# BAB III
# METODE PENELITIAN

## 3.1 Objek Penelitian

Komentar TikTok tentang mobil listrik dikumpulkan melalui teknik *crawling* dalam dua tahap. Tahap pertama pada 31 Maret 2024, pukul 11:11 – 20:44 WIB, menghasil 1.032 komentar untuk publikasi jurnal. Penambahan data dilakukan pada 11-15 Desember 2024 untuk memenuhi kebutuhan Tugas Akhir. Total keseluruhan data yang digunakan mencapai 2.567 komentar. Rincian dataset dapat dilihat pada Tabel 3.1.

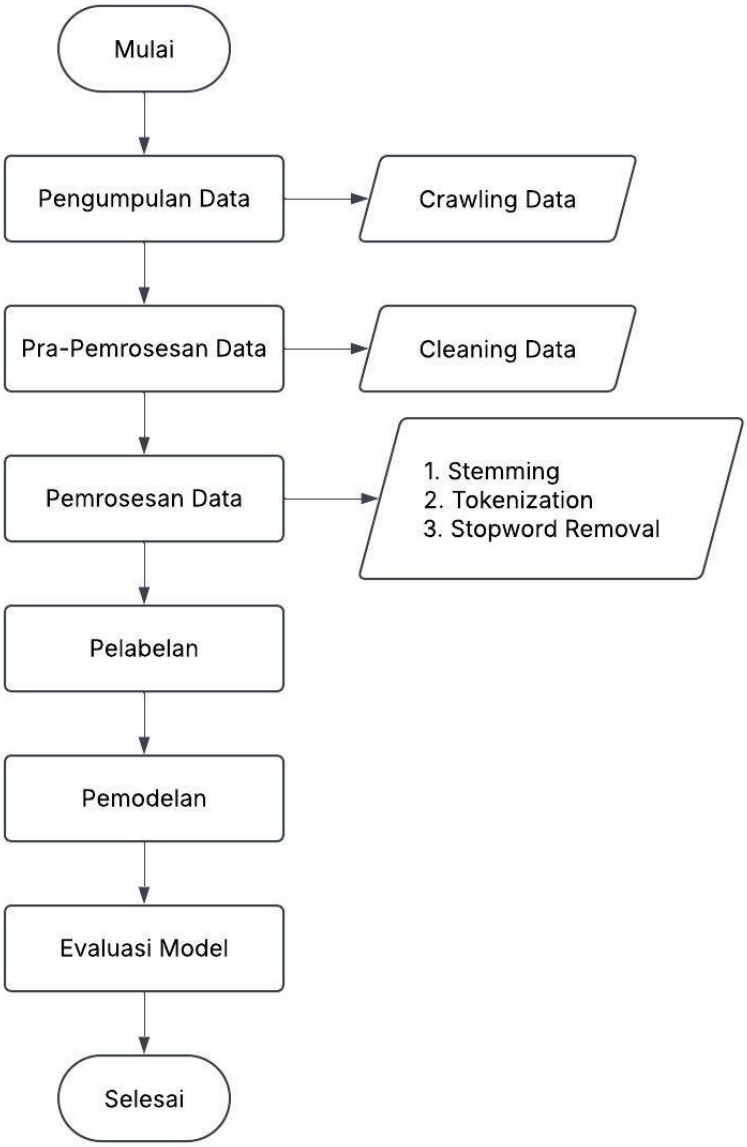
Tabel 3. 1 Rincian Dataset

| full_text | username |
|---|---|
| cuan terus cuannya di pake buat benerin lingku... | Marrlz \| 𝔐𝔦𝔠𝔠 |
| lah gw tinggal di Halmahera 🗿 | BG aris |
| guwah sih lingkungan bang dari padq cuan | zan ART⚡ |
| … | … |
| gua pilih lingkungan aja bg | marvel |
| kemaren ad org wuling makan di warteg pake bin... | Marchel |

## 3.2 Prosedur Penelitian

Tahap pertama dalam penelitian ini adalah melakukan analisis kebutuhan untuk menentukan data yang diperlukan, diikuti pengumpulan komentar pada platform TikTok menggunakan metode *crawling*. Tahap pra-pemrosesan data dilakukan untuk membersihkan data, diikuti tahap pemrosesan data seperti tokenisasi, menghilangkan kata-kata yang tidak berperan besar dalam pemahaman makna, serta normalisasi untuk memastikan keseragaman format dan ejaan.

Penelitian dilanjutkan dengan tahap pelabelan, di mana data yang telah melalui tahap pemrosesan diklasifikasikan berdasarkan kategori yang telah ditentukan menggungunakan metode TF-IDF untuk pembobotan kata. Setelah itu, data yang telah diberi label digunakan dalam tahap pemodelan untuk analisis lebih lanjut. Adapun tahap proses pengolahan data dapat dilihat pada Gambar 3.1.




Gambar 3. 1 Alur Penelitian

### 3.2.1 Pengumpulan Data

Data diperoleh melalui metode *crawling*, sebuah proses otomatis untuk mengumpulkan data dari sumber online seperti situs web dan media sosial

(Muliawan & Dazki, 2023). Dalam penelitian ini, proses *crawling* difokuskan pada platform TikTok untuk mengumpulkan komentar yang relevan terkait mobil listrik. Total jumlah data yang berhasil dikumpulkan adalah sebanyak 2.567.

### 3.2.2 Pra-Pemrosesan Data

Langkah berikutnya adalah pra-pemrosesan data, yang bertujuan mengubah data mentah menjadi data yang siap untuk analisis lebih lanjut (Riyadi et al., 2021). Data mentah yang diperoleh dari tahap *crawling* akan melalui proses pembersihan untuk menghilangkan komentar yang tidak bernilai atau null, serta menghapus data yang duplikat.

Proses ini dilanjutkan dengan penghapusan elemen-elemen yang tidak relevan, seperti simbol, tagar, URL, dan alamat email.

### 3.2.3 Pemrosesan Data

Tahap pemrosesan data dilakukan setelah data melalui pra-pemrosesan, dengan tujuan utama untuk menyempurnakan kualitas data agar siap digunakan dalam proses analisis. Dalam tahap ini, terdapat tiga langkah utama yang akan dilakukan yaitu:

a. *Stemming,* tahap ini bertujuan menormalisasi kata dengan mengubahnya ke bentuk dasar, seperti menghapus tanda baca dan kata-kata umum yang tidak relevan, sehingga meningkatkan efisiensi analisis teks (Rinaldi et al., 2024).
b. *Tokenization,* tahap ini dilakukan untuk membagi teks atau data menjadi unit-unit yang lebih kecil. Proses ini memecah kalimat menjadi potongan-potongan kata (Fikri et al., 2020). Tujuan utama dari tahap ini untuk mempermudah analisis pada tahap *stopword removal.*
c. *Stopword Removal,* pada tahap ini dilakukan panyaringan pada daftar kata yang tidak memiliki nilai atau relevansi, karena tidak memberikan kontribusi informasi pada kalimat (Alin et al., 2023). Kata-kata tersebut kemudian dihapus dari teks. Setelah itu, dilanjutkan ke tahap *normalization*, yang bertujuan untuk menstandarkan teks agar memiliki format yang seragam, proses ini mencakup penghapusan tanda baca,

penyesuaian ejaan konsisten dengan bentuk baku, serta perbaikan kata yang mengalami kesalahan ketik sehingga dapat dikenali dan diproses dengan labih akurat oleh model (Aurelly et al., 2024).

### 3.2.4 Pelabelan

Proses pelabelan atau pemberian label adalah langkah untuk menentukan respon dari setiap komentar dalam dataset secara terperinci (Santoso et al., 2022). Padat tahap ini, setiap komentar diberi label untuk diklasifikasikan ke dalam tiga kategori: positif, negatif, dan netral. Bobot kata ditentukan berdasarkan nilai *Term Frequency-Inverse Document Frequency* (TF-IDF).

Untuk memastikan keakuratan pembobotan, validasi dilakukan dengan melibatkan pakar ahli bahasa. Validasi ini bertujuan untuk mengevaluasi pembobotan kata yang dilakukan menggunakan TF-IDF telah sesuai dengan konteks linguistik dan makna yang ada dalam dataset.

### 3.2.5 Pemodelan

Tahap ini, melibatkan penggunaan data komentar yang telah diproses dan dilabeli untuk menguji model klasifikasi. Pemodelan ini menerapkan algoritma *Support Vector Machine* dan *K-Nearest Neighbor*. Kedua algoritma tersebut dipilih karena keunggulannya dalam menganalisis dan mengklasifikasikan data teks yang kompleks.

*a. Support Vector Machine*

Dalam *machine learning*, *Support Vector Machine* (SVM) digunakan untuk menyelesaikan masalah klasifikasi dan regresi. Prinsip utama SVM adalah menemukan *hyperplane* yang memisahkan dua kelas data dengan margin maksimal, sehingga menghasilkan keputusan yang optimal. Model ini bekerja pada ruang fitur berdimensi tinggi dengan memanfaatkan fungsi *linear* sebagai dasar pembentukan hipotesis (Rina Noviana & Isram Rasal, 2023).

Dalam penelitian ini, SVM diterapkan pada tahap pelatihan dan pengujian setelah data melalui tahap pelabelan. Tahap penerapan model secara rinci dijelaskan pada Tabel 3.2.

Tabel 3. 2 Pseudocode Algoritma *Support Vector Machine*

| **Algoritma *Support Vector Machine*** |
|---|
| 1. Ganti nilai kosong pada kolom “full_text” dengan string kosong |
| 2. Vektorisasi kolom “full_text” menjadi matriks numerik untuk algoritma *machine learning* |
| 3. Bagi data menjadi 70% data latih dan 30% data uji |
| 4. Siapkan model SVM dengan tipe kernel *linear* |
| 5. Latih model SVM dengan data latih |
| 6. Prediksi label untuk data uji |
| 7. Hitung *confusion matrix*, metrik evaluasi, dan akurasi |
| 8. Tampilkan hasil evaluasi |

*b. K-Nearest Neighbor*

*K-Nearest Neighbor* (K-NN) adalah algoritma dalam pembelajaran terawasi yang digunakan untuk tugas klasifikasi dan regresi. Cara kerja algoritma ini adalah dengan menemukan K titik data terdekat dari sampel yang dianalisis, lalu memanfaatkan label atau nilai dari titik-titik tersebut untuk memprediksi kelas atau nilai sampel tersebut. Nilai K harus dipastikan tidak melebihi jumlah total data yang ada (Syahril Dwi Prasetyo et al., 2023).

Penerapan K-NN dilakukan pada tahap pelatihan dan pengujian setelah proses pelabelan. Rincian mengenai tahap penerapan model K-NN dapat dilihat pada Tabel 3.3.

Tabel 3. 3 Pseudocode Algoritma *K-Nearest Neighbor*

| **Algoritma *K-Nearest Neighbor*** |
| --- |
| 1. Ganti niai kosong pada kolom “full_text” dengan string kosong. |
| 2. Ubah kolom “full_text” menjadi matriks fitur numerik. |
| 3. Bagi data menjadi data latih (70%) data uji (30%). |
| 4. Inisialisasi model K-NN. |
| 5. Latih model menggunakan data latih. |
| 6. Prediksi model yang telah dilatih untuk mengklasifikasikan data uji. |
| 7. Hitung *confusion matrix*, metrik evaluasi, dan akurasi |
| 8. Tampilkan hasil evaluasi |

### 3.2.6 Evaluasi Model

*Confusion matrix* digunakan untuk menilai performa model klasifikasi dalam pembelajaran mesin dan memberikan gambaran tentang seberapa efektif model tersebut dalam memprediksi berbagai label. Dalam penelitian ini, *confusion matrix* berfungsi sebagai alat untuk mengukur metrik daya seperti *precision*, *recall*, *f1-score*, dan akurasi (Rahayu et al., 2021).