BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Gojek merupakan aplikasi layanan transportasi *online* berdasarkan permintaan yang berbasis utama di Indonesia. Berdiri sejak tahun 2010, Gojek menjadi raja layanan transportasi *online* di Indonesia sampai sekarang. Kendati demikian, pada tahun 2021 sampai 2023, terjadi penurunan jumlah pengguna Gojek. Pada penelitian ini, akan dilakukan prediksi mengenai keberlanjutan Gojek dimasa mendatang.

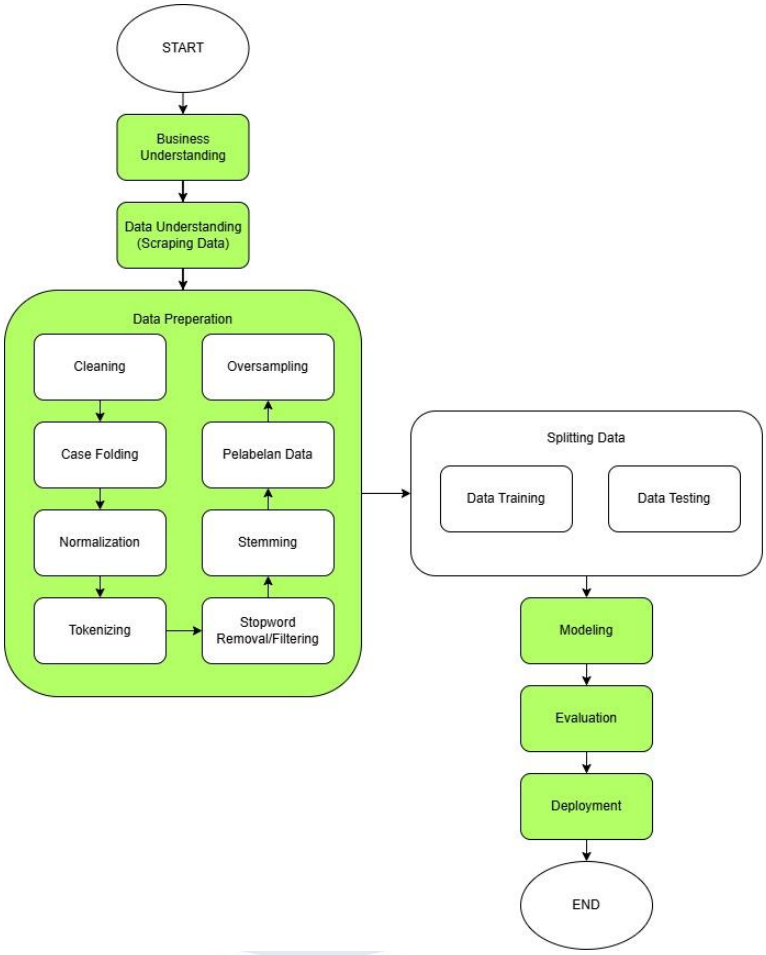
Studi ini akan menggunakan data ulasan dari pengguna Gojek yang diambil dari *platform Google Play Store*. Platform tersebut dipilih karena merupakan platform untuk mengunduh suatu aplikasi dan terdapat review dari suatu aplikasi sehingga data yang dihasilkan dapat terfokus pada review aplikasinya saja. Data tersebut diambil dengan cara data scraping mengenai review gojek yang ada di *Google Play Store*.

Studi ini menerapkan tiga algoritma dalam melakukan analisis sentimen kepuasan pelanggan yaitu algoritma *Naïve Bayes*, *K-Nearest Neighbors* (KNN), dan *Support Vector Machine* (SVM). Data yang ada nantinya akan diproses dengan ketiga algoritma tersebut yang nantinya berpedoman pada CRISP-DM (*Cross Industry Standard Process – Data Mining*). *Output* dari analisis sentimen tersebut nantinya akan menjadi pedoman masyarakat untuk mengetahui keberlanjutan Gojek di pasar transportasi *online* Indonesia.

3.2 Metode Penelitian

3.2.1 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 3.1.



Gambar 3. 1 Alur Penelitian

3.2.2 Metode *Data Mining*

Teknik *data mining* yang dilakukan dalam studi ini ialah perbandingan algoritma. Beberapa algoritma yang dibandingkan dalam studi ini adalah *Naive Bayes*, KNN, dan SVM, yang tujuannya menentukan algoritma yang paling efektif. Setiap algoritma memiliki kelebihan dan kekurangan yang berbeda-beda, yang dipaparkan pada Tabel 3.1.

Tabel 3. 1 Perbandingan Algoritma Penelitian

	Naive Bayes	***K-Nearest Neighbors* (KNN)**	***Support Vector Machine* (SVM)**
Konsep	Algoritma yang didasarkan pada probabilitas dan menggunakan Teorema Bayes dengan asumsi bahwa fitur-fitur saling independen.	Algoritma berbasis instance yang mengklasifikasikan data berdasarkan kedekatan atau jarak ke titik data lain yang telah dilabeli.	Algoritma yang berupaya menemukan hyperplane optimal yang memisahkan kelas-kelas dalam ruang fitur dengan margin terluas.
Kelebihan	1. Cepat dan efisien untuk data dalam jumlah besar. 2. Dapat bekerja dengan baik bahkan dengan data yang sedikit. 3. Mudah untuk diimplementasikan.	1. Tidak memerlukan asumsi distribusi data. 2. Sederhana dan intuitif. 3. Dapat bekerja dengan baik untuk *dataset* dengan batasan non-linear.	1. Efektif pada ruang fitur berdimensi tinggi. 2. Dapat bekerja dengan baik untuk data dengan margin yang jelas. 3. Memiliki kemampuan generalisasi yang baik.
Kekurangan	1. Asumsi independensi antar fitur jarang terjadi di dunia nyata. 2. Sensitif terhadap data yang tidak relevan atau fitur yang berkorelasi.	1. Memiliki kinerja yang lambat pada *dataset* besar karena perlu menghitung jarak ke semua titik data lain. 2. Memori intensif karena perlu menyimpan semua data	1. Kurang efisien pada *dataset* yang sangat besar. 2. Bisa sulit diinterpretasikan karena kompleksitas model.

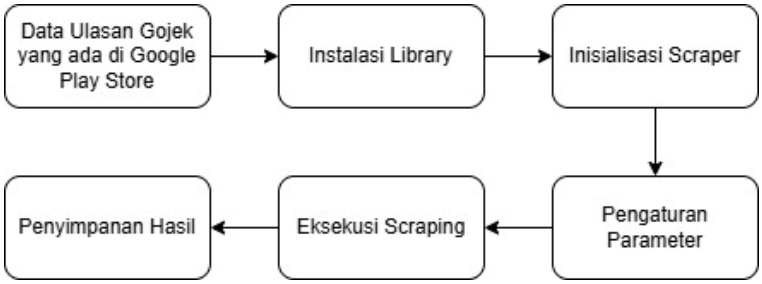
		pelatihan. 3. Kinerja sangat bergantung pada pemilihan nilai k dan metrik jarak.	

3.3 Teknik Pengumpulan Data

Dalam penelitian ini, data dikumpulkan menggunakan menggunakan data primer yang diperoleh langsung dari sumber aslinya [6], yaitu ulasan pengguna Gojek yang tersedia di *Google Play Store*. Proses pengumpulan data dilakukan menggunakan teknik *web Scraping* dengan bantuan *library google_play_scraper* dimana program dapat mengekstrak data dari halaman web *Google Play Store* secara sistematis. *Dataset* yang diambil sebesar 10.000 [22], seperti yang ada pada penelitian terdahulu.

Tahap pertama dalam pengumpulan data adalah melakukan instalasi *library* yang dibutuhkan. *Pandas* dan *google-play-scraper* akan menjadi *library* yang digunakan pada tahap ini. *Library Pandas* digunakan untuk manipulasi, analisis, dan pengolahan data. *Library google-play-scraper* digunakan untuk mengumpulkan data aplikasi dari *Google Play Store*.

Tahap selanjutnya adalah inisialisasi *scraper* dilakukan dengan melakukan *import library* yang telah diunduh sebelumnya. Setelah itu, akan dilakukan pengaturan parameter yang terdiri dari ID aplikasi, bahasa, dan negara. Setelah parameter berhasil diatur, *scraping* akan dieksekusi sesuai dengan parameter yang telah ditentukan sebelumnya. Hasil dari *scraping* tersebut akan disimpan pada *data frame* dan akan disimpan dalam format CSV.



Gambar 3. 2 Tahapan Pengumpulan Data

3.4 Variabel Penelitian

Variabel merupakan elemen kunci yang menentukan arah dan tujuan dari studi yang dilakukan. Terdapat 2 variabel penelitian dalam studi ini yaitu:

1. Variabel Independen

   Variabel independen pada studi ini ialah opini atau *review* pengguna Gojek yang diambil dari *platform Google Play Store*.

2. Variabel Dependen

   Variabel dependen pada studi ini ialah sentimen kepuasan pelanggan. Sentimen akan dikategorikan sebagai positif, negatif, dan netral [13]. Ketiga kategori tersebut digunakan untuk meningkatkan akurasi pada algoritma karena dapat mengurangi bias pada klasifikasi yang dilakukan.

3.5 Teknik Analisis Data

Pada studi ini, terdapat beberapa tahap untuk memproses, menganalisa, serta memprediksi data. Langkah pertama dalam penelitian ini adalah pengumpulan data ulasan aplikasi Gojek dengan *library google-play-scraper*. *Library* ini digunakan untuk mengambil data *review* pengguna secara otomatis dari *Google Play Store*, termasuk informasi seperti rating, ulasan teks, tanggal ulasan, dan versi aplikasi. Data yang berhasil dikumpulkan akan disimpan dalam format CSV guna memudahkan pengolahan selanjutnya.

Tahap berikutnya adalah persiapan data yang mencakup beberapa langkah yaitu *cleaning, Case Folding, Normalization, Tokenizing, Stopword Removal*, dan *Stemming*. *Cleaning* merupakan proses menghilangkan atau memperbaiki data yang kotor, seperti menghapus duplikasi, mengisi data yang hilang, atau menghapus karakter khusus seperti *emoji, symbol* atau tanda baca, dan menghapus angka. *Case folding* merupakan proses mengubah seluruh huruf menjadi huruf kecil. *Normalization* adalah proses standar yang memiliki tujuan merubah teks ke bentuk dasar dan mengubah kata-kata tidak baku menjadi kata baku dengan menggunakan kamus kata baku yang diambil dari situs penyedia dataset [36]. *Tokenizing* adalah proses memecah teks menjadi unit-unit kecil. *Stopword Removal* adalah proses penghapusan kata-kata umum yang tidak memiliki makna. *Stemming* adalah proses mengubah kata-kata ke bentuk dasarnya.

Setelah data melewati tahap persiapan data, langkah selanjutnya adalah pemberian label terhadap opini yang telah melewati tahap persiapan data. Pemberian label ini menggunakan *Lexicon Based* dengan bantuan kamus *Indonesia Lexicon Based* yang dibuat oleh pakar [37]. Opini akan dihitung berdasarkan poin yang ada pada kamus leksikon yang kemudian poin tersebut akan menentukan label pada setiap opini.s

Tahap berikutnya adalah pembagian data atau *data splitting.* Pada tahap ini, *dataset* akan dibagi menjadi dua bagian yaitu data latih dan data uji. Sebanyak 80% data digunakan untuk melatih model, sedangkan 20% sisanya digunakan untuk menguji kinerja model. Pembagian ini dilakukan secara acak untuk memastikan distribusi data yang merata pada kedua set.

Data yang sudah dibagi, akan memasuki tahap modeling sesuai dengan algoritma yang digunakan. Pada tahap ini, data ulasan yang telah dilabeli digunakan untuk melatih model analisis sentimen dengan algoritma *Naïve Bayes*, KNN, dan SVM. Algoritma *Naïve Bayes* digunakan karena kemampuannya dalam menangani data teks, KNN untuk pendekatan berbasis jarak, dan SVM untuk klasifikasi yang lebih presisi dengan menggunakan

kernel tertentu. Setiap model dioptimalkan melalui proses validasi untuk mendapatkan hasil terbaik.

Tahap berikutnya adalah melakukan evaluasi terhadap model yang sudah dibuat dengan tiga algoritma yang berbeda. Model yang telah diimplementasikan akan dievaluasi dengan metrik seperti akurasi, *precision*, *recall*, dan *F1-score*. Evaluasi ini dilakukan untuk membandingkan performa ketiga algoritma dalam melakukan analisis sentimen *review*. Hasil evaluasi diperlihatkan dalam bentuk tabel dan grafik guna memudahkan interpretasi dan analisis.

Tahap terakhir pada analisis data pada studi ini adalah mengembangkan aplikasi berbasis web untuk analisis sentimen menggunakan *Streamlit*. Aplikasi ini akan menampilkan *dashboard* analisis sentimen. Aplikasi ini berguna untuk mempermudah pengguna melihat data yang sudah dilabeli sehingga dapat membantu pengguna dalam implementasi model untuk analisis sentimen.