijadikan dasar dalam proses pembentukan label sentimen. Dalam penelitian ini, rating 1–3 diklasifikasikan sebagai sentimen negatif, sedangkan rating 4–5 diklasifikasikan sebagai sentimen positif. Pengelompokan ini dilakukan karena rating 1–3 secara umum merepresentasikan ketidakpuasan pelanggan, sedangkan rating 4–5 menunjukkan pengalaman yang cenderung memuaskan. Dengan menggunakan aturan tersebut, diperoleh total 2.789 data Negatif dan 2.523 data Positif seperti yang terlihat pada Gambar 4. Dengan demikian, tahap EDA tidak hanya memberikan gambaran awal mengenai kondisi data, tetapi juga membantu memastikan bahwa proses pengkategorian sentimen dilakukan berdasarkan pola nyata yang terdapat pada dataset.



Gambar 4. Distribusi Rating Positif dan Negatif

C. *Preprocessing Data*

Pada tahap preprocessing, dilakukan serangkaian proses pembersihan dan transformasi teks untuk memastikan bahwa data ulasan pelanggan berada dalam bentuk yang siap digunakan pada tahap pemodelan. Proses preprocessing ini penting karena kualitas data teks sangat mempengaruhi performa model analisis sentimen. Secara umum, tahap preprocessing yang diterapkan dalam penelitian ini meliputi *text cleaning* (pembersihan teks), *case folding*, *tokenizing*, *stopword removal*, dan *stemming* menggunakan library dari sastrawi. Contoh hasil penerapan setiap tahapan preprocessing ditunjukkan pada Tabel 1.

TABEL I
HASIL PREPROCESSING

Tahapan	Tweet
Original Text	Pesanan Saya TIDAK SAMPAI KE ALAMAT!!! SAYA MAU UANG SAYA KEMBALI!! SAYA BELUM MENERIMA PESANAN SAYA
Text Cleaning	Pesanan Saya TIDAK SAMPAI KE ALAMAT SAYA MAU UANG SAYA KEMBALI SAYA BELUM MENERIMA PESANAN SAYA
Case Folding	pesanan saya tidak sampai ke alamat saya mau uang saya kembali saya belum menerima pesanan saya
Tokenizing	[pesanan, saya, tidak, sampai, ke, alamat, saya, mau, uang, saya, kembali, saya, belum, menerima, pesanan, saya]
Stopword Removal	[pesanan, tidak, alamat, mau, uang, menerima, pesanan]
Stemming	pesan tidak alamat mau uang terima pesan

Dari hasil tahapan preprocessing tersebut, terlihat bahwa setiap proses memberikan kontribusi dalam menyederhanakan teks agar lebih mudah diproses oleh model analisis sentimen. Tahap cleaning berhasil menghilangkan karakter yang tidak diperlukan seperti tanda seru, sehingga teks menjadi lebih bersih. Pada tahap case folding, seluruh huruf diubah menjadi huruf kecil untuk menjaga konsistensi dan mencegah duplikasi makna antar kata yang sama. Proses tokenizing kemudian memecah kalimat menjadi unit-unit kata sehingga dapat diproses secara individual.

Selanjutnya, stopword removal menghapus kata-kata umum yang tidak memiliki kontribusi signifikan terhadap penentuan sentimen, seperti “saya”, “ke”, dan “belum”. Namun, kata “tidak” tetap dipertahankan karena memiliki peran penting sebagai penanda sentimen negatif. Tahap terakhir yaitu stemming menggunakan Sastrawi mengubah kata menjadi bentuk dasarnya, seperti “menerima” menjadi “terima” dan “pesanan” menjadi “pesan”. Hasil akhir ini menghasilkan representasi teks yang lebih ringkas dan padat, namun tetap mempertahankan informasi penting yang relevan untuk proses klasifikasi sentimen pada tahap pemodelan. Dengan demikian, model yang dibangun dapat bekerja lebih efisien dalam mempelajari pola sentimen dari data teks yang telah direpresentasikan.

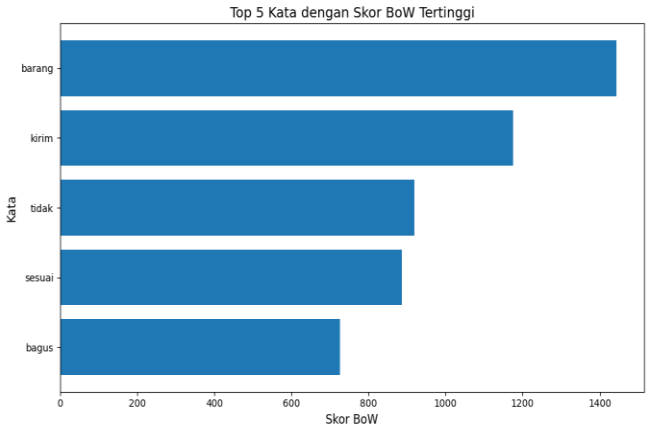
D. *Ekstraksi Fitur*

Pada penelitian ini, proses ekstraksi fitur dilakukan dengan dua pendekatan utama, yaitu Bag-of-Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF). Kedua metode tersebut diterapkan pada teks yang telah melalui tahap preprocessing sehingga diperoleh representasi numerik yang dapat diolah oleh model klasifikasi. Sebelum tahap ekstraksi fitur dilakukan, dataset terlebih dahulu dibagi menjadi data latih dan data uji dengan proporsi 80% untuk data latih dan 20% untuk data uji. Pembagian data ini bertujuan untuk memastikan bahwa proses pelatihan dan evaluasi model dilakukan secara objektif untuk menghindari terjadinya kebocoran data. Selain itu, proses pembagian data dilakukan secara acak dengan pengaturan random state tertentu untuk menjaga konsistensi proses pengolahan data dan kestabilan hasil pengujian. Selain itu, penelitian ini menambahkan fitur tambahan berupa panjang ulasan dan jumlah tanda baca untuk memperkaya informasi yang diberikan kepada model. Seluruh proses ini dilakukan dengan pengaturan parameter tertentu yang dipilih berdasarkan pertimbangan efektivitas dan kesesuaian terhadap karakteristik data.

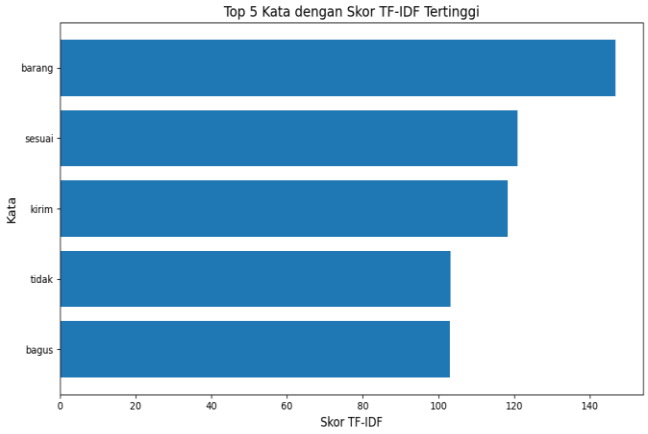
Pada metode BoW, proses vectorization dilakukan menggunakan *CountVectorizer* dengan konfigurasi max_df = 0.5, min_df = 2, dan rentang n-gram dari unigram hingga bigram. Penggunaan parameter max_df = 0.5 dimaksudkan untuk membuang kata-kata yang muncul terlalu sering pada lebih dari 50% dokumen. Kata yang terlalu sering muncul biasanya tidak memberikan informasi yang berguna dalam membedakan sentimen suatu ulasan, karena bersifat terlalu umum dan kurang memberikan kontribusi yang signifikan dalam membedakan sentimen. Sementara itu, min_df = 2 digunakan untuk mengabaikan kata yang hanya muncul satu kali pada keseluruhan corpus. Kata yang sangat jarang muncul cenderung bersifat noise dan tidak memberikan kontribusi yang berarti dalam pembelajaran model. Penggunaan n-gram dengan rentang (1,2) dipilih agar model tidak hanya mempelajari kata tunggal, tetapi juga frasa dua kata yang sering kali lebih merepresentasikan konteks sentimen.

Hasil transformasi BoW kemudian dikonversi menjadi DataFrame agar setiap fitur lebih mudah dianalisis. Data ini kemudian tidak langsung digunakan sebagai input model, tetapi terlebih dahulu digabungkan dengan dua fitur tambahan, yaitu *review length* (panjang ulasan) dan *punctuation* (jumlah tanda baca). Kedua fitur ini ditambahkan agar representasi yang diperoleh tidak hanya berbasis teks, tetapi juga mempertimbangkan karakteristik struktural dari ulasan yang kemudian digunakan sebagai masukan model untuk mengetahui efektivitas BoW dalam mendukung proses klasifikasi sentimen.

Selanjutnya, pendekatan TF-IDF diterapkan dengan konsep yang hampir serupa namun dengan perbedaan mendasar pada pembobotan kata. Pada metode ini digunakan *TfidfVectorizer* dengan konfigurasi yang sama seperti BoW, yaitu max_df = 0.5, min_df = 2, dan rentang n-gram (1,2). Setelah proses ekstraksi, hasil TF-IDF kemudian dikonversi menjadi sebuah DataFrame sehingga dapat digabungkan dengan fitur tambahan, yaitu *review length* dan *punctuation* sebagaimana dilakukan pada BoW. Penggabungan ini menghasilkan representasi fitur yang lebih lengkap karena mengombinasikan bobot TF-IDF yang bersifat semantik dengan karakteristik non-teks yang memberikan konteks tambahan. Representasi ini kemudian digunakan sebagai masukan model untuk mengetahui efektivitas TF-IDF dalam mendukung proses klasifikasi sentimen. Untuk memberikan gambaran mengenai distribusi kata yang paling berpengaruh dalam proses pembobotan, ditampilkan visualisasi hasil ekstraksi fitur menggunakan BoW pada Gambar 5 dan TF-IDF pada Gambar 6.



Gambar 5. Kata dengan Skor BoW Tertinggi

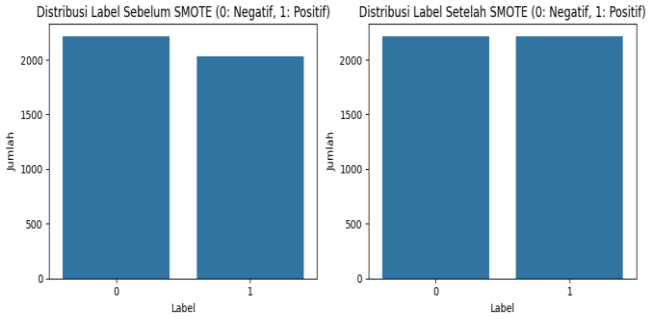


Gambar 6. Kata dengan Skor TF-IDF Tertinggi

Berdasarkan visualisasi BoW dan TF-IDF, terlihat bahwa kata "*barang*", "*kirim*", "*tidak*", "*sesuai*", dan "*bagus*" muncul sebagai kata yang paling dominan dalam ulasan, baik dari sisi frekuensi kemunculan maupun bobot kepentingannya. Pada grafik BoW, kata "*barang*" menjadi kata dengan kemunculan tertinggi, diikuti "*kirim*" dan "*tidak*", yang menunjukkan bahwa topik mengenai barang dan proses pengiriman menjadi fokus utama pengalaman pengguna. Sementara itu, pada grafik TF-IDF, kata "*barang*" kembali memiliki bobot tertinggi, tetapi diikuti oleh "*sesuai*" dan "*kirim*", menunjukkan bahwa kata-kata tersebut dianggap lebih informatif dan berperan penting dalam membedakan isi ulasan. Secara keseluruhan, kedua grafik tersebut menunjukkan konsistensi kata-kata yang sering muncul dalam ulasan, dengan masing-masing metode memberikan perspektif yang berbeda, BoW dari sisi frekuensi kemunculan, dan TF-IDF dari sisi tingkat kepentingan relatif kata dalam dokumen.

E. Imbalance Data (SMOTE)

Pada tahap awal analisis terlihat bahwa distribusi sentimen pada dataset tidak seimbang, dengan selisih jumlah yang cukup jelas antara masing-masing kelas. Untuk mengatasi ketidakseimbangan ini, digunakan teknik SMOTE pada data latih. Sebelum SMOTE diterapkan, kelas 0 (negatif) memiliki 2.216 data, sedangkan kelas 1 (positif) hanya berjumlah 2.033 data, sehingga berpotensi membuat model lebih condong pada kelas mayoritas. Melalui proses SMOTE, sampel baru secara sintetis ditambahkan pada kelas minoritas (kelas 1) dengan menghasilkan data yang memiliki pola serupa dengan data positif yang telah ada. Setelah proses ini dilakukan, jumlah sampel kelas positif meningkat hingga menyamai kelas negatif, sehingga distribusinya menjadi seimbang, yaitu masing-masing 2.216 data. Perubahan ini terlihat pada Gambar 7, di mana grafik sebelum SMOTE menunjukkan ketidakseimbangan label, sementara grafik setelah SMOTE memperlihatkan kedua kelas telah memiliki jumlah yang sama. Dengan distribusi yang lebih seimbang, model diharapkan dapat belajar secara lebih objektif tanpa bias terhadap salah satu kelas.



Gambar 7. Visualisasi Sebelum dan Sesudah SMOTE

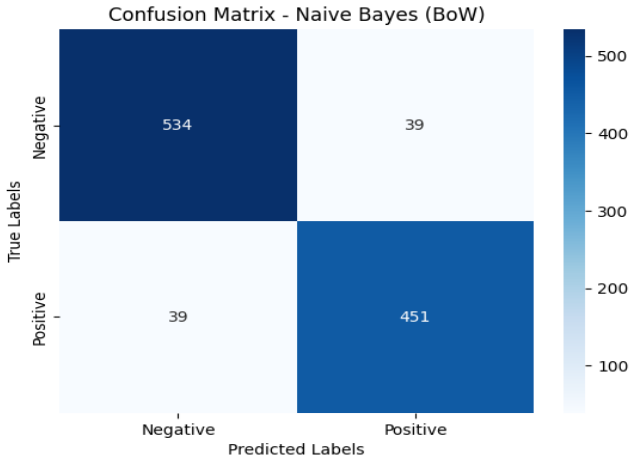
F. Klasifikasi dengan Naïve Bayes

Setelah data melalui tahap preprocessing, dataset terlebih dahulu dibagi menjadi 80% data latih dan 20% data uji. Pembagian data ini bertujuan agar model dapat dilatih menggunakan sebagian besar data dan diuji menggunakan data yang tidak pernah dilihat sebelumnya. Selanjutnya, proses ekstraksi fitur dilakukan pada data teks menggunakan metode Bag-of-Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF) untuk menghasilkan representasi numerik yang dapat diproses oleh model. Data latih yang telah melalui proses ekstraksi fitur kemudian diseimbangkan menggunakan metode SMOTE yang berguna untuk mengatasi ketidakseimbangan kelas pada data pelatihan. Setelah itu, proses klasifikasi dilakukan menggunakan algoritma Naïve Bayes, di mana model dilatih menggunakan data latih yang telah diseimbangkan dan selanjutnya digunakan untuk melakukan prediksi terhadap data uji.

G. Evaluasi Model

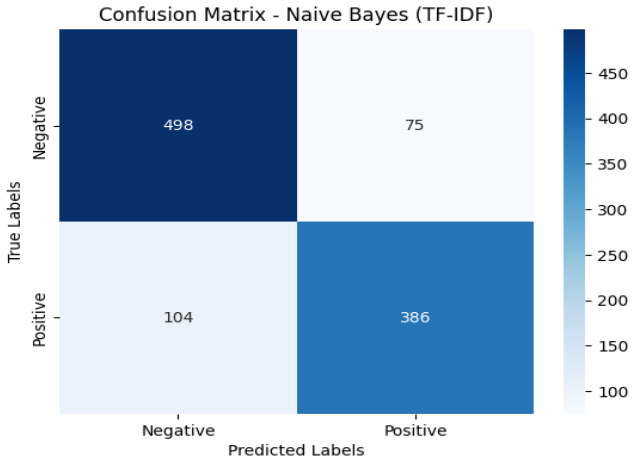
Pada tahap evaluasi, performa model Naïve Bayes diuji menggunakan dua pendekatan ekstraksi fitur, yaitu BoW dan TF-IDF. Evaluasi dilakukan menggunakan metrik accuracy, precision, recall, dan f1-score untuk mengukur kemampuan model dalam mengklasifikasikan sentimen ulasan ke dalam kelas Negative dan Positive. Proses pelatihan dan pengujian dilakukan setelah data melalui tahap preprocessing serta penyeimbangan kelas menggunakan SMOTE. Hasil evaluasi menunjukkan bahwa perbedaan metode representasi fitur memberikan pengaruh terhadap performa klasifikasi yang dihasilkan.

1) *Evaluasi Model Naïve Bayes dengan BoW:* Berdasarkan confusion matrix yang ditampilkan pada Gambar 8, model Naïve Bayes dengan representasi fitur BoW menunjukkan performa yang sangat baik. Model berhasil mengklasifikasikan 534 ulasan negatif dan 451 ulasan positif dengan benar. Sementara itu, kesalahan klasifikasi relatif rendah, dengan 39 ulasan negatif yang diprediksi sebagai positif dan 39 ulasan positif yang diprediksi sebagai negatif. Hasil ini berkontribusi pada nilai accuracy sebesar 93%, serta nilai precision, recall, dan f1-score yang relatif seimbang pada kedua kelas. Hal tersebut menunjukkan bahwa pendekatan BoW mampu menangkap pola frekuensi kata secara efektif pada data ulasan yang digunakan, sehingga model dapat membedakan sentimen negatif dan positif dengan baik.



Gambar 8. Confusion Matrix BoW

2) *Evaluasi Model Naïve Bayes dengan TF-IDF:* Untuk Confusion matrix dengan pendekatan TF-IDF yang ditampilkan pada Gambar 9. Dari hasil tersebut, terlihat bahwa model berhasil mengklasifikasikan 498 ulasan negatif dan 386 ulasan positif dengan benar. Namun, jumlah kesalahan klasifikasi lebih tinggi dibandingkan dengan BoW, khususnya pada kelas positif, di mana terdapat 104 ulasan positif yang salah diprediksi sebagai negatif. Kondisi ini menyebabkan nilai accuracy model menurun menjadi 83%, dengan penurunan yang cukup signifikan pada nilai recall kelas Positive. Hal ini mengindikasikan bahwa meskipun TF-IDF memberikan bobot kata yang lebih informatif secara teoritis, pendekatan ini kurang optimal dalam menangkap karakteristik sentimen positif pada dataset ulasan yang digunakan dalam penelitian ini.



Gambar 9. Confusion Matrix TF-IDF

H. Perbandingan Ekstraksi Fitur BoW dan TF-IDF

Berdasarkan hasil evaluasi yang disajikan pada Tabel II, terlihat adanya perbedaan performa antara model Naïve Bayes yang menggunakan ekstraksi fitur Bag-of-Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF). Secara keseluruhan, pendekatan BoW menunjukkan performa yang lebih unggul dibandingkan dengan TF-IDF dalam mengklasifikasikan sentimen ulasan pelanggan.

Model dengan ekstraksi fitur BoW berhasil mencapai akurasi sebesar 93%, yang menunjukkan kemampuan klasifikasi yang sangat baik. Pada kelas positif, BoW memperoleh nilai presisi, recall, dan f1-score masing-masing sebesar 92%, yang menandakan bahwa model mampu mengidentifikasi ulasan positif secara konsisten dan seimbang. Sementara itu, pada kelas negatif, performa BoW bahkan sedikit lebih tinggi dengan nilai presisi, recall, dan f1-score sebesar 93%, yang menunjukkan bahwa model sangat efektif dalam mengenali sentimen negatif pada data ulasan.

Sebaliknya, pendekatan TF-IDF menghasilkan akurasi sebesar 83%, yang lebih rendah dibandingkan dengan BoW. Pada kelas positif, nilai presisi sebesar 84% menunjukkan bahwa sebagian besar prediksi positif sudah tepat, namun nilai recall yang menurun menjadi 79% mengindikasikan bahwa masih terdapat cukup banyak ulasan positif yang salah diklasifikasikan sebagai negatif. Hal ini berdampak pada nilai f1-score sebesar 81%, yang mencerminkan ketidakseimbangan antara presisi dan recall pada kelas positif. Pada kelas negatif, TF-IDF menunjukkan performa yang relatif lebih baik dengan nilai presisi sebesar 83%, recall sebesar 87%, dan f1-score sebesar 85%. Meskipun demikian, secara keseluruhan, performa TF-IDF masih berada di bawah BoW, terutama dalam mendeteksi ulasan dengan sentimen positif.

TABEL II
HASIL EVALUASI EKSTRAKSI FITUR BOW DAN TF-IDF

Ekstraksi Fitur	Akurasi	Kelas	Presisi	Recall	F1-score
BoW	93%	Positif	92%	92%	92%
		Negatif	93%	93%	93%
TF-IDF	83%	Positif	84%	79%	81%
		Negatif	83%	87%	85%

Dari hasil perbandingan tersebut dapat disimpulkan bahwa Performa BoW yang lebih unggul dibandingkan TF-IDF pada penelitian ini dapat dipengaruhi oleh karakteristik dataset ulasan Tokopedia yang didominasi oleh kata-kata sentimen yang sering muncul dan memiliki makna polar yang jelas, seperti *"bagus"*, *"tidak"*, *"sesuai"*, dan *"cepat"*. Pendekatan BoW yang mempertahankan frekuensi kemunculan kata mampu merepresentasikan kontribusi kata-kata tersebut secara langsung dalam proses klasifikasi. Sebaliknya, pada metode TF-IDF, kata-kata yang sering muncul justru mengalami penurunan bobot sehingga pengaruhnya terhadap proses pengambilan keputusan model menjadi lebih kecil. Selain itu, penggunaan algoritma Naïve Bayes yang berbasis probabilistik lebih selaras dengan representasi fitur BoW dibandingkan bobot kontinu pada TF-IDF, sehingga menghasilkan performa klasifikasi yang lebih stabil dan konsisten pada dataset yang digunakan. Hal ini menunjukkan bahwa representasi berbasis frekuensi kata sederhana lebih efektif dalam menangkap pola sentimen pada dataset ulasan yang digunakan.

Penelitian ini memiliki keterbatasan pada penggunaan dataset yang hanya berasal dari ulasan Tokopedia sehingga hasilnya belum tentu dapat digeneralisasikan ke platform lain. Pelabelan sentimen didasarkan pada *Customer Rating* dengan dua kelas, yaitu positif dan negatif, yang berpotensi tidak sepenuhnya merepresentasikan isi teks ulasan. Selain itu, keragaman bahasa ulasan serta penggunaan metode ekstraksi fitur berbasis statistik (BoW dan TF-IDF) dapat membatasi kemampuan model dalam menangkap konteks semantik secara mendalam.

IV. KESIMPULAN

Pada Penelitian ini yang bertujuan untuk membandingkan performa algoritma Naïve Bayes dengan dua metode ekstraksi fitur, yaitu Bag of Words (BoW) dan Term Frequency–Inverse Document Frequency (TF-IDF), dalam klasifikasi sentimen pada ulasan produk Tokopedia. Berdasarkan hasil evaluasi yang telah dilakukan, seluruh rangkaian proses mulai dari preprocessing, ekstraksi fitur, penyeimbangan data dengan SMOTE, hingga pelatihan model menunjukkan bahwa pemilihan ekstraksi fitur memiliki pengaruh signifikan terhadap hasil klasifikasi. Ekstraksi fitur dengan BoW memberikan performa terbaik dengan akurasi mencapai 93%, serta nilai precision, recall, dan f1-score yang seimbang pada kedua kelas sentimen. Hasil ini menunjukkan bahwa BoW lebih efektif dalam menangkap pola frekuensi kata yang dominan pada ulasan pelanggan dibandingkan TF-IDF, yang dalam penelitian ini hanya menghasilkan akurasi sebesar 83% dan menunjukkan kecenderungan salah klasifikasi pada ulasan positif. Dengan demikian, tujuan penelitian untuk menentukan metode representasi fitur yang paling sesuai bagi algoritma Naïve Bayes telah tercapai, dan dapat disimpulkan bahwa metode BoW merupakan pilihan yang lebih optimal untuk analisis sentimen pada dataset ulasan Tokopedia.

Meskipun demikian, penelitian ini masih memiliki peluang untuk dikembangkan lebih lanjut. Penggunaan algoritma pembanding seperti Support Vector Machine, Logistic Regression, atau model berbasis deep learning dapat dievaluasi untuk memperoleh pemahaman yang lebih komprehensif mengenai performa klasifikasi pada dataset serupa. Selain itu, eksplorasi teknik representasi fitur modern seperti Word2Vec, FastText, atau kombinasi TF-IDF dengan n-gram yang lebih kompleks berpotensi meningkatkan pemahaman konteks kata dalam ulasan. Penelitian mendatang juga dapat mempertimbangkan pengembangan klasifikasi

multikategori dengan menambahkan kategori sentimen netral atau melakukan analisis sentimen berdasarkan aspek tertentu untuk mendapatkan pemahaman yang lebih jelas mengenai opini pelanggan. Dengan pengembangan tersebut, hasil penelitian diharapkan dapat memberikan manfaat yang lebih baik dalam mendukung proses pengambilan keputusan pada platform e-commerce.

DAFTAR PUSTAKA

[1] Y. Putri, R. Kusumadewi, and E. Saefulloh, "Pengaruh Kredibilitas Influencer dan Brand Awarness Terhadap Minat Pembelian di Tokopedia (studi Pada Pelanggan Tokopedia Yang Bertransaksi Melalui Bank Syariah Indonesia)," *Entrep. J. Bisnis Manaj. Dan Kewirausahaan*, vol. 4, pp. 205–225, May 2023, doi: 10.31949/entrepreneur.v4i2.5651.

[2] M. Idris, A. Rifai, and K. D. Tania, "Sentiment Analysis of Tokopedia App Reviews using Machine Learning and Word Embeddings," *Sink. J. Dan Penelit. Tek. Inform.*, vol. 9, no. 1, pp. 210–219, Jan. 2025, doi: 10.33395/sinkron.v9i1.14278.

[3] D. G. Nugroho, Y. H. Chrisnanto, and A. Wahana, "Analisis Sentimen Pada Jasa Ojek Online Menggunakan Metode Naïve Bayes," *Pros. Sains Nas. Dan Teknol.*, vol. 1, no. 1, Sept. 2016, doi: 10.36499/psnst.v1i1.1526.

[4] "Analisis Sentimen Berdasarkan Opini Pengguna pada Media Twitter Terhadap BPJS Menggunakan Metode Lexicon Based dan Naïve Bayes Classifier," *J. Ilm. Komputasi*, vol. 20, no. 1, Mar. 2021, doi: 10.32409/jikstik.20.1.401.

[5] A. I. Tanggraeni and M. N. N. Sitokdana, "Analisis Sentimen Aplikasi E-Government pada Google Play Menggunakan Algoritma Naïve Bayes | JATISI," June 2022, Accessed: Dec. 16, 2025. [Online]. Available: https://jurnal.mdp.ac.id/index.php/jatisi/article/view/1835

[6] O. Bellar, A. Baina, and M. Ballafkih, "Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers," *Mathematics*, vol. 12, no. 15, p. 2403, Jan. 2024, doi: 10.3390/math12152403.

[7] V. Gooljar, T. Issa, S. Hardin-Ramanan, and B. Abu-Salih, "Sentiment-based predictive models for online purchases in the era of marketing 5.0: a systematic review," *J. Big Data*, vol. 11, no. 1, p. 107, Aug. 2024, doi: 10.1186/s40537-024-00947-0.

[8] N. Nasrabadi, H. Wicaksono, and O. Fatahi Valilai, "Shopping marketplace analysis based on customer insights using social media analytics," *MethodsX*, vol. 13, p. 102868, Dec. 2024, doi: 10.1016/j.mex.2024.102868.

[9] "Pengaruh Feature Selection Dan Feature Extraction Dalam Peningkatan Akurasi Klasifikasi Kebakaran Hutan | Armaya | Jurnal Teknologi Informasi." Accessed: Dec. 16, 2025. [Online]. Available: https://ejournal.akakom.ac.id/index.php/JuTI/article/view/1039/pdf

[10] D. Setiawan, N. Umar, and M. A. Nur, "Feature Extraction Optimization to Improve Naïve Bayes Accuracy in Sentiment Analysis of Bulukumba Tourism Objects," *Sist. J. Sist. Inf.*, vol. 13, no. 5, pp. 2209–2221, Sept. 2024, doi: 10.32520/stmsi.v13i5.4580.

[11] A. Firdaus, A. Hadiana, and A. Ningsih, "Klasifikasi Sentimen pada Aplikasi Shopee Menggunakan Fitur Bag of Word dan Algoritma Random Forest," *Ranah Res. J. Multidiscip. Res. Dev.*, vol. 6, pp. 1678–1683, July 2024, doi: 10.38035/rrj.v6i5.994.

[12] C. Rosanti, F. A. Artanto, and R. E. Saputra, "Regresi Dengan Ekstrasi Fitur Neural Bag of Words Pada Analisis Sentimen Pengguna Aplikasi Bank Digital Syariah," *JIPI J. Ilm. Penelit. Dan Pembelajaran Inform.*, vol. 10, no. 3, pp. 2418–2425, Aug. 2025, doi: 10.29100/jipi.v10i3.6508.

[13] R. Al Rasyid and D. H. U. Ningsih, "Penerapan Algoritma TF-IDF dan Cosine Similarity untuk Query Pencarian Pada Dataset Destinasi Wisata," *J. JTIK J. Teknol. Inf. Dan Komun.*, vol. 8, no. 1, pp. 170–178, Jan. 2024, doi: 10.35870/jtik.v8i1.1416.

[14] M. Adha, F. Freddy, and F. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *J. Inf. Technol.*, vol. 2, pp. 36–40, Mar. 2022, doi: 10.46229/jifotech.v2i1.330.

[15] S. M. P. Tyas, B. S. Rintyarna, and W. Suharso, "The Impact of Feature Extraction to Naïve Bayes Based Sentiment Analysis on Review Dataset of Indihome Services," *Digit. Zone J. Teknol. Inf. Dan Komun.*, vol. 13, no. 1, pp. 1–10, Apr. 2022, doi: 10.31849/digitalzone.v13i1.9158.

[16] "Analisis Sentimen Pelanggan Tokopedia Menggunakan Metode Naïve Bayes Classifier | Jurnal Minfo Polgan." Accessed: Dec. 16, 2025. [Online]. Available: https://jurnal.polgan.ac.id/index.php/jmp/article/view/11640

[17] B. Darmawan and A. D. Laksito, "Analisis Perbandingan Ekstraksi Fitur Teks pada Sentimen Analisis Kenaikan Harga BBM", Accessed: Dec. 16, 2025. [Online]. Available: https://core.ac.uk/works/150334040/

[18] R. Suryanti and P. Prasetyaningrum, "Perbandingan Metode TF-IDF dan Bag of Words dalam Analisis Sentimen Diet Kopi Americano di Media Sosial Twitter Menggunakan Naïve Bayes," *Build. Inform. Technol. Sci. BITS*, vol. 7, no. 1, pp. 104–115, June 2025, doi: 10.47065/bits.v7i1.7244.

[19] A. Ernawati, A. O. Sari, S. N. Sofyan, M. Iqbal, and R. F. W. Wijaya, "Implementasi Algoritma Naïve Bayes dalam Menganalisis Sentimen Review Pengguna Tokopedia pada Produk Kesehatan," *Bull. Inf. Technol. BIT*, vol. 4, no. 4, pp. 533–543, Dec. 2023, doi: 10.47065/bit.v4i4.1090.

[20] A. Gerliandeva, Y. Chrisnanto, and H. Ashaury, "Optimasi Klasifikasi Sentimen pada Komentar Online menggunakan Multinomial Naïve Bayes dan Ekstraksi Fitur TF-IDF serta N-gramsOptimization of Sentiment Classification on Online Comments using Multinomial Naïve Bayes and TF-IDF Feature Extraction and N-grams," *J. Pekommas*, vol. 9, pp. 260–272, Dec. 2024, doi: 10.56873/jpkm.v9i2.5585.

[21] M. A. Palomino and F. Aider, "Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis," *Appl. Sci.*, vol. 12, no. 17, p. 8765, Jan. 2022, doi: 10.3390/app12178765.

[22] R. Wati, S. Ernawati, and H. Rachmi, "Pembobotan TF-IDF Menggunakan Naïve Bayes pada Sentimen Masyarakat Mengenai Isu Kenaikan BIPIH," *J. Manaj. Inform. JAMIKA*, vol. 13, no. 1, pp. 84–93, Apr. 2023, doi: 10.34010/jamika.v13i1.9424.

[23] M. Kamruzzaman and G. Kim, "Efficient Sentiment Analysis: A Resource-Aware Evaluation of Feature Extraction Techniques, Ensembling, and Deep Learning Models," in *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, L.-W. Ku and C.-T. Li, Eds., Bali, Indonesia: Association for Computational Linguistics, Nov. 2023, pp. 9–20. doi: 10.18653/v1/2023.socialnlp-1.2.

[24] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "Smote: Potensi Dan Kekurangannya Pada Survei," *E-J. Mat.*, vol. 10, no. 4, pp. 235–240, Nov. 2021, doi: 10.24843/MTK.2021.v10.i04.p348.

[25] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput.*, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.

[26] A. Nurhopipah and C. Magnolia, "Perbandingan Metode Resampling Pada Imbalanced Dataset Untuk Klasifikasi Komentar Program Mbkm," *J. Publ. Ilmu Komput. Dan Multimed.*, vol. 2, no. 1, pp. 9–22, Jan. 2023, doi: 10.55606/jupikom.v2i1.862.

[27] H. Susana, "Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet," *J. Ris. Sist. Inf. Dan Teknol. Inf. JURSISTEKNI*, vol. 4, no. 1, pp. 1–8, Feb. 2022, doi: 10.52005/jursistekni.v4i1.96.

[28] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, Mar. 2024, doi: 10.1038/s41598-024-56706-x.