

## BAB III

### METODOLOGI PENELITIAN

#### 3.1 Gambaran Umum Objek Penelitian

Penelitian ini berfokus pada analisis sentimen terhadap tiga aplikasi *Online Travel Agent* (OTA) yang populer di Indonesia, yaitu Traveloka, Tiket.com, dan Agoda. Data diperoleh melalui metode web scraping terhadap 1.000 ulasan pengguna dari masing-masing aplikasi di Google Play Store.

Ulasan yang dikumpulkan digunakan untuk membangun model klasifikasi sentimen menggunakan lima algoritma supervised learning, yaitu Naïve Bayes, Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Random Forest. Penelitian ini bertujuan untuk membandingkan performa masing-masing algoritma dalam mengklasifikasikan sentimen pengguna berdasarkan metrik akurasi.

#### 3.2 Metode Penelitian

Metode KDD digunakan dalam penelitian ini karena memiliki alur yang sistematis dalam pengolahan data, mulai dari tahap seleksi hingga evaluasi. Untuk memperkuat pemilihan metode, dilakukan perbandingan antara KDD dan CRISP-DM sebagaimana ditunjukkan pada Tabel 3.1 [90]. Tabel tersebut memperlihatkan bahwa meskipun keduanya memiliki struktur yang mirip, KDD lebih fokus pada proses penemuan pengetahuan (knowledge discovery) yang sesuai dengan kebutuhan penelitian ini.

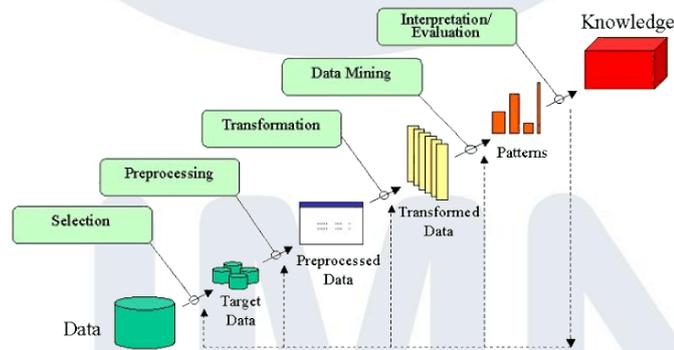
*Tabel 3.1 KDD vs CRISP DM*

	<b>KDD</b>	<b>CRISP-DM</b>
Fleksibilitas	Memberikan fleksibilitas dalam pendekatan analisis dan pemodelan data	Lebih terstruktur dan memberikan panduan yang jelas dalam setiap tahapan
Pengembangan model	Pemilihan model dan evaluasi dapat lebih fleksibel	Memberikan panduan yang lebih spesifik untuk pemilihan model dan evaluasi tahap tertentu
Integrasi dengan algoritma yang digunakan	Dapat diintegrasikan dengan berbagai algoritma, termasuk naïve bayes dan knn	Tidak mengikat secara khusus dengan algoritma tertentu, memberikan kebebasan dalam pemilihan algoritma

Skala	Lebih efisien untuk proyek skala menengah dan volume data yang besar	Lebih efisien untuk proyek skala besar, namun tergantung algoritma yang digunakan
Kompleksitas	Lebih sederhana dan mudah dipelajari	Lebih kompleks dan membutuhkan pemahaman lebih yang mendalam
Deployment	Tidak Ada	Ada

Table 3.1 menjelaskan bahwa metode KDD dan juga CRISP DM. Pada penelitian kali ini, digunakan metode KDD karena lebih fleksibel dan terintegrasi dengan konsep penemuan secara umum. Metodologi KDD tidak memerlukan tahap *deployment* sebagaimana yang terdapat dalam CRISP-DM. Hal ini dikarenakan KDD lebih berfokus pada proses ekstraksi pola dan pengetahuan dari data, sedangkan CRISP-DM mencakup tahapan implementasi model ke dalam sistem operasional. Untuk yang baru memulai penelitian analisis sentiment, metode KDD cocok karena metode ini lebih sederhana dan mudah dipelajari, sehingga dapat lebih cepat memahami proses analisis sentiment.

### 3.2.1 Metode KDD



Gambar 3.1 Framework KDD[91]

Penelitian ini menggunakan pendekatan KDD sebagai framework utama dalam proses analisis data. Framework KDD digunakan untuk mengekstraksi pola dan wawasan dari data ulasan pengguna aplikasi OTA[92]. Berikut merupakan tahapan penelitian yang dilakukan dalam proses KDD:

#### a. Selection

Tahap ini melibatkan pemilihan data yang relevan untuk analisis sentiment dari sumber data Traveloka, Tiket.com dan Agoda seperti

ulasan pengguna atau feedback pelanggan. Berikut merupakan tabel detail isi dari dataset yang digunakan untuk penelitian:

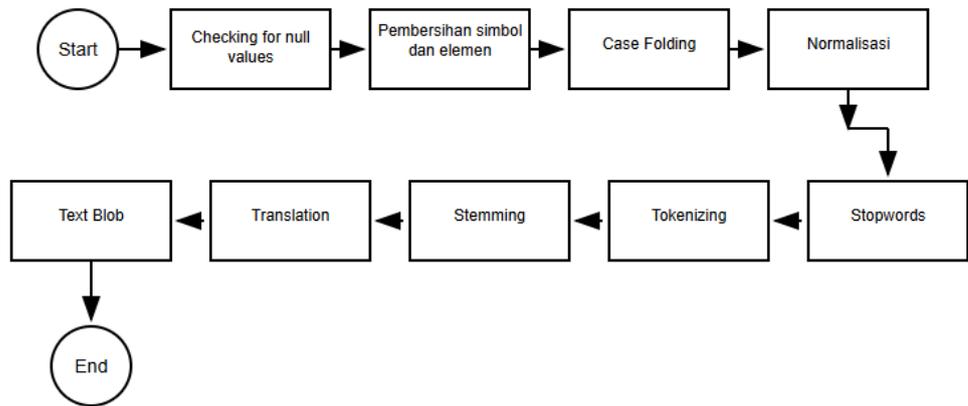
Tabel 3. 2 Variabel Dataset

Variabel	Jenis Variabel	Keterangan
Username	Categorical	Variabel ini berisikan nama pengguna pada masing-masing aplikasi di Google Play Store.
Score	Numerical	Variabel ini berisikan peringkat atau skor rating dari masing-masing aplikasi yang diberikan oleh user.
At	Time Series	At disini berisikan kapan user memberikan komentar dan rating pada masing-masing aplikasi
Content	Categorical	Content berisikan isi dari ulasan yang diberikan oleh user terhadap masing-masing aplikasi

b. *Pre Processing*

Preprocessing merupakan tahap penting dalam analisis data teks yang bertujuan untuk membersihkan dan menyiapkan data agar dapat diproses secara efektif oleh algoritma klasifikasi. Pada penelitian ini, data yang digunakan berupa ulasan teks dari pengguna aplikasi OTA, sehingga perlu melalui serangkaian tahapan preprocessing untuk menghilangkan elemen-elemen yang tidak relevan dan menyatukan format teks agar lebih terstruktur.

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



Gambar 3. 2 Flowchart Tahap Preprocessing

Pada Gambar 3.2 menunjukkan bahwa tahap preprocessing akan mengolah data untuk memastikan kualitas dan konsistensi sebelum masuk ke tahap analisis sentimen. Proses ini mencakup pembersihan data, case folding, normalisasi, stopword removal, tokenizing, stemming, translasi dan analisis sentiment menggunakan text blob (positif/negatif)

c. Transformation

Tahap *transformation* merupakan proses mengubah data teks yang telah melalui preprocessing menjadi representasi numerik yang dapat dipahami dan diproses oleh algoritma pembelajaran mesin. Karena algoritma klasifikasi tidak dapat bekerja langsung pada data dalam bentuk teks mentah, maka diperlukan transformasi ke dalam bentuk fitur numerik menggunakan teknik representasi teks.



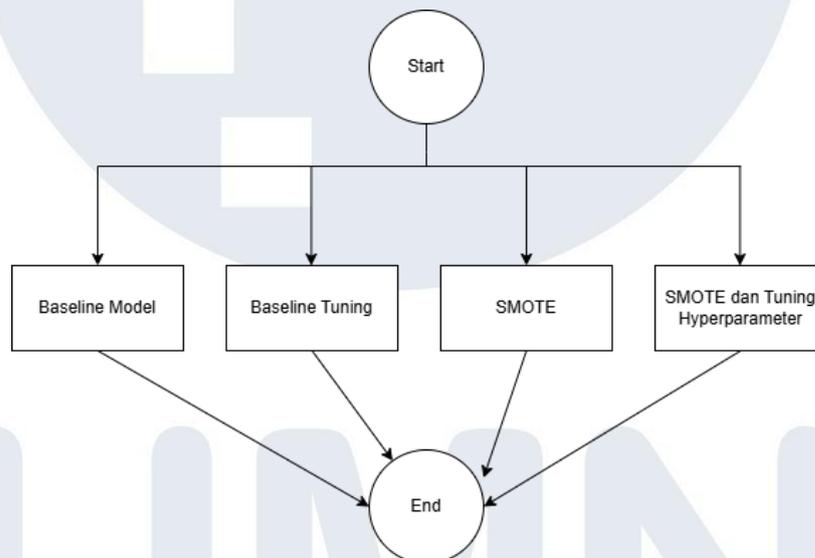
Gambar 3. 3 Flowchart Tahap Transformation

Pada Gambar 3.3 menunjukkan bahwa tahap transformation akan menyesuaikan format dan struktur data agar dapat diaplikasikan pada model algoritma. Proses ini meliputi *splitting data* fitur X dan y, serta

penerapan metode TF-IDF untuk mengubah teks dalam bentuk numerik dan juga dilakukannya deklarasi SMOTE untuk menangani data yang *imbalanced* atau tidak seimbang.

d. *Data Mining*

Tahap data mining merupakan inti dari proses analisis, yaitu proses penerapan algoritma klasifikasi untuk membangun model yang mampu mengenali pola sentimen dalam data ulasan pengguna. Dalam penelitian ini, proses data mining dilakukan untuk membangun model prediksi sentimen berdasarkan ulasan aplikasi OTA (Traveloka, Tiket.com, dan Agoda) yang telah melalui tahapan preprocessing dan transformasi.



Gambar 3. 4 Flowchart Tahap Data Mining

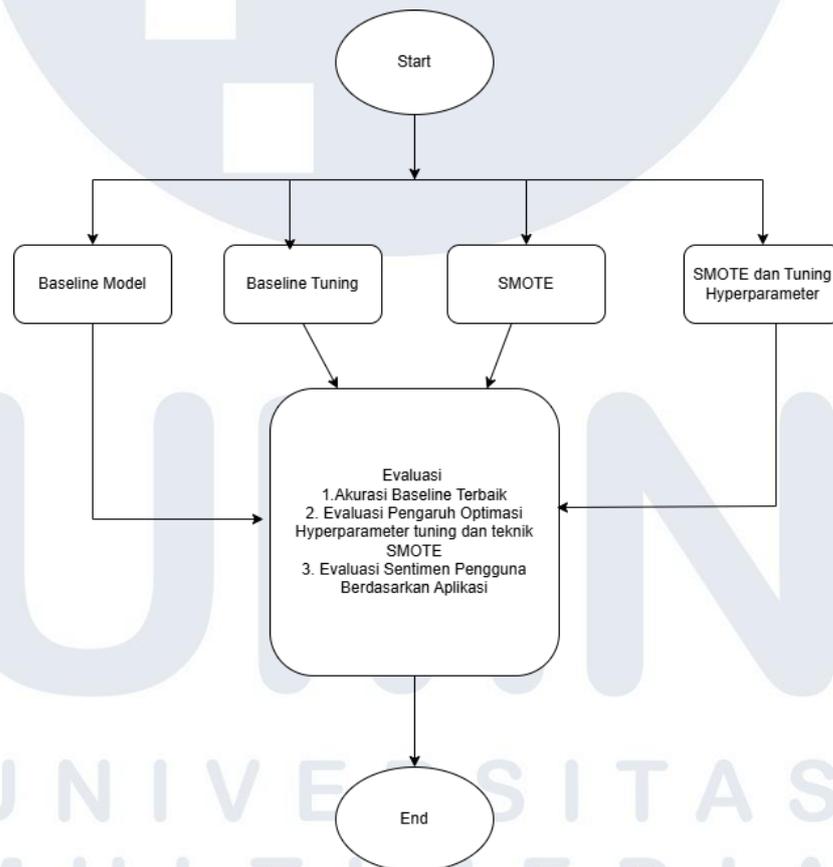
Pada Gambar 4.4 menunjukkan proses penerapan algoritma Naive Bayes, KNN, SVM, Decision Tree dan Random Forest untuk mengekstrak pola sentimen dari data ulasan ketiga aplikasi OTA.

Metode yang diterapkan pada masing-masing aplikasi:

- Ketiga aplikasi akan menggunakan Baseline Model – Tuning (Baseline) – SMOTE – Tuning (dengan SMOTE)

e. *Evaluation*

ada tahapan evaluasi dalam proses *Knowledge Discovery in Databases* (KDD), dilakukan pengukuran kinerja model klasifikasi untuk menentukan seberapa baik algoritma *machine learning* yang digunakan mampu mengenali pola sentimen dalam data ulasan. Evaluasi ini mencakup analisis terhadap metrik performa seperti akurasi, presisi, recall, dan F1-score, serta membandingkan hasil antar skenario—mulai dari baseline hingga penerapan tuning hyperparameter dan penanganan imbalance data menggunakan SMOTE. Hasil evaluasi berperan penting dalam mengidentifikasi pendekatan terbaik yang dapat menghasilkan klasifikasi sentimen yang paling akurat dan stabil.



Gambar 3.5 Flowchart Tahap Evaluation

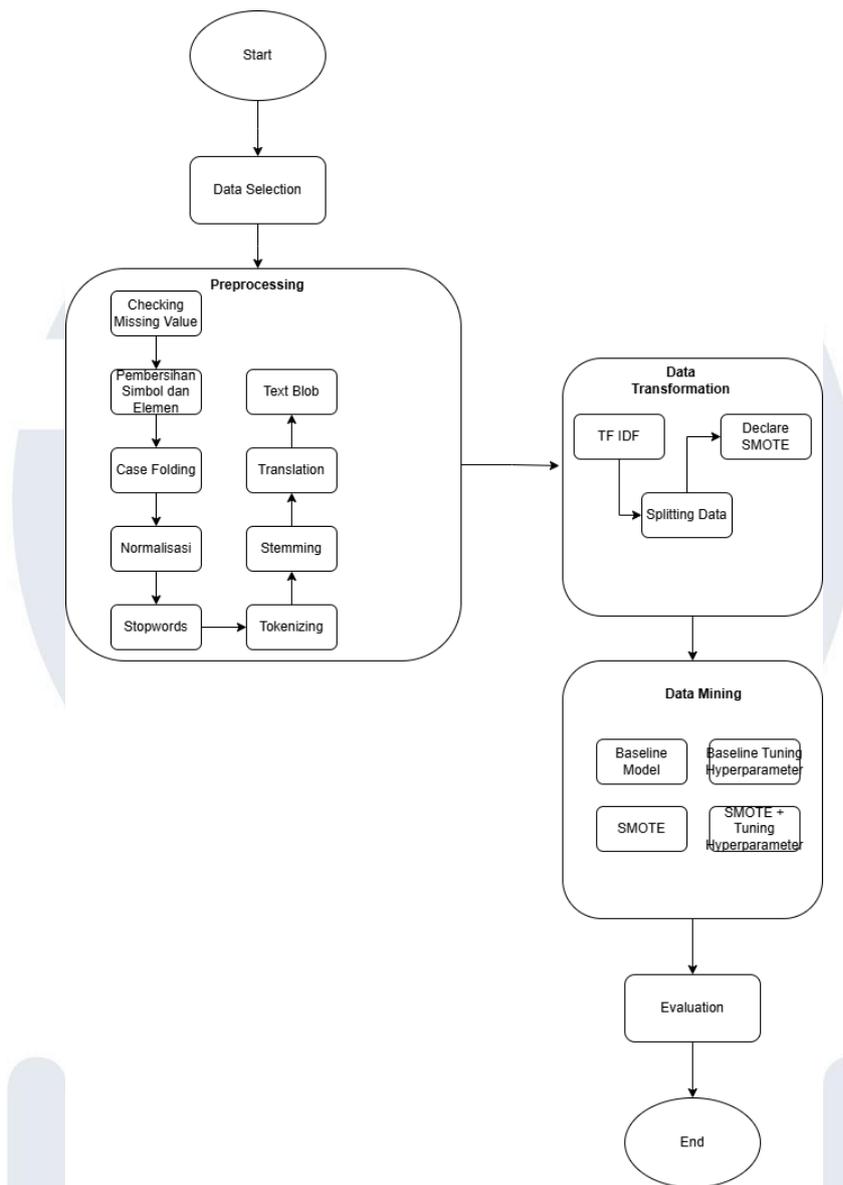
Pada Gambar 3.5 menunjukkan bahwa nantinya pada tahap evaluation akan menilai kinerja model keseluruhan dan hasil analisis sentimen

untuk memahami sejauh mana model dapat mengklasifikasikan sentimen dengan benar.

### 3.2.2 Alur Penelitian

Alur penelitian sangat penting untuk memberikan panduan yang jelas bagi peneliti, agar proses penelitian dapat berjalan dengan lancar. Hal ini membantu peneliti untuk tetap fokus dan tidak kehilangan arah. Alur penelitian tersebut dapat dilihat pada Gambar 3.6.

Gambar 3.6 menunjukkan alur pada penelitian kali ini. Penelitian diawali dengan tahap data selection yang dimana nanti data akan didapatkan dengan cara *web scraping* pada 3 aplikasi (Traveloka, Tiket.com dan Agoda), lalu data yang sudah didapat akan dilakukan teknik *preprocessing* dengan berbagai cara. Langkah selanjutnya adalah Data Transformation yang dimana TF-IDF akan digunakan untuk mengubah kata kata menjadi numerik, *splitting data* serta *declare SMOTE*. Pada tahap Data Mining nantinya akan dilakukan menggunakan 5 algoritma klasifikasi (Naïve Bayes, Decision Tree, Random Forest, KNN, SVM) dengan skenario *baseline – baseline tuning – SMOTE – SMOTE + Tuning*. Tahap selanjutnya adalah evaluasi model untuk menilai performa setiap algoritma yang digunakan. Evaluasi dilakukan dengan membandingkan akurasi model sebelum dan setelah optimasi, guna mengetahui sejauh mana peningkatan yang terjadi.



Gambar 3. 6 Alur Penelitian

### 3.3 Variabel Penelitian

Terdapat 2 variabel penelitian dalam penelitian ini, yaitu independent dan dependen. Diantaranya adalah:

#### 3.3.1 Variabel Independen

Dalam penelitian ini, variabel dependen yang digunakan adalah sentimen ulasan pengguna, yang dikategorikan menjadi positif dan negatif. Kategori tersebut ditentukan berdasarkan analisis teks ulasan yang diperoleh dari tiga aplikasi OTA, yaitu Traveloka, Tiket.com, dan

Agoda. Sentimen ini menjadi target utama yang ingin diprediksi oleh model klasifikasi dalam penelitian.

### 3.3.2 Variabel Dependen

Variabel independen terdiri dari berbagai fitur yang diekstrak dari teks ulasan pengguna. Salah satu fitur utama yang digunakan adalah representasi teks dalam bentuk TF-IDF, yang berfungsi mengubah kumpulan kata menjadi format numerik. Representasi ini memungkinkan model machine learning untuk memproses dan mengenali pola dalam data teks secara matematis.

### 3.4 Teknik Pengumpulan Data

Teknik pengumpulan data ulasan pengguna aplikasi berdasarkan Google Play Store dilakukan menggunakan Google Colab. Dalam proses ini, Google Colab dimanfaatkan untuk menjalankan skrip scraping secara otomatis dengan bantuan pustaka seperti Selenium. Pendekatan ini memungkinkan pengambilan data dilakukan secara efisien, terstruktur, dan dapat disesuaikan dengan kebutuhan analisis sentimen. Perbedaan antara kedua aplikasi dapat dilihat dalam Tabel 3.3.

Tabel 3. 3 Google Colab vs Selenium

	Web Scraping dengan Google Colab	Web Scraping dengan Selenium
Kompleksitas	Lebih sederhana dan mudah dipelajari	Lebih kompleks dan membutuhkan pemahaman tentang web browser
Kecepatan	Cenderung lebih cepat karena tidak memerlukan rendering browser	Cenderung lebih lambat karena melibatkan browser dan rendering halaman web
Kemudahan Penggunaan	Lebih mudah karena tersedianya library seperti BeautifulSoup atau Request	Memerlukan pemahaman lebih lanjut dan lebih rumit karena melibatkan browser dan interaksi dinamis
Ketersediaan	Gratis dan tersedia secara online	Berbayar dan harus di install di komputer lokal
Fitur tambahan	Mendukung pemrosesan data di cloud, sehingga memiliki skalabilitas tinggi	Tidak mendukung pemrosesan data di cloud

Berdasarkan Tabel 3.3 menunjukkan teknik pengumpulan data ulasan dilakukan dengan cara web scraping menggunakan Google Colab dengan bahasa pemrograman Python. Pemilihan Google Colab didasarkan pada kemudahan penggunaannya, ketersediaan yang gratis dan berbasis cloud, serta dukungannya terhadap pustaka scraping seperti BeautifulSoup dan Requests. Selain itu, Google Colab memungkinkan pemrosesan data secara efisien tanpa memerlukan instalasi lokal, sehingga lebih praktis dibandingkan dengan penggunaan Selenium yang membutuhkan browser rendering dan setup tambahan.

```
df_busu[['userName', 'score', 'at', 'content']].head() #preview userName, rating,
```

	userName	score	at	content
0	Dina Mrdn	1	2025-06-12 20:20:07	kenapa kalau mau melakukan pembayaran harganya...
1	Anindra	5	2025-06-20 17:57:24	Liburan ke Malang bareng temen, semua bookinga...
2	Rapil Rtt	5	2025-06-16 19:12:49	Salah satu fitur terbaik adalah informasi posi...
3	Whiska Syahputra	1	2025-06-16 13:21:04	pelayanan buruk sekali, komplain hanya dilempa...
4	Ready multimedia	3	2025-06-13 12:41:36	saya mau pesan tiket kereta dari jember ke jak...

Gambar 3. 7 kolom yang diambil untuk dataset

Gambar 3.7 menjelaskan bahwa kolom yang diambil dalam proses pengumpulan data terdiri dari username, score, at, serta content. Username sebagai pengguna, score sebagai rating, at merupakan tahun, bulan dan tanggal ulasan dan content merupakan isi konten dari ulasan pengguna.

Tabel 3. 4 Isi Dataset Scraping Traveloka

Username	Score	At	content
Muh Saleh	5	2/2/2025 15:47	Traveloka sangat bagus dan membantu saya untuk memesan tiket dan lain lain nya
Heri Ana		2/2/2025 13:08	Kenapa ya traveloka sekarang suka gangguan. Dan susah utk memuat ulang pdhal update sudah? Agak ocewa pdhal pelanggan sudah lama dan sllu pakai
Santoso Priartono	5	2/2/2025 12:05	Dikasih limit paylater,dan bisa dicicil 3x pembayaran atau 3

Username	Score	At	content
			bulan. Sangat bermanfaat. Traveloka kalau qris buat belanja paylater tambah mantap kayanya
Elyati Elyati	5	2/2/2025 10:35	baik dan sangat membantu saya dalam perjalanan
Moch Cuncunnurrosit	1	2/2/2025 9:49	Harga tiket keluar negeri lebih murah dari pada penerbangan domestik

Tabel 3.4 menunjukkan lima contoh ulasan pengguna aplikasi Traveloka yang diperoleh melalui proses web scraping dari Google Play Store. Data ini merupakan bagian dari total 1000 ulasan yang dikumpulkan selama periode scraping pada Januari hingga Februari 2025. Setiap entri mencakup nama pengguna, skor rating yang diberikan (1–5), waktu ulasan diposting, serta isi komentar. Komentar-komentar ini nantinya diproses lebih lanjut dalam tahap preprocessing dan digunakan sebagai input untuk analisis sentimen menggunakan berbagai algoritma klasifikasi.

### 3.5 Teknik Analisis Data

Pengumpulan data dilakukan melalui web scraping pada periode Januari hingga Februari 2025. Sebanyak 1.000 ulasan berhasil dikumpulkan dari komentar pengguna aplikasi OTA di Google Play Store. Rentang waktu ulasan yang diambil mencakup Desember 2018 hingga Februari 2025. Berikut merupakan tahap tahap Teknik analisis data pada penelitian kali ini:

#### a) Tahap *Pre Processing*

Tahap ini akan dilakukan pembersihan data yang mencakup proses penghapusan duplikasi, normalisasi teks, stopwords removal, tokenisasi, stemming, serta penerjemahan ke dalam bahasa Inggris. Setelah tahap preprocessing selesai, data kemudian dianalisis sentimennya menggunakan library TextBlob untuk menentukan apakah ulasan bersentimen positif atau

negatif. Proses ini bertujuan untuk menyiapkan data teks dalam format yang bersih dan terstruktur sebelum digunakan dalam pemodelan.

b) Tahap Labeling sentiment

Tahap labeling sentimen pada penelitian ini dilakukan menggunakan pendekatan lexicon. Proses ini dilakukan secara otomatis dengan memanfaatkan kamus kata yang sudah memiliki nilai sentimen, baik positif maupun negatif. Dengan cara ini, setiap ulasan bisa diberi label sentimen tanpa perlu melakukan pelabelan manual satu per satu.

c) Tahap Pembagian data

Setelah proses labeling selesai, data dibagi menjadi dua set, yaitu data training dan data testing. Data training digunakan untuk melatih model agar dapat mengenali pola dalam data, sedangkan data testing digunakan untuk mengukur sejauh mana model mampu melakukan prediksi terhadap data yang belum pernah dilihat sebelumnya. Pembagian ini penting agar evaluasi model lebih objektif dan tidak overfitting terhadap data latih.

d) Tahap Ekstraksi fitur

Tahap ini nantinya akan menggunakan metode TF-IDF untuk mengubah teks ulasan menjadi representasi vektor numerik yang dapat digunakan oleh algoritma klasifikasi. Dengan pendekatan ini, setiap kata dalam dokumen diberikan bobot berdasarkan frekuensi kemunculannya di satu dokumen dan keseluruhan korpus. Representasi vektor ini memungkinkan model machine learning mengenali pola dari teks ulasan secara matematis dan melakukan klasifikasi secara lebih akurat.

e) Evaluation

Tahap pembuatan model dilakukan dengan menerapkan lima algoritma klasifikasi, yaitu Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, dan Random Forest. Masing-masing algoritma digunakan untuk membangun model klasifikasi sentimen berdasarkan teks ulasan pengguna, dengan pendekatan dan

karakteristik yang berbeda-beda. Penjelasan mengenai cara kerja kelima algoritma tersebut disajikan pada Tabel 3.5.

Tabel 3. 5 Model algoritma

<b>Naïve Bayes</b>	<b><i>K-Nearest Neighbors (KNN)</i></b>	<b><i>SVM</i></b>	<b><i>Decision Tree</i></b>	<b><i>Random Forest</i></b>
Melatih model Naive Bayes untuk klasifikasi sentimen positif dan negative berdasarkan probabilitas kata dalam teks[93].	Menerapkan algoritma KNN untuk menentukan kelas sentimen dengan membandingkan kemiripan ulasan terhadap ulasan lain di sekitarnya[94].	Memisahkan ulasan berdasarkan vektor di ruang multidimensi untuk menemukan hyperplane terbaik yang membagi sentimen positif dan negatif[95].	Membangun pohon keputusan yang membagi data berdasarkan fitur kata tertentu untuk mengklasifikasikan ulasan ke dalam kategori sentimen[96].	Menggunakan beberapa pohon keputusan secara bersamaan untuk meningkatkan akurasi dalam menentukan sentimen berdasarkan fitur teks yang ada[83].

#### f) Evaluasi Model

Pada tahap evaluasi nantinya akan dilakukan perbandingan akurasi antara masing-masing algoritma dan aplikasi untuk mengetahui model mana yang memberikan performa terbaik dalam analisis sentimen. Selain itu, akan dibandingkan pula pengaruh penerapan teknik optimasi seperti tuning hyperparameter dan SMOTE terhadap peningkatan akurasi model. Dengan demikian, evaluasi ini diharapkan dapat memberikan gambaran menyeluruh mengenai efektivitas setiap algoritma dan strategi optimasinya.

U N I V E R S I T A S  
M U L T I M E D I A  
N U S A N T A R A