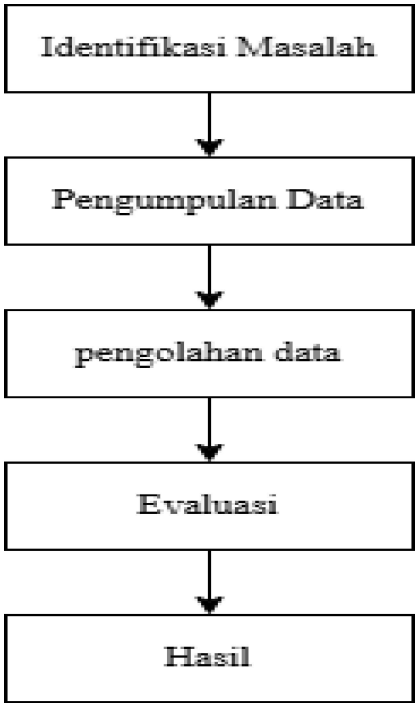
BAB III
METODE PENELITIAN

3.1 Desain Penelitian

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



Gambar 3. 1 Desain Penelitian

Sumber: Data Penelitian 2025

3.1.1 Identifikasi Masalah

Peda penelitian ini mengidentifikasi masalah dalam rekomendasi produk *e-commerce* Tokopedia, terutama kurang nya data histori pada produk sehingga produk yang di pasarkan mengalami penurunan pembeli. Berikutnya produk yang

di promosikan tidak sesuai dengan harapan pembeli sehingga menimbulkan tidak mempercaya platform tersebut atau memberikan ulasan negatif pada produk.

3.1.2 Pengumpulan Data

Data yang digunakan adalah hasil dari pembelian yang memberikan umpan balik pada setiap produk atau ulasan produk Tokopedia

3.1.3 Pengolahan Data

Setalah data di kumpulkan kemudian di oleh menggunakan aplikasi jupyter dengan menggunakan algoritma *naive bayes*.

3.1.4 Evaluasi

Setelah di oleh dengan algoritma *naive bayes* maka akan mendapatkan hasil dengan *confusion matrix* dan hasil Tingkat akurasi, presisi, recall, F1_score.

3.1.5 Hasil

Kemudian mendapat hasil dari confusion matrix dan Tingkat akurasi pada data ulasan produk Tokopedia yang dijadikan sebagai tolak ukur rekomendasi terhadap pelanggan baru.

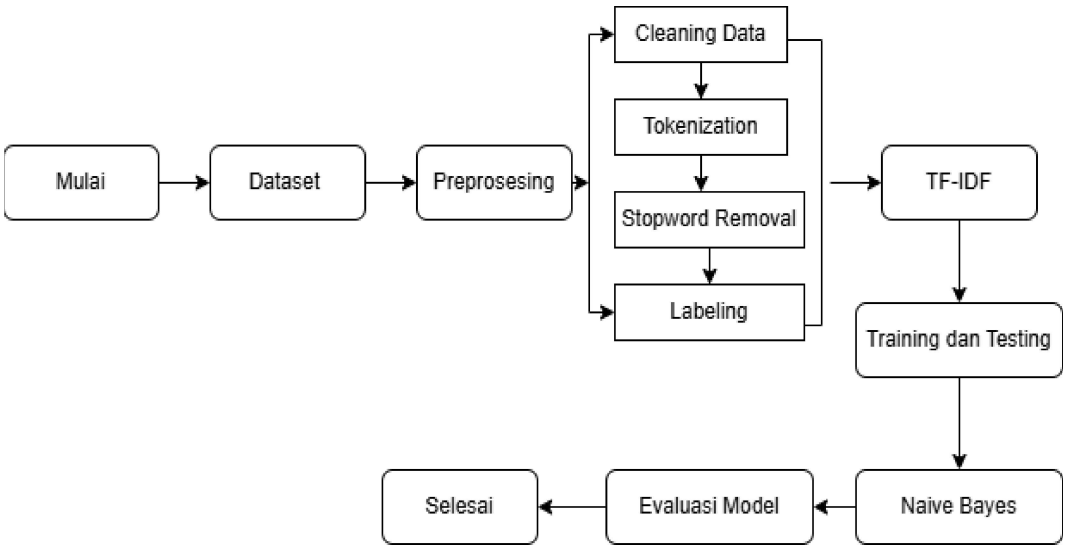
3.2 Metode Pengumpulan Data

Data yang digunakan adalah ulasan produk Tokopedia. Yang di ambil dari *google kaggel* dengan jumlah data sebanyak 40607 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/farhan999/tokopedia-product-reviews

3.3 Metode Perancangan

Pada penelitian ini, metode perencanaan dengan algoritma *naive bayes* baisanya mengacu pada tahapan sistematis dalam membangun model yang cukup pemilihan data, *preprocessing*, pemilihan fitur, pelatihan model, dan evaluasi (Hermiani et al. 2023). Berikut alur algoritma *naïve bayes* dengan menggunakan *flowchart* seperti di bawah ini.



Gambar 3. 9 Perencanaan Model
Sumber: Penelitian 2025

3.3.1 Pengumpulan Data

Sumber data diambil dari situs *google kaggel*. 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.

3.3.2 *preprocessingss*

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 mempengaruhi akurasi dalam proses klasifikasi

d. *Labeling* (pemberian label sentimen)

Proses ini memberikan label berdasarkan nilai rating. Contoh rating 4-5 label positif (1), rating 3 lebel netral, dan rating 1-2 label negatif (0).

3.3.3 Ekstraksi Fitur

Ekstraksi fitur ini bertujuan untuk mengubah data teks (ulasan pengguna) menjadi representasi numerik agar bisa diolah oleh model *naive bayes*. Sebab *naive bayes* bekerja dengan probabilitas berdasarkan frekuensi kata, maka perlu mengubah kedalam bentuk numerik. Ekstraksi fitur menggunakan metode TF-IDF (*tern frequency-inverse document frequency*) (Hermiani et al. 2023).

1. TF (*Term Frequency*) adalah mengukur seberapa sering sebuah kata muncul dalam dokumen tertentu.

Rumus 6 TF (*Term Frequency*)

$$\text{TF}\ (t, d) = \frac{\text{jumlah kemunculan kata } t \text{ dalam dokumen } d}{\text{jumlah total kata dalam dokumen } d}$$

2. IDF (*Inverse Document Frequency*) adalah megukur kata yang jarang muncul dalam seluru data.

$$IDF\ (t) = log\left(\frac{N}{df(t)}\right)$$

Rumus 7 IDF (*Inverse Docement Frequency*)

Keterangan:

N : jumlah total dokumen

Df (t) : jumlah dokumen yang mengadung kata t

3. TF-IDF Score adalah model akan menghitung probabilitas suatu kelas berdasarkan nilai fitur TF-IDF

$$TF - IDF\ (t, d) = TF(t, d)x\ IDF(t)$$

Rumus 8 TF-IDF Score

3.3.4 Treaning dan Testing

Training dan testing merupakan dua tahapan utama dalam proses pembangunan model machine learning. Pada tahap ini, dataset yang telah melalui proses preprocessing dan transformasi (misalnya dengan TF-IDF) akan dibagi menjadi dua bagian: data latih (training set) dan data uji (testing set). Tujuan utama

pembagian ini adalah untuk melatih model dengan sebagian data, lalu menguji performanya pada data yang belum pernah dilihat sebelumnya.

Pada penelitin ini data dibagi dengan 80% untuk training dan 20% untuk testing, pada jumlah data yang tersedia dan strategi evaluasi yang digunakan. Proporsi ini penting agar model tidak hanya belajar dari semua data (yang berisiko menyebabkan overfitting), tetapi juga diuji pada data yang realistis untuk mengetahui kemampuannya dalam generalisasi.

3.3.5 Klasifikasi *Naive Bayes*

Bayesian classification adalah suatu pengklasifikasian statisik yang dapat digunakan untuk memprediksi probabilitas keangotaan. Menerut penelitian (Putro, Vulandari, and Saptomo 2020) *Naïve bayes classification* salah satu algoritma yang sederhana namun memiliki kemapuan dan akurasi tinggi.

Dalam penelitian ini, varian Multinomial Naive Bayes (MultinomialNB) dari pustaka *scikit-learn* digunakan. Varian ini sangat cocok untuk data yang direpresentasikan dalam bentuk hitungan diskrit (seperti frekuensi kata atau bobot *TF-IDF*), di mana fitur-fitur merepresentasikan frekuensi kemunculan *term* dalam dokumen.

3.3.6 Evaluasi Model

Menurut penelitian (K. S. Putri, Setiawan, and Pambudi 2023) evaluasi dalam klasifikasi *naive bayes* dilakukan dengan matrik statistic seperti akurasi, presisi , recall dan F1_score serta confusion matrix. Evaluasi ini memberikan Gambaran seberapa andal dan akurat model sebelum di implementasikan.

3.4 Jadwal Peneltian

Berikut jadwal penelitian yang di lakukan oleh peneliti selama satu semester.

Tabel 1 Jadwal Penelitian

Pertemuan	1	2	3	4	5	6	7	8	9	10	11	12	13	14
BAB I														
BAB II														
BAB III														
BAB IV														
BAB V														