

REKOMENDASI PRODUK *E-COMMERCE* BERBASIS KLASIFIKASI MENGGUNAKAN ALGORITMA *MACHINE LEARNING*

Intensif Nehe¹, Hotma Pangaribuan²

¹Program Studi Teknik Informatika, Universitas Putera Batam

²Program Studi Teknik Informatika, Universitas Putera Batam

email: pb210210072@upbatam.ac.id

ABSTRACT

The growth of e-commerce is driving an increase in the number of product reviews, but not all reviews are informative and relevant. This study aims to build a Tokopedia review classification system using the Naive Bayes algorithm to filter relevant reviews. The dataset used totaled 40,607 product review lines. The preprocessing process includes cleaning, tokenization, stopword removal, and feature transformation using TF-IDF with 5,000 feature words. Labels are determined based on ratings, where a rating of 4-5 is considered relevant. The Multinomial Naive Bayes model yielded an accuracy of 93.71%, a precision of 0.94, a recall of 0.99 for the relevant class, and an F1-score of 0.96. Although performance in irrelevant classes is still low, this model is effective in supporting product recommendations.

Keywords: Naive Bayes, Product Reviews, Tokopedia, Classification

PENDAHULUAN

Pertumbuhan pengguna internet yang sedemikian pesatnya membuat internet salah satu media yang efektif bagi perusahaan maupun perseorangan untuk memperkenalkan dan menjual barang atau jasa kepada konsumen khusus nya di Indonesia. *E-commerce* adalah salah satu situs web yang menyediakan atau dapat melakukan transaksi online, dan juga merupakan suatu cara berbelanja atau berdagang secara online yang memanfaatkan fasilitas internet dimana yang menyediakan layanan “*get and deliver*” [1]. Berdasarkan analisis kementerian perdagangan republik Indonesia Jumlah pengguna *e-commerce* di Indonesia pada tahun 2024 sebanyak 65,65 juta pengguna, dan di tahun 2025 meningkat sebesar 11,2% menjadi 73,06 juta pengguna.

Menurut [2]. popularitas *e-commerce* menjadi tantangan baru untuk memberikan pengalaman berbelanja seperti keamanan data saat melakukan transaksi, kepercayaan terhadap produk yang tidak sesuai dengan ekspektasi, ketergantungan pada koneksi internet dan banyak platform *e-commerce* website atau aplikasi yang lambat, navigasi yang membingungkan atau kurangnya informasi yang detail pada produk dapat menghambat pengalaman berbelanja. Tokopedia salah satu *platform* yang mudah dipahami untuk berbelanja dengan fitur-fitur yang mudah di pahami, keaaman privasi yang kuat dan informasi produk sangat jelas pada deskripsi produk.

Rekomendasi adalah suatu yang menyarankan untuk menemukan produk atau item (disuatu kasus *e-commerce*) terhadap pengguna [3]. Salah satu tantangan utama dalam rekomendasi

produk *e-commerce* adalah meningkatnya jumlah data dan kompleksitas preferensi pengguna, yang membuat sulit untuk memberikan rekomendasi yang relevan. Masalah dalam merekomendasikan produk ketika produk belum banyak ulasan, rating dan data transaksi lainnya sehingga pengguna baru ragu dengan produk yang tersebut. Maka dalam mengatasi tantangan tersebut, peneliti menggunakan algoritma *machine learning* dapat membantu mengelompokkan pengguna dan produk secara lebih efektif, sehingga meningkatkan akurasi rekomendasi.

. *Machine learning* adalah Teknik pendekatan dari *Artificial Intelligent* (AI) yang meniru perilaku manusia dan menggantikan peran aktivitas manusia untuk menyelesaikan masalah [4]. pada penelitian ini mengusulkan algoritma *Naïve Bayes* untuk menyelesaikan masalah klasifikasi produk Tokopedia dengan tujuan untuk memprediksi produk yang relevan dapat di rekomendasikan kepada pengguna. *Naive Bayes* adalah algoritma klasifikasi berbasis probabilitas dan statistik yang sangat efektif. Algoritma *Naive Bayes* digunakan untuk melakukan klasifikasi pada data dengan mengestimasi probabilitas dari kelas tertentu berdasarkan fitur-fitur atau atribut yang berada dalam data [5]. Hasil dari klasifikasi akan di lakukan pengolahan untuk mendapatkan rekomendasi produk *e-commerce* yang relevan.

KAJIAN TEORI

2.1 E-commerce

E-commerce singkatan dari *elektronik commerce*, yang diartikan sebagai sistem pembelian yang di lakukan secara elektronik atau media

elektronik. *E-commerce* juga adalah proses pembelian dan penjualan produk, jasa dan informasi melalui jaringan komputer atau jaringan internet [6]. Salah satu *platform e-commerce* yang berkembang pesat saat ini adalah Tokopedia. Tokopedia menjadi salah satu yang sukses dalam memanfaatkan potensi *platform e-commerce*. Dibantu dengan teknologi yang canggih dan strategi inovatif, Tokopedia memainkan peran yang signifikan dalam mendorong pertumbuhan ekonomi digital Indonesia, kehadiran Tokopedia ini tidak hanya untuk mempengaruhi cara konsumen berbelanja, tetapi juga memberikan dampak bagi UKM untuk bersaing dipasar global yang semakin maju saat ini [6]. Perilaku pengguna *e-commerce* yang sering melakukan pencarian produk, membeli untuk memenuhi kebutuhan nya [7].

2.2 Algoritmat Machine Learning

Machine learning atau pembelajaran mesin adalah bagian dari kecerdasan buatan/*Artificial Intelligence* (AI) yang di kembangkan sistem komputer dapat meniru perilaku manusia, dan dapat memproses Bahasa alami, pengenalan suara, persepsi visual, dan pengambilan Keputusan [8] Dalam rekomendasi produk *e-commerce* pada Tokopedia, kecerdasan buatan (AI) dapat dimanfaatkan untuk menemukan produk yang relevan berdasarkan hasil dari preferensi dan pengguna sebelumnya serta memberikan peningkatan pengalaman berbelanja secara daring terhadap pengguna baru.

2.2.1 Metode Naive Bayes

Naive Bayes adalah algoritma *machine learning* yang yang digunakan untuk klasifikasi. Metode ini didasarkan

pada teorema bayes, yang digunakan untuk menghitung probabilitas suatu kejadian yang berdasarkan informasi sebelumnya [9]. Menurut [10] Algoritma *Naïve Bayes* adalah salah satu pengklasifikasi statistik yang digunakan untuk memprediksi sebuah kelas. Pengklasifikasi dilakukan perhitungan nilai probabilitas dari setiap label kelas yang ada terhadap masukan yang diberikan label kelas diambil dari nilai probabilitas yang sangat besar. Algoritma *naive bayes* mudah digunakan karena memiliki perhitungan yang tidak panjang [11].

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

Keterangan:

- $P(H|X)$: probabilitas hipotesis H (misalnya, kelas tertentu) diberikan data X (*posterior probability*).
- $P(H)$: probabilitas awal dari hipotesis H sebelum melihat data (*prior probability*)
- $P(X|H)$: probabilitas data X diberikan hipotesis H (*likelihood*)
- $P(X)$: probabilitas data X (*evidence*)

2.2.2 Klasifikasi

Klasifikasi adalah proses yang mencari model atau fungsi yang dapat membedakan dan menjelaskan objek atau kelas-kelas data berdasarkan atribut atau fitur-fitur yang dimilikinya. Klasifikasi mencakup pengenalan tulisan tangan, analisis sentiment, deteksi penyakit melalui gambar medis, pemisahan spam pada email dan prediksi harga saham. Dalam klasifikasi ini menggunakan algoritma *naive bayes* untuk memprediksi produk *e-commerce* Tokopedia [12]. Pada penelitian ini *jupyter notebook*

sebagai software yang digunakan untuk melakukan analisis untuk memprediksi produk yang relevan.

2.2.3 Confusion matrix

Menurut [13] *Confusion matrix* bertujuan untuk mengetahui performa algoritma *naive bayes* dalam mengklasifikasi produk apakah relevan atau tidak relevan bagi pengguna. Matrix ini memberikan gambaran keseluruhan performa model dengan menampilkan TN, TP, FN dan FP pada konteks klasifikasi rekomendasi produk. Dari hasil *confusion matrix* maka akan mengevaluasi hasil akurasi, presisi, recall, dan F1-Score.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Presisi = \frac{TP}{TP + FP}$$

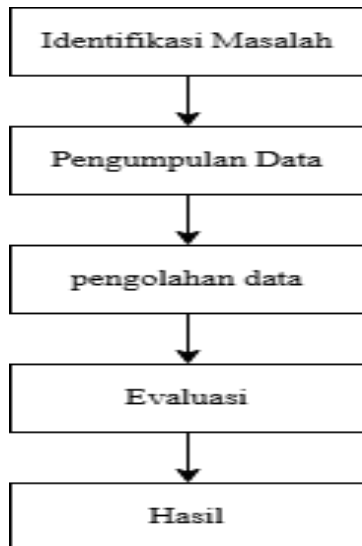
$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall}$$

METODE PENELITIAN

3.1 Desain Penelitian

Desain penelitian adalah rancangan atau strategi yang akan digunakan dalam penelitian. Desain penelitian ini menjelaskan alur kegiatan secara lengkap. Berikut tahap-tahap desain penelitian.



Gambar 1 Desain Penelitian

Sumber : Penelitian 2025

3.2 Metode Pengumpulan Data

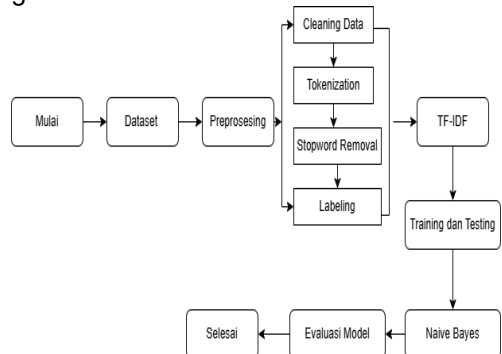
Data yang digunakan adalah ulasan produk Tokopedia. Yang di ambil dari *google kaggle* dengan jumlah data sebanyak 40893 ribu. Data di simpan dalam bentuk CSV, sebelum digunakan untuk tahap proses berikutnya. Data diimport dalam bentuk file *Excel* dengan format .xlsx. Data bisa diakses dengan mengklik linkn di bawah ini.

<https://www.kaggle.com/datasets/farhan99/tokopedia-product-reviews>

3.3 Metode Perancangan

Pada penelitian ini, metode perancangan klasifikasi dengan algoritma *naive bayes* baisanya mengacu pada tahapan sistematis dalam membangun model yang cukup pemilihan data, *preprocessing*, pemilihan fitur, pelatihan model, dan evaluasi [9]. Berikut alur algoritma *naive bayes* dengan

menggunakan *flowchart* seperti pada gambar.

Gambar 2 perencanaan Model
Sumber : Penelitian 2025

1. Pengumpulan data

Sumber data diambil dari situs *google kaggle*. Dimana data ini merupakan hasil ulasan produk tokopedia, isi dari data tersebut terdiri dari ulasan teks, rating, kategori, nama produk, id_produk, terjual, id_toko dan produk_url.

2. Preprocessing

pada penelitian ini, *Preprocessing* adalah proses mengambil data awal pada teks untuk mempersiapkan teks menjadi data yang dapat diolah. Terdapat tahap-tahap yang dilakukan pada preprocessing ini sebagai berikut:

a. cleaning data

proses *cleaning* data ini dilakukan penghapusan karakter seperti tanda baca, simbol, angka, url atau link dan emoji atau karakter khusus.

b. Tokenization

Proses ini memecahkan teks menjadi unit-unit kata sebagai dasar analisis.

c. Stopword removal

Proses ini menghapus kata yang sering muncul namun tidak



	text	rating
0	Barang sesuai pesanan dan cepat sampai	5
1	Barang bagus harga murah	5
2	Paket rapi...mantap....cepat....sampe ke tujuan	5
3	ya saya puas dgn barangnya	5
4	Responya luar biasa b mantap	5
...	...	...
40602	gak bisa ditutup.. ukuran gak pas..	1
40603	Penjual krg peduli dan brg jelek. Brg yg dikir...	1
40604	pelayanan kurang sigap, berhari2 dr nanya baik...	1
40605	barang yg dikirim tsk sesuai wm, kd memang dr...	1
40606	Pihak penjual memberikan respon yang buruk kep...	1

40607 rows x 2 columns

Gambar 4 Kolom Text dan Rating
Sumber : Data Penetian 2025

4.2 Hasil Preprocessing

Proses *preprocessing* data merupakan tahapan krusial dalam analisis teks untuk mengubah data mentah menjadi format yang bersih, terstruktur, dan siap untuk diolah oleh algoritma *naive bayes*

1. Cleaning

Tabel 1 Hasil Cleaning

Sebelum clean	Hasil
Barang sesuai pesanan dan cepat sampai	barang sesuai pesanan dan cepat sampai
Barang bagus harga murah	barang bagus harga murah
Paket rapi...mantap....cepat....sampe ke tujuan	paket rapimantapcepatsa mpe ke tujuan

Sumber : Data Penelitian 2025

2. Tokenization

Tabel 2 Hasil Tokenization

Sebelum token	Hasil
barang sesuai pesanan dan cepat sampai	[barang, sesuai, pesanan, dan, cepat, sampai]
barang bagus harga murah	[barang, bagus, harga, murah]

Sumber : Data Penelitian 2025

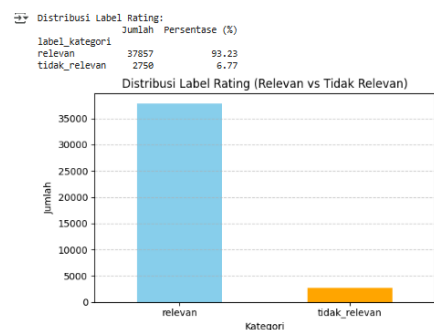
3. Stopword removal

Tabel 3 hasil Stopword Removal

Sebelum stopwords	Hasil
-------------------	-------

[barang, sesuai, pesanan, dan, cepat, sampai]	barang sesuai
barang, bagus, harga, murah]	barang bagus harga murah

Sumber : Data Penelitian 2025
4. Hasil pelebalan rating



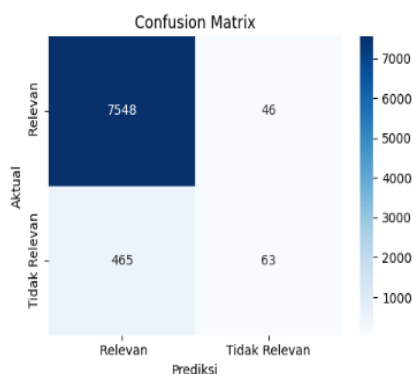
Gambar 5 Hasil Peleban Rating
Sumber : Data Penelitian 2025

4.3 Evaluasi Kinerja Model

Setelah model Naive Bayes digunakan untuk melakukan prediksi terhadap data uji, langkah selanjutnya adalah melakukan evaluasi terhadap performa model. Evaluasi ini bertujuan untuk mengukur seberapa baik model mampu mengklasifikasikan data dengan benar.

1. Hasil Confusion matrix

Confusion Matrix adalah tabel yang menyediakan gambaran visual dan detail tentang kinerja algoritma klasifikasi. Tabel ini menunjukkan jumlah *true positive (TP)*, *true negative (TN)*, *false positive (FP)*, dan *false negative (FN)* yang dibuat oleh model.



Gambar 6 confusion matrix
Sumber : Data Penelitian 2025

Berdasarkan gambar diatas True Positive (TP): 7548 ulasan diprediksi 'Relevan' dan memang benar-benar 'Relevan'. True Negative (TN): 63 ulasan diprediksi 'Tidak Relevan' dan memang benar-benar 'Tidak Relevan'. False Positive (FP): 463 ulasan diprediksi 'Relevan' padahal sebenarnya 'Tidak Relevan'. False Negative (FN): 46 ulasan diprediksi 'Tidak Relevan' padahal sebenarnya 'Relevan'.

2. Hasil Classification Report

Laporan Klasifikasi (*Classification Report*): Laporan ini memberikan detail *precision*, *recall*, dan *F1-score* untuk setiap kelas (Relevan dan Tidak Relevan), serta nilai rata-ratanya (*macro avg* dan *weighted avg*).

Classification Report:				
	precision	recall	f1-score	support
Relevan	0.94	0.99	0.97	7594
Tidak Relevan	0.58	0.12	0.20	528
accuracy			0.94	8122
macro avg	0.76	0.56	0.58	8122
weighted avg	0.92	0.94	0.92	8122

Gambar 7 Hasil *Classification Report*
Sumber : Data Penelitian 2025

$$Akurasi =$$

$$\frac{TP+TN}{TP+TN+FP+FN} \times 100\% = \frac{7611}{8122} \times 100\% = 93.71\%$$

$$Presisi = \frac{TP}{TP+FP} = \frac{7548}{7548+465} = 0.94$$

$$Recall = \frac{TP}{TP+FN} = \frac{7548}{7548+63} = 0.99$$

$$F1 \text{ Score} =$$

$$2 \times \frac{\text{Presisi} \times \text{Recall}}{\text{Presisi} + \text{Recall}} = 2 \times \frac{0.94 \times 0.99}{0.94 + 0.99} = 0.97$$

Hasil pengujian menunjukkan bahwa klasifikasi naive bayes dalam memberikan kinerja yang baik dengan tingkat akurasi 93.71% (94%) dan nilai presisi 0,94%, recall 0,99% dan F1_Score 0,97% menunjukkan bahwa model mampu mengklasifikasi produk dengan tingkat relevan yang tinggi berdasarkan ulasan dan rating yang di uji.

KESIMPULAN

Proses preprocessing teks yang mencakup pembersihan data (*cleaning*), *tokenisasi*, penghapusan *stopword*, dan transformasi TF-IDF sangat efektif dalam menyiapkan data untuk klasifikasi. Hasilnya adalah matriks fitur sebanyak 40.607 baris dengan 5.000 fitur kata yang mewakili informasi penting dari ulasan pengguna. Representasi ini memungkinkan algoritma Naive Bayes untuk mengidentifikasi pola-pola bahasa yang berkorelasi dengan relevansi ulasan secara efisien. hasil analisis evaluasi *matrix* model *naive bayes* mampu

mengatasi permasalahan dalam rekomendasi produk pada transformasi lebeling ulasan dan reteng memberikan tingkat presentasi relevan 93.71% dari total ulasan, sementara presentasi tidak relevan 6.29% dan disusul dengan nilai presisi 0.94, Recall 0.99 dan F1-Score 0.97. mampu mengidentifikasi hampir seluruh ulasan positif dengan baik, sehingga cocok untuk rekomendasi produk berbasis klasifikasi.

DAFTAR PUSTAKA

- A. Pramesti, C. Novitasari, and D. Oktaviani, "Penerapan Manajemen Operasional Di Era Digital Dan Perkembangan E – Commerce," *Econ. Bus. Financ. Entrep.*, vol. 1, no. 1, pp. 88–97, 2023.
- I. M. Sumual, J. Supriadi, E. Effendy, and A. Wijaya, "KLASIFIKASI KATEGORI PRODUK TERLARIS PADA E-COMMERCE," vol. 5, no. 2, pp. 51–60.
- S. Z. Putri, M. Hariadi, and R. F. Rachmadi, "Rekomendasi Barang pada E-commerce menggunakan Content Based Recommender System berbasis Ethereum Blockchain," *J. Tek. ITS*, vol. 12, no. 2, 2023, doi: 10.12962/j23373539.v12i2.115263.
- S. Sylvia and S. Lestari, "Implementasi K-Means Dalam Mengatasi Masalah Cold Star Pada Collaborative Filtering," *Semin. Nas. Has. Penelit. dan Pengabd. Masy.*, pp. 58–63, 2022.
- L. Rahmawati and D. B. Santoso, "Implementasi Metode Naive Bayes Untuk Klasifikasi Ulasan Aplikasi E-Commerce Tokopedia," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 6, no. 1, pp. 116–124, 2023, doi: 10.31539/intecom.v6i1.5515.
- U. Kolbia and N. Dahliyanti, "Di Situs E-Commerce Menggunakan Random Forest ;," *J. Ilm. Inform. Komput.*, vol. 29, no. 2, pp. 182–196, 2024.
- W. Siregar, "Analisis Perilaku Konsumen dalam Transaksi di E-commerce," *Indones. J. Comput. Sci.*, vol. 12, no. 6, 2023, doi: 10.33022/ijcs.v12i6.3483.
- I. Boukrouh and A. Azmani, "Explainable machine learning models applied to predicting customer churn for e-commerce," *IAES Int. J. Artif. Intell.*, vol. 14, no. 1, pp. 286–297, 2025, doi: 10.11591/ijai.v14.i1.pp286-297.
- M. Hermiani *et al.*, "Analisis Ulasan Produk Menggunakan Metode Naive Bayes Classifier," *Jnatia*, vol. 2, no. 1, pp. 177–184, 2023.
- T. Fatiah Rahmadanti, M. Jajuli, and I. Purnamasari, "Klasifikasi Pengguna Shopee Berdasarkan Promosi Menggunakan Naïve Bayes,"

<i>Gener. J.</i>, vol. 5, no. 2, pp. 81–90, 2021, doi: 10.29407/gj.v5i2.15998. H. F. Putro, R. T. Velandari, and W. L. Y. Saptomo, “Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan,” <i>J. Teknol. Inf. dan Komun.</i>, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500. P. Pangestu, R. Novita, and M. Mustakim, “Systematic Literature Review: Perbandingan Algoritma Klasifikasi,” <i>INOVTEK Polbeng - Seri Inform.</i>, vol. 8, no. 2, p. 431, 2023, doi: 10.35314/isi.v8i2.3698. K. S. Putri, I. R. Setiawan, and A. Pambudi, “Analisis Sentimen Terhadap Brand Skincare Lokal Menggunakan Naïve Bayes Classifier,” <i>Technol. J. Ilm.</i>, vol. 14, no. 3, p. 227, 2023, doi: 10.31602/tji.v14i3.11259.	 Intensif Nehe, merupakan mahasiswa Prodi Teknik Informatika Universitas Putera Batam
	 Hotma Pangaribuan S.kom.,M.SI, merupakan Dosen Prodi Teknik Informatika Universitas Putera Batam, yang aktif dan <i>expert</i> di bidang Teknik Informatika