BAB III
METODE PENELITIAN

3.1 Objek Penelitian

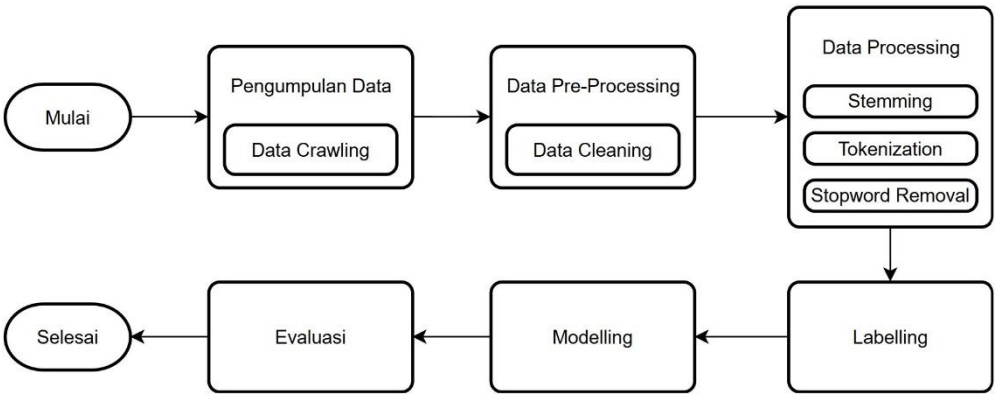
Objek penelitian yang digunakan berupa data teks komentar dari TikTok. Sebanyak 1032 komentar berhasil dikumpulkan dengan menggunakan *keyword* "mobil listrik" pada tanggal 31 Maret 2024, antara pukul 11:11 hingga 20:44. Kemudian, pada tanggal 9 Desember 2024 dari pukul 13:40 hingga 18:20 dikumpulkan data tambahan sebanyak 1231. Total data teks komentar yang berhasil dikumpulkan yaitu 2263. Data yang terkumpul selanjutnya disimpan dalam format Excel untuk analisis lebih lanjut. Rincian dataset dapat dilihat pada Tabel 3.1.

Tabel 3. 1 Dataset Komentar TikTok

username	full_text
Mie333ay4eunmmm	33 menit ngecas doang? Klo di pikir2 buang2 waktu bgtt ya apa lg klo musim mudik
Luvias	pengen beli tapi inget listrik ku cman 1300😂
Harfy 🆔	mending beli bensin, ga sampe 10 menit kelar. listrik aman ga jebol
ngokkkkk	kalo mobil listrik setuju,tapi kalo motor.....
Icha rahayu Skincare	masih belom rekomen bgt 🙏

3.2 Prosedur Penelitian

Prosedur penelitian ini dilakukan melalui beberapa tahapan. Dimulai dengan pengumpulan data (*data crawling*), *pre-processing data*, *processing data*, *labelling*, *modelling*, dan evaluasi. Prosedur penelitian ditunjukkan pada Gambar 3.1.



Gambar 3. 1 Prosedur Penelitian

3.2.1 Pengumpulan Data

Data dikumpulkan menggunakan metode *crawling*, yaitu teknik yang digunakan untuk mengakses dan mengumpulkan informasi dari web, atau berbagai sumber lainnya di internet (Riskiyah et al., 2024). *Crawling* bekerja secara otomatis, mengumpulkan informasi berdasarkan kata kunci yang ditentukan oleh pengguna (Maria et al., 2023). Teknik ini banyak dimanfaatkan dalam penelitian berbasis data digital karena mampu mengakses data dalam jumlah besar secara efisien dan akurat.

Proses *crawling* dilakukan untuk menemukan komentar di TikTok dengan kata kunci "mobil listrik". Data teks komentar berhasil dikumpulkan dengan total sebanyak 2263 data. Data tersebut kemudian menjadi sumber utama dalam analisis sentimen untuk mengetahui persepsi pengguna TikTok terhadap mobil listrik di Indonesia.

3.2.2 *Data Pre-processing*

Pre-processing dalam *machine learning* meliputi pembersihan dan normalisasi data untuk mempermudah tahap pemrosesan berikutnya (Meilana, 2023). Langkah ini bertujuan mengubah data mentah menjadi data yang siap untuk analisis lebih lanjut (Riyadi et al., 2021).

Dalam penelitian ini, data hasil *crawling* dari platform TikTok terlebih dahulu dinormalisasi dengan mengubah seluruh karakter menjadi huruf kecil (*lowercase*) serta menghapus komentar yang mengandung nilai *null* maupun data duplikat. Proses ini juga mencakup penghapusan elemen tidak relevan, seperti

simbol, tagar, nama pengguna, URL, dan alamat email karena tidak memberikan kontribusi bermakna terhadap analisis sentimen.

3.2.3 *Data Processing*

Pada tahapan *data processing*, terdapat tiga langkah utama yang dilakukan, yaitu *stemming*, *tokenization*, dan *stopword removal*. Langkah-langkah tersebut bertujuan untuk mengoptimalkan pengolahan data teks secara efisien. Tahapan ini penting untuk memastikan data yang diolah menjadi lebih bersih, terstruktur dengan baik, dan siap untuk analisis lebih lanjut.

Penerapan proses ini menjadi esensial mengingat karakteristik data teks yang tidak terstruktur dan sering mengandung informasi yang berlebihan. Melalui pembersihan data yang sistematis, teks menjadi lebih terfokus, ringkas, dan relevan terhadap tujuan analisis.

3.2.3.1 *Stemming*

Stemming adalah proses mengubah kata-kata menjadi bentuk dasarnya, sehingga berbagai variasi bentuk kata dapat dikenali sebagai kata yang sama (Saputra & Hasan, 2024). Proses ini dilakukan dengan menghapus semua imbuhan (*affixes*) yang melekat pada kata, seperti akhiran (*suffixes*), awalan (*prefixes*), serta gabungan antara awalan dan akhiran (*confixes*) pada kata turunan, sehingga diperoleh kata dasar (*root* atau *stem*) (Sinaga & Nainggolan, 2023). Tahap ini dilakukan untuk meningkatkan efektivitas analisis teks.

Tahapan *stemming* sangat penting dalam pemrosesan bahasa alami karena mampu menyatukan berbagai bentuk kata ke dalam satu representasi standar. Hal ini tidak hanya meningkatkan efisiensi analisis, tetapi juga membantu mengurangi kompleksitas data yang dapat memperberat kinerja model.

3.2.3.2 *Tokenization*

Tokenization merupakan proses memecah kalimat, paragraf, atau halaman menjadi unit-unit kata tunggal yang terpisah dan berdiri sendiri (Agustina et al., 2022). Selain memisahkan teks, proses ini juga dapat menafsirkan dan mengelompokkan token yang terpisah untuk membentuk unit token yang lebih kompleks (Wati et al., 2023).

Tahapan tokenisasi dilakukan untuk mempermudah proses pengolahan teks (Utama et al., 2024). Dengan memecah teks menjadi token, setiap kata atau frasa dapat dianalisis secara individual, sehingga memudahkan proses ekstraksi fitur dan pengembangan model. Selain mempercepat proses analisis, *tokenizatio*n juga membantu meningkatkan konsistensi representasi data, mengurangi ambiguitas bahasa alami, serta membentuk dasar yang kuat untuk tahap *data processing* selanjutnya.

3.2.3.3 *Stopword Removal*

Stopword adalah kata-kata yang tidak penting dan dapat diabaikan dalam pemrosesan bahasa alami (Wibawa et al., 2021). *Stopword removal* bertujuan untuk menghilangkan kata-kata yang tidak memiliki makna penting dalam analisis teks, seperti "dan", "atau", "yang", dan sebagainya. Proses ini umumnya dilakukan setelah tokenisasi, di mana teks sudah dipisah menjadi kata-kata. Selanjutnya, setiap kata diperiksa dan kata-kata yang ada dalam daftar *stopwords* akan dihapus (Toresa et al., 2024).

Setiap token diperiksa dan dibandingkan dengan daftar *stopwords*; jika terdapat kecocokan, token tersebut dihapus dari data. Dengan mengurangi jumlah kata yang tidak relevan, proses ini membantu meningkatkan efisiensi pengolahan data, mempercepat proses analisis, dan dapat berkontribusi pada peningkatan akurasi model.

3.2.4 *Labelling*

Labelling merupakan tahap penting dalam analisis sentimen yang melibatkan proses penetapan label sentimen pada dataset (Pane & Ramdan, 2022), (Oktavia Praneswara & Cahyono, 2023). Dalam *supervised learning*, label merujuk pada hasil atau *output* yang ingin diprediksi oleh model (Amirah & Karimah, 2023). Pada penelitian ini, pelabelan dilakukan untuk mengklasifikasikan komentar pengguna ke dalam kategori sentimen positif, negatif, dan netral.

Proses ini bertujuan untuk menetapkan bobot pada setiap kata dengan menggunakan metode TF-IDF. Kemudian, dilakukan validasi oleh ahli Bahasa Indonesia, yaitu Delvia Pepy Anggraeni, S.Pd. untuk memastikan pelabelan sesuai dengan kategori sentimen yang sebenarnya.

3.2.5 *Modelling*

Pada proses *modelling*, pengujian model klasifikasi dilakukan menggunakan data komentar yang sebelumnya telah di proses dan dilabeli. Data kemudian di *split* menjadi *data training* sebesar 80%, dan *data testing* sebesar 20%. Proses *modelling* melibatkan algoritma Logistic Regression dan SVM. Kedua algoritma tersebut digunakan karena terbukti memberikan hasil yang optimal ketika digunakan dalam analisis teks. Berikut contoh penerapan model Logistic Regression dan SVM pada salah satu komentar dari dataset pada Tabel 3.2.

Tabel 3. 2 Contoh Vektor Kata

No.	Kata	TF-IDF
1.	irit	1.445
2.	bensin	1.122
3.	boros	1.578
4.	listrik	0.594

Maka vektor dari komentar tersebut:

$x = [1.445, 1.122, 1.578, 0.594\ ]$

a. Logistic Regression

Logistic Regression menghitung probabilitas dengan fungsi sigmoid berdasarkan persamaan (2) dan (3):

$$z = B_0 + \sum_{i=1}^{n} B_i X_i$$

$$p = \frac{1}{1 + e^{-z}}$$

Asumsikan bobot model hasil *training* sebagai berikut:

$B_0 = -0{,}5$ (bias)

$B = [0.3, -0.6, -0.7, 0.1]$

Maka:

$$z = -0.5 + (0.3 \times 1.445) + (-0.6 \times 1.122) + (-0.7 \times 1.578) + (0.1 \times 0.594)$$

$$z = -0.5 + 0.4335 - 0.6732 - 1.1046 + 0.0594 = -1.785$$

$$p = \frac{1}{1 + e^{-(1.785)}} = \frac{1}{1 + e^{1.785}} = 0.143$$

Karena $p < 0.5$, maka komentar diklasifikasikan sebagai sentimen negatif.

b. Support Vector Machine

Pada implementasi SVM linear, fungsi kernel pada Tabel 2.1:

$$K(x, y) = x.y$$

Diterapkan dalam bentuk fungsi keputusan persamaan (4):

$$f(x) = w \times x + b$$

Asumsikan parameter model hasil *training* adalah:

$w = [0.5, -0.8, -1, 0.2]$ dan $b = -0.3$

Maka:

$$f(x) = (0.5 \times 1.445) + (-0.8 \times 1.122) + (-1 \times 1.578) + (0.2 \times 0.594) - 0.3$$

$$f(x) = 0.7225 - 0.8976 - 1.578 + 0.1188 - 0.3 = -1.9343$$

Hasilnya, $f(x) < 0$, maka komentar diklasifikasikan sebagai negatif.

3.2.6 Evaluasi

Pada tahap evaluasi, performa model dinilai berdasarkan kemampuannya dalam mengklasifikasikan data. *Confusion matrix* digunakan untuk menghitung performa kedua model yang diterapkan setelah tahap *modelling* selesai. *Confusion matrix* menentukan performa model dengan nilai-nilai yang ada di dalamnya, yaitu *accuracy*, *precision*, *recall*, *dan F1-Score*.

Data yang telah dipisahkan sebelumnya menjadi *data train* dan *data test* kemudian digunakan untuk menguji kinerja model. Melalui pengujian ini, diperoleh nilai *accuracy*, *precision*, *recall*, dan *F1-Score*, yang masing-masing mencerminkan aspek berbeda dari performa model dalam mengklasifikasikan data. Evaluasi menggunakan metrik ini penting untuk memahami kelebihan dan kelemahan model secara lebih menyeluruh, sehingga dapat menjadi dasar dalam melakukan perbaikan atau optimasi model.