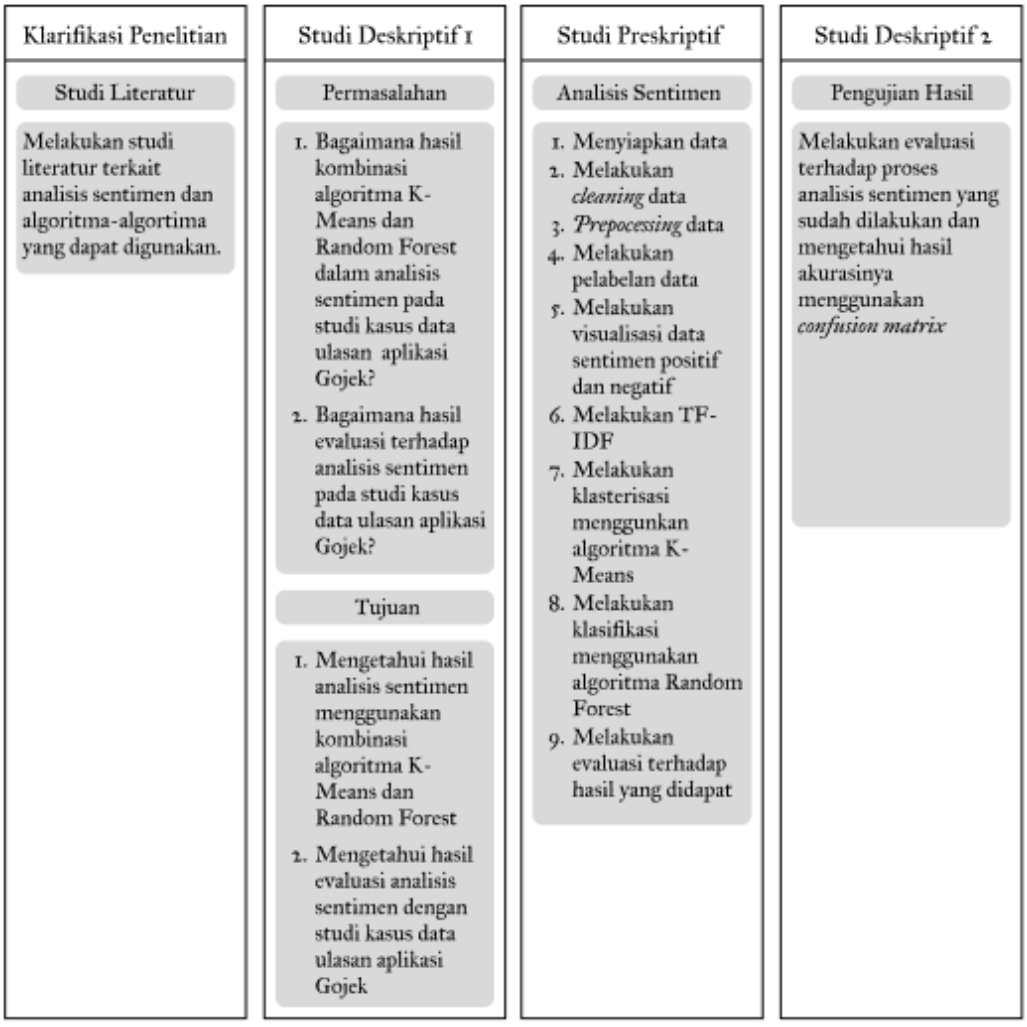
BAB III
METODE PENELITIAN

3.1 Desain Penelitian

Desain penelitian yang digunakan pada penelitian ini adalah *Design Research Methodology*. Metode ini adalah sebuah pendekatan yang terdiri dari serangkaian metode dan panduan yang berfungsi sebagai kerangka kerja dalam melaksanakan penelitian desain (Pratama & Sukma, 2024).



Gambar 3.1

Desain Penelitian

3.1.1 Klasifikasi Penelitian

Pada tahap awal ini terdapat proses studi literatur terkait dengan analisis sentimen. Proses ini dapat membantu dalam menentukan permasalahan

penelitian, tujuan penelitian, agar dapat digunakan sebagai latar belakang penelitian yang akan dilakukan. Studi literatur yang dilakukan terkait tentang analisis sentimen dan algoritma-algoritma *machine learning* yang dapat digunakan untuk melakukan proses analisis sentimen.

3.1.2 Studi Deskriptif 1

Pada tahap ini melakukan pembuatan rancangan permasalahan penelitian dan juga tujuan dari penelitian itu sendiri. Setelah mempertimbangkan hasil dari studi literatur yang sebelumnya telah dilakukan maka permasalahan yang akan diangkat pada penelitian ini adalah bagaimana hasil kombinasi algoritma *K-Means* dan *Random Forest* dalam analisis sentimen pada studi kasus data ulasan aplikasi Gojek serta bagaimana hasil evaluasi terhadap analisis sentimen pada studi kasus data ulasan aplikasi Gojek. Tujuan penelitiannya adalah mengetahui hasil analisis sentimen menggunakan algoritma *K-Means* dan *Random Forest* serta mengetahui hasil evaluasi analisis sentimen dengan studi kasus data ulasan aplikasi Gojek.

3.1.3 Studi Preskriptif

Pada tahapan ini melakukan analisis sentimen. Tahapan pertama yaitu menyiapkan data yang akan digunakan, pada penelitian ini data yang akan digunakan adalah data ulasan pengguna aplikasi Gojek yang didapatkan dari situs Kaggle. Setelah itu melakukan *cleaning* data dari data yang sudah didapatkan akan dilakukan beberapa proses agar mendapatkan data yang bersih untuk digunakan dalam penelitian (Safitri et al., 2021), proses *cleaning* data tersebut seperti penghapusan data yang duplikat, penghapusan data yang kosong, dan penghapusan kolom username (Lubis & Yudertha, 2024). Selanjutnya melakukan *preprocessing* data yang terbagi menjadi beberapa tahapan, yaitu (Herjanto & Carudin, 2024) :

a. *Case Folding* merupakan proses merubah seluruh kalimat menjadi *lower case*.
b. *Tokenizing* merupakan proses pemisahan kata sesuai dengan spasi yang ada.

c. *Filtering* merupakan proses penghapusan kata-kata yang tidak diperlukan atau memiliki nilai informasi yang rendah, seperti kata yang, di, dan yang lainnya.
d. *Stemming* merupakan proses perubahan suatu kata menjadi kata dasarnya.

Tahapan selanjutnya adalah melakukan pelabelan terhadap data apakah data tersebut termasuk ke dalam label positive, negative, atau netral. Proses tersebut menggunakan Lexicon Base, dimana proses pembobotannya tersebut menggunakan kamus lesikon. Setiap kata akan dilakukan perhitungan *score* sentimen, jika mendapatkan *score* ≥ 1 maka termasuk ke dalam sentimen *positive*, *score* = 0 termasuk ke dalam netral, dan *score* ≥ -1 termasuk ke dalam *negative* (Manullang et al., 2023). Kamus yang digunakan pada penelitian ini adalah kamus sentistrength_id dimana memiliki nilai antara -5 hingga +5 (Abdillah et al., 2021)

Setelah itu dilakukan pembobotan kata menggunakan TF-IDF yang akan digunakan untuk proses analisis selanjutnya (Wicaksono et al., 2023). Pembobotan kata menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) ini memiliki fungsi untuk merubah data teks menjadi numerik, hasil yang didapatkan dari TF-IDF adalah dapat mengidentifikasi kata-kata penting yang terdapat pada suatu dokumen. Penjelasan dari masing-masing perhitungannya dapat dilihat sebagai berikut (Wati et al., 2023):

a. *Term Frequency* (TF)

   Perhitungan ini menghasilkan nilai frekuensi dari setiap kemunculan suatu kata yang terdapat dalam dokumen. Untuk perhitungannya dapat menggunakan rumus sebagai berikut:

$$TF = \frac{(jumlah\ kemunculan\ kata\ pada\ dokumen)}{(jumlah\ kata\ pada\ dokumen)}$$

b. *Inverse Document Frequency* (IDF)

   Perhitungan ini merupakan kebalikan dari perhitungan TF, karena pada perhitungan IDF ini nantinya akan menghasilkan nilai frekuensi atau bobot terhadap tiap kata yang kemunculan kata tersebut jarang pada suatu dokumen. Untuk perhitungannya dapat menggunakan rumus sebagai berikut:

$$IDF = log\frac{N}{n}$$

Keterangan:

N = Jumlah dokumen yang terdapat pada dokumen

n = Jumlah dokumen yang mempunyai kata tersebut

Nantinya, jika kedua nilai tersebut sudah didapatkan akan dilakukan perkalian antara nilai TF dan IDF. Bila nilai yang didapatkan tinggi, hal tersebut mencerminkan bahwa kata tersebut dianggap penting dan memberikan kontribusi yang besar dalam penentuan suatu topik di dalam suatu dokumen.

Proses analisis lebih lanjut menggunakan algoritma K-Means untuk melakukan clustering, dimana algoritma K-Means ini akan mengelompokkan data berdasarkan kemiripan pola yang dimiliki setiap data, dapat mengurangi ambiguitas pada data, dan dapat memberikan representasi data yang lebih terstruktur sehingga dapat meningkatkan nilai akurasi pada klasifiaksi dan analisis sentimen (Iparraguirre-Villanueva et al., 2022). Kemudian algoritma *Random Forest* itu sendiri digunakan untuk melakukan klasifikasi, pada proses klasifikasi tersebut data akan dibagi menjadi membagi 80% data untuk pelatihan *(training)* dan 20% data untuk pengujian *(testing)* seperti pada jurnal (Manullang et al., 2023) yang digunakan sebagai acuan. Setelah itu dilakukan proses evaluasi dilakukan dengan menggunakan *confusion matrix* untuk menilai tingkat akurasi terhadap proses anlisis sentimen (Que et al., 2020).

3.1.4 Studi Deskriptif 2

Pada tahapan terakhir ini melakukan evaluasi terhadap hasil analisis sentimen yang sudah dilakukan. Evaluasi yang dilakukan menggunakan *confusion matrix* yang berisikan nilai prediksi dan nilai sebenarnya, *confusion matrix* juga memberikan rangkuman kinerja dari proses klasifikasi dan menampilkannya ke dalam bentuk visual (Sani et al., 2022). Terdapat istilah-istilah yang perlu diketahui seperti *True Positive* (TP), *True Negative* (TN), *False Positif* (FP), dan *False Negative* (FN) digunakan untuk melakukan representasi dari hasil proses klasifikasi yang sudah dilakukan dan berguna untuk dapat mengetahui jumlah hasil prediksi yang benar dan berapa banyak yang salah dari model tersebut (Weny Indah Kusumawati & Adisaputra Zidha Noorizki, 2023).

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini merupakan data ulasan aplikasi Gojek yang dari situs Kaggle. Data tersebut diambil dari bulan November tahun 2021 hingga Februari tahun 2024 sebanyak 225.002 data yang terkumpul. Tetapi setelah melakukan proses *cleaning* data, hanya terdapat 121.721 data bersih yang akan digunakan dalam penelitian. Dataset keseluruhan berisikan username, content, score, at, dan appVersion. Berikut pada tabel 3.1 merupakan contoh data yang akan digunakan dalam penelitian ini.

Tabel 3.1 Data ulasan aplikasi Gojek

No	username	content	score	at	appVersion
1	username1	akun gopay saya di blok	1	2022-01-21 10:52:12	4.9.3
2	username2	Baru download gojek dan hape baru trus ditop u gopay transaksi dialfamart transaksi bloked transaksilaporan di pusat bantuan gak jelas yang ditanyakan apa jawaban lainlama lama gojek dikelola Tokopedia udah nyimpangapa gojek anak bangsa seperti dulu apa punya Tokopedia	1	2022-09-03 15:21:17	4.9.3
3	username3	Mantap	5	2022-01-15 10:05:27	4.9.3

4	username4	Coba dulu	2	2021-12-10 22:40:45	4.9.3
5	username5	Gimana ini kak pin saya salah terus padahal udah di ubah masih salah	1	2022-12-17 08:56:52	4.9.3

3.3 Alat dan Bahan Penelitian

Dalam penelitian ini, alat yang digunakan meliputi perangkat keras dan perangkat lunak. Perangkat keras atau *hardware* yang digunakan sebagai berikut:

1. Processor Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.20 GHz
2. RAM 8,00 GB
3. Monitor 15.6” dengan resolusi 1920x1080 pixel

Selain itu untuk perangkat lunak atau *software* yang digunakan sebagai berikut:

1. *Operating System*: Microsoft windows 11
2. *Tools Editor*: Google Colaboratory
3. Bahasa Pemrograman: Phyton versi 3.10.12
4. *Library* Phyton:
    - Pandas
    - Regular Expressions (re)
    - Natural Language Toolkit (nltk)
    - Sastrawi
    - WordCloud
    - Stemmer Factory
    - Stopwords
    - Numpy
    - Matplotlib.pyplot

Sedangkan bahan penelitian yang digunakan adalah data ulasan pengguna aplikasi Gojek yang terdapat di google play store, data tersebut diambil dari situ Kaggle. Data tersebut diambil dari bulan November tahun 2021 hingga Februari tahun 2024.

3.4 Instrumen Penelitian

Instrumen penelitian yang digunakan ialah *Confusion Matrix* yang digunakan untuk melakukan evaluasi dan mengukur hasil dari model klasifikasi machine learning dari segi peformanya (Safitri et al., 2021). Penilaian yang dilakukan pada *confusion matrix* nantinya dapat digunakan untuk menentukan nilai *Accuracy, Precision, Recall,* dan *F1 Score*. Tabel untuk *confusion matrix* dapat dilihat pada tabel 3.2 berikut.

Tabel 3.2 *Confusion Matrix*

Predicted Class	*Observed*	
	True	*False*
True	*True Positive* (TP)	*False Positive* (FP)
False	*False Negative* (FN)	*True Negative* (TN)

Dimana keterangan untuk tabel 3.2 sebagai berikut:

1. *True Positive* (TP), jumlah data dengan nilai positif yang diklasifikasikan dengan benar oleh sistem.
2. *True Negative* (TN), jumlah data dengan nilai negatif yang diklasifikasikan dengan benar oleh sistem.
3. *False Negative* (FN), jumlah data dengan nilai negatif yang diklasifikasikan secara salah oleh sistem.
4. *False Positive* (FP), jumlah data dengan nilai positif yang diklasifikasikan secara salah oleh sistem.

Sehingga untuk mengetahui nilai *accuracy, precision, recall,* dan *f1 score* yang digunakan sebagai parameter penelitian dapat dilakukan perhitungan sebagai berikut (Arminda et al., 2023):

a. *Accuracy* adalah perbandingan nilai antara nilai prediksi dengan nilai yang sebenarnya. Sehingga jika nilai yang didapatkan tinggi, sistem tersebut akan semakin bagus dalam melakukan prediksi. Perhitungan *accuracy* tersebut dapat dihitung menggunakan rumus berikut:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \times 100\%$$

b. *Precision* adalah nilai yang didapatkan dari segi tingkat kesesuaian antara informasi yang diberikan oleh sistem dan yang diminta oleh pengguna. Perhitungan *precision* tersebut dapat dihitung menggunakan rumus berikut:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \times 100\%$$

c. *Recall* adalah perhitungan terkait dengan keakuratan prediksi, dimana nilai yang didapatkan digunakan sebagai tolak ukur kerberhasilan sistem untuk memperoleh kembali sebuah informasi. Perhitungan *recall* tersebut dapat dihitung menggunakan rumus berikut:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \times 100\%$$

d. *F1 Score* adalah perhitungan untuk menentukan rata-rata dari *precision* dan *recall* untuk dapat memberikan keseluruhan gambaran terkait kinerja model. *F1 Score* dapat dihitung menggunakan rumus berikut:

$$F1\ Score = 2\ \frac{(precision \times recall)}{precision + recall} \times 100\%$$