

BAB II

METODOLOGI PENELITIAN

2.1. Obyek Penelitian

Objek penelitian dalam studi ini adalah ulasan pengguna aplikasi DANA yang tersedia di Google Play Store. Penelitian ini bertujuan untuk menganalisis sentimen ulasan pengguna terhadap layanan dompet digital DANA dengan menggunakan dua metode klasifikasi, yaitu K-Nearest Neighbor (KNN) dan Naive Bayes. Data yang digunakan berasal dari teknik *web scraping* menggunakan bahasa pemrograman Python, yang mengumpulkan sebanyak 700 ulasan dalam bahasa Indonesia.

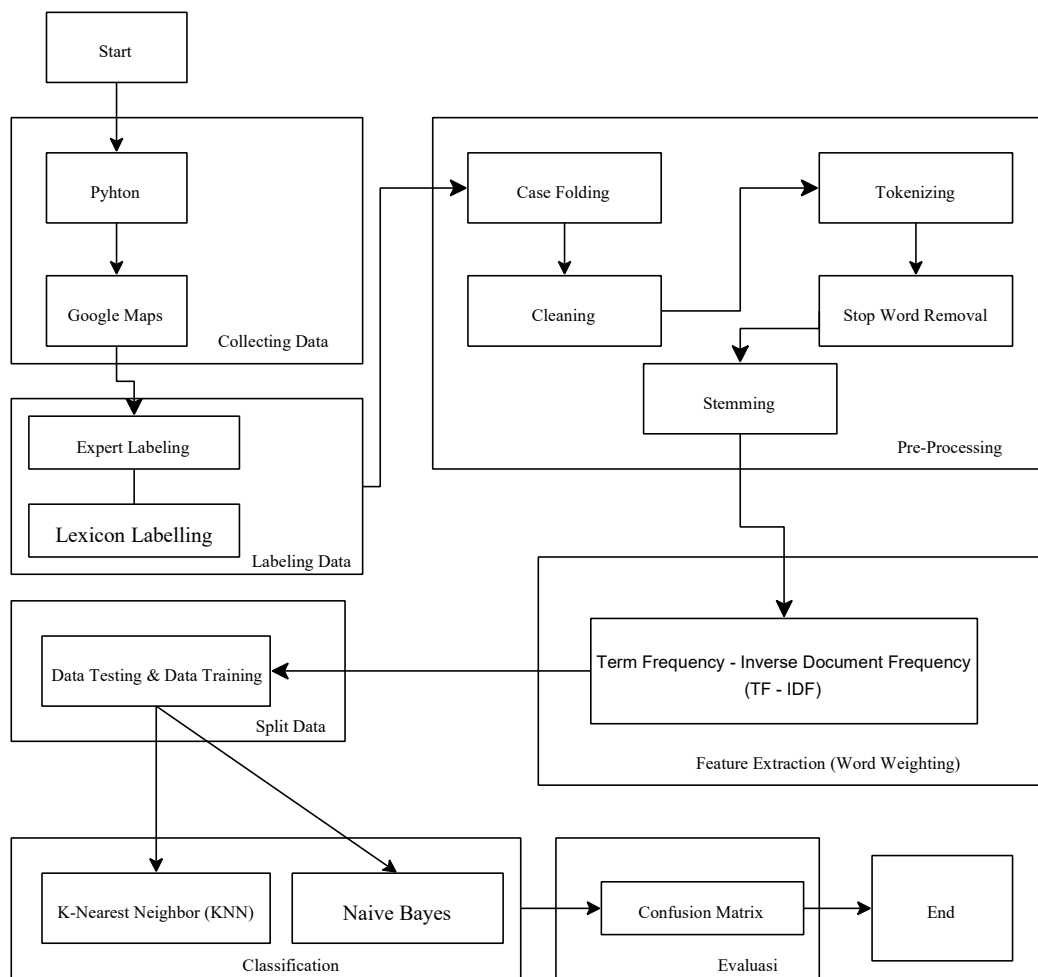
Teknik *web scraping* akan digunakan untuk mengumpulkan data ulasan pengguna, yang kemudian akan diproses melalui beberapa tahap seperti *Case Folding*, *Cleaning*, *Tokenizing*, *Stopword Removal*, dan *Stemming* sebelum dilakukan klasifikasi. Data yang telah diproses kemudian akan dianalisis untuk mengklasifikasikan sentimen menjadi kelas positif, negatif, dan netral. Evaluasi model akan dilakukan menggunakan metrik akurasi. Penelitian ini diharapkan dapat memberikan wawasan yang berguna bagi peningkatan kualitas layanan aplikasi DANA berdasarkan analisis sentimen dari ulasan pengguna.

2.2. Alat Dan Bahan

Dalam penelitian ini, alat dan bahan yang digunakan meliputi perangkat keras Laptop dengan spesifikasi prosesor intel CORE i5 Gen 10, RAM 4GB, dan penyimpanan SSD 512GB. Untuk perangkat lunak digunakan Google Colab versi 1.0.0 (<https://colab.research.google.com/>) dengan Python versi 3.8.10. Library Python yang digunakan mencakup Google-play-scraper-1.2.7, Pandas versi 2.0.3, NumPy versi 1.25.2, NLTK versi 3.8.1, Scikit-learn versi 1.2.2, Matplotlib versi 3.7.1, serta Sastrawi versi 1.0.1.

2.3. Prosedur Penelitian

Dalam penelitian ini, dilakukan teknik *web scraping* pada Google Play Store dengan menggunakan bahasa pemrograman Python kemudian mengambil link aplikasi DANA untuk proses analisis lebih lanjut. Metode yang dipakai dalam penelitian ini bertujuan untuk melakukan analisis sentimen terhadap data pengguna aplikasi DANA di Google Play Store dengan membandingkan dua metode, yaitu K-Nearest Neighbor dan Naive Bayes. Diagram alur penelitian pada Gambar 2.1.



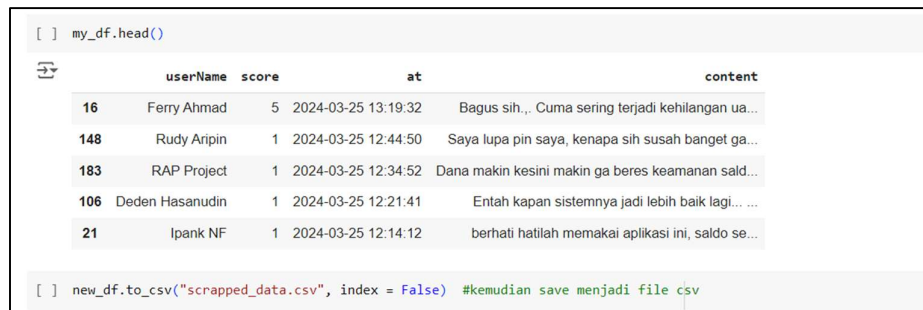
Gambar 2.1 Alur Penelitian

Pada Metode Penelitian terdiri dari beberapa tahapan utama, mulai dari Pengumpulan data melalui *scrapping* menggunakan Python, Labeling Data oleh ahli bahasa (*expert*)

dan menggunakan model *lexicon based*, Pre-processing, Pembobotan kata (TF-IDF), Pembagian data (*Split data*), Klasifikasi menggunakan metode KNN dan Naive Bayes, serta Evaluasi kinerja model menggunakan *confusion matriks*. Setiap tahapan akan dijelaskan secara rinci dengan memaparkan teori, konsep, dan algoritma yang digunakan.

2.3.1. Pengumpulan Data (*Scrapping*)

Dalam penelitian ini pengumpulan data dengan menggunakan metode teknik *web scraping* yang bersumber dari kolom komentar pada Google Play Store DANA. dengan menggunakan metode *web scraping* yaitu suatu kemampuan untuk pengambilan data dalam jumlah besar dari sebuah website. Proses *scrapping* dilakukan dengan mengambil label nama, waktu, komentar, dan jumlah data yang diambil sebanyak 700 data. Setelah proses *scrapping* selesai, Data yang berhasil diekstrak dapat disimpan dalam format file CSV atau Excel (Saputro et al., 2023) Hasil crawling data pada Gambar 2.2.



```
[ ] my_df.head()
```

	userName	score	at	content
16	Ferry Ahmad	5	2024-03-25 13:19:32	Bagus sih,.. Cuma sering terjadi kehilangan ua...
148	Rudy Aripin	1	2024-03-25 12:44:50	Saya lupa pin saya, kenapa sih susah banget ga...
183	RAP Project	1	2024-03-25 12:34:52	Dana makin kesini makin ga beres keamanan sald...
106	Deden Hasanudin	1	2024-03-25 12:21:41	Entah kapan sistemnya jadi lebih baik lagi... ..
21	lpank NF	1	2024-03-25 12:14:12	berhati hatilah memakai aplikasi ini, saldo se...

```
[ ] new_df.to_csv("scrapped_data.csv", index = False) #kemudian save menjadi file csv
```

Gambar 2.2 Hasil Pengumpulan Data

2.3.2. Labelling Data

Dalam proses klasifikasi teks pada data komentar untuk skripsi tugas akhir, peneliti memerlukan ahli bahasa yang berpengalaman dalam pelabelan data dan memiliki pengetahuan mendalam tentang bahasa Indonesia. Untuk itu, peneliti mengajukan permintaan di *website* project.co.id untuk mencari tenaga ahli yang memenuhi kriteria tersebut. Dalam pengajuan ini, peneliti menjelaskan bahwa diperlukan lulusan jurusan Bahasa Indonesia yang saat ini bekerja di bidang terkait, seperti guru/dosen bahasa Indonesia, penulis, atau ahli bahasa. Calon tenaga ahli diminta untuk mengajukan penawaran dengan menyertakan pekerjaan saat ini, pengalaman relevan dengan bahasa Indonesia, serta gelar akademik yang dimiliki. Hasil pelabelan data *expert* pada Gambar 2.3.

1	User Nama	Waktu	Komentar	SENTIMEN
2	Nitong99	2024-03-25 08:47:39	Buat APLIKASI DANA kebanyakan update, tapi masih lola lambat saat login m.	NEGATIF
3	Suci Afriany	2024-03-24 03:52:21	Aplikasinya membantu, tapi cape setiap mau transaksi system error, dan nyo	NEGATIF
4	M Hasan	2024-03-18 02:34:05	Lagi dan lagi, di saat pembayaran e-wallet ke shopeepay bermasalah lagi ales:	NEGATIF
5	DMS ALFARIZI	2024-03-15 03:32:58	Untuk fitur seperti transfer,nabung,inves,transaksi dll semua sudah oke men	POSITIF
6	Agung Nursetyav	2024-03-18 14:08:24	Brengsek isi saldo malah gak bisa dipake gara² minta upgrade premium. Coba	NEGATIF
7	Y	2024-03-22 17:54:04	Sebenarnya aplikasinya sudah bagus enak buat transaksi gak pake lama udah	NEGATIF
8	Vemi Love	2024-03-23 04:23:27	Sebelum update tidak ada masalah dan sangat membantu dalam bertransaks	NEGATIF

Gambar 2.3 Pelabelan Data Expert

Setelah mendapatkan hasil pelabelan data menggunakan *expert*, peneliti juga menggunakan model *lexicon based* untuk memberikan label secara otomatis. *Lexicon* adalah metode yang menggunakan pendekatan kamus untuk menganalisis sentimen. Kamus *lexicon* memproses bobot kata berdasarkan kamus yang memudahkan klasifikasi otomatis kalimat yang akan dianalisis. Oleh karena itu, *lexicon* dapat digunakan untuk melabeli data sebelum dilakukan pelatihan model (Fernanda & Fathoni, 2024). Hasil pelabelan data *lexicon* pada Gambar 2.4.

1	User Nama	Waktu	Komentar	Positif	Negatif	Netral	Compound sentiment
2	Nitong99	2024-03-25 08:47:39	Buat APLIKASI DANA	0,112	0	0,888	0,6981 Positif
3	Suci Afriany	2024-03-24 03:52:21	Aplikasinya membar	0	0,153	0,847	-0,6597 Negatif
4	M Hasan	2024-03-18 02:34:05	Lagi dan lagi, di saat	0	0	1	0 Netral
5	DMS ALFARIZI	2024-03-15 03:32:58	Untuk fitur seperti tr	0	0	1	0 Netral

Gambar 2.4 Pelabelan Data Lexicon

2.3.3. *Pre-Processing*

Sebelum melakukan *text mining* dokumen teks perlu disiapkan terlebih dahulu agar dapat digunakan dalam proses utama persiapan ini dikenal sebagai tahap *text preprocessing*, *Text preprocessing* berperan dalam mengubah data mentah yang tidak terstruktur menjadi terstruktur (Khairunnisa et al., 2021). Langkah-langkah yang dilakukan dalam tahap *preprocessing* (Dwiki et al., 2021) :

- a) *Case Folding* adalah langkah dalam pemrosesan teks yang bertujuan untuk mengubah semua karakter menjadi huruf kecil.
- b) *Cleaning* adalah proses membersihkan dokumen dan menyeleksi kata yang tidak diperlukan seperti menghapus HTML, emoticon, hashtag, mention, dan URL dari teks.
- c) *Tokenizing* adalah proses memecah teks menjadi token, yang bisa berupa kata-kata atau bahkan karakter tergantung pada kebutuhan. Selama proses tokenisasi, karakter pembatas seperti spasi dapat dihilangkan, angka dapat dihapus, dan tanda baca atau karakter non-huruf lainnya juga dapat dihapus, sesuai dengan kebutuhan pemrosesan teks yang diinginkan.
- d) *Stopwords* adalah proses menghapus kata-kata yang tidak memiliki makna atau relevansi penting.
- e) *Steaming* adalah Proses untuk menyaring kata-kata seperti kata sambung, kata ganti, dan kata depan, serta mengubahnya menjadi bentuk kata dasar dengan menghilangkan awalan atau akhiran.

2.3.4. **Pembobotan Kata (TF-IDF)**

Pada tahap ini, dilakukan pembobotan kata dalam teks menggunakan metode *Term Frequency-Inverse Document Frequency* (TF-IDF). Langkah ini bertujuan untuk menentukan bobot kata dalam sebuah dokumen atau seberapa sering kata tersebut

muncul dalam dokumen tersebut, TF-IDF adalah metode yang umum digunakan dalam analisis teks untuk mengevaluasi pentingnya sebuah kata dalam sebuah koleksi dokumen berdasarkan frekuensinya dalam dokumen dan dalam seluruh koleksi (Fadiyah Basar et al., 2022).

Penggunaan TF dapat menggunakan rumus pada Persamaan (2.1).

$$tf_{ij} = \frac{f_d(i)}{\max f_d(j)} \quad (2.1)$$

TF menunjukkan dokumen (d) seberapa banyak kata (t) yang muncul. Dan terkait untuk rumus IDF dapat dilihat pada Persamaan (2.2)

$$idf_t = \log \frac{N}{df_t} \quad (2.2)$$

N melambangkan jumlah kata dalam teks, df adalah jumlah teks yang memiliki kata t. Dengan menggabungkan TF dan IDF dalam pengerjaan dapat membantu meningkatkan performa. Terkait rumus pembobotan TF-IDF pada Persamaan (2.3).

$$W_{t,d} = tf_{d,t} \times idf_t \quad (2.3)$$

Keterangan :

t = Kata kunci, term

d = Dokumen

W_{d,t} = Bobot d terhadap t

Tf = Banyaknya t (kata) yang dicari dalam dokumen

Idf = Banyak t kebalikan dari kata yang dicari

2.3.5. Split Data

Pada tahap ini, split data adalah proses membagi dataset yang digunakan dalam penelitian menjadi dua bagian, dataset tersebut umumnya dibagi menjadi data latih

(*training*) dan data uji (*testing*). Data latih digunakan untuk melatih algoritma, sedangkan data uji digunakan untuk mengevaluasi kinerja algoritma tersebut (Putri et al., 2023). Dalam penelitian ini, dataset dibagi menjadi dua tahap dengan menggunakan rasio 90:10, 80:20, dan 70:30 untuk mengevaluasi pengaruh rasio tersebut terhadap kinerja model yang digunakan, yaitu Metode KNN dan Naïve Bayes dengan pembobotan fitur menggunakan TF-IDF. Hasil dari pembagian data ini menunjukkan bagaimana perbedaan rasio dapat mempengaruhi tingkat akurasi model dalam mengklasifikasikan sentimen.

2.3.6. Klasifikasi

Klasifikasi merupakan salah satu tahap penting dalam text mining yang bertujuan untuk mengelompokkan data atau objek baru ke dalam kelas atau label berdasarkan atribut-atribut tertentu. Proses ini melibatkan penggunaan teknik yang melihat variabel dari kelompok data yang sudah ada untuk menentukan aturan pengelompokan. Dengan mempelajari pola dari data yang sudah diberi label, klasifikasi memungkinkan kita untuk memprediksi kelas dari suatu objek yang belum diketahui sebelumnya (Azzahra Nasution et al., 2019)

2.3.7. K-Nearest Neighbor (KNN)

Dalam proses klasifikasi peneliti menggunakan dua metode yaitu KNN (K-Nearest Neighbors) dan Naive Bayes, KNN merupakan metode klasifikasi yang sederhana dan non-parametrik yang digunakan untuk mengklasifikasikan data. Meskipun algoritma ini sederhana, kinerjanya sangat baik dan menjadi parameter yang penting klasifikasi KNN membutuhkan metrik dan integer positif (K). Aturan KNN memegang posisi sampel pelatihan beserta kelas mereka ketika menghadapi data masuk baru, tujuan dari algoritma ini adalah untuk mengklasifikasikan objek baru berdasarkan nilai atribut dan data latih yang ada (Putra et al., 2022).

Berikut adalah langkah-langkah klasifikasi algoritma KNN:

- a) Tentukan parameter nilai k = banyaknya jumlah tetangga terdekat.
- b) Hitung jarak antara data *training* dan data *testing*, rumus dapat dilihat pada Persamaan (2.4).

$$euc = \sqrt{(\sum_{i=1}^n (p_i - q_i)^2)} \quad (2.4)$$

Keterangan :

p_i = sample data / data training

q_i = data uji / data testing

I = variabel data

n = dimensi data

- a) Urutkan jarak-jarak tersebut dan tetapkan tetangga terdekat berdasarkan berdasarkan jarak minimum hingga ke- k .
- b) Periksa kelas dari tetangga terdekat.
- c) Gunakan mayoritas sederhana dari kelas tetangga sebagai nilai prediksi untuk data baru.

2.3.8. Naive Bayes

Dan algoritma kedua yang digunakan untuk klasifikasi adalah Naive Bayes. Algoritma ini menggunakan perhitungan probabilitas dan statistik dalam pengoperasiannya. Metode klasifikasi Naive Bayes digunakan untuk mengambil keputusan dengan melakukan prediksi suatu kasus berdasarkan hasil dari klasifikasi yang telah diperoleh. Pada penelitian ini, metode Naive Bayes digunakan untuk menentukan sentimen dari data *Google Playstore* (Astari et al., 2020).

Berikut ini adalah langkah – langkah rumus Naive Bayes pada Persamaan (2.5).

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (2.5)$$

Keterangan :

A = Hipotesis data merupakan suatu *class* spesifik.

B = Data dengan kelas yang masih belum diketahui.

P(A|B) = Probabilitas hipotesis berdasar kondisi.

P(A) = Probabilitas hipotesis berdasar kondisi.

P(B|A) = Probabilitas berdasarkan kondisi pada hipotesis.

P(B) = Probabilitas B.

2.3.9. Evaluasi

Evaluasi digunakan untuk melihat kinerja dari model klasifikasi yang digunakan. Metode evaluasi yang digunakan adalah confusion matrix. Berdasarkan confusion matrix pengujian model dilakukan dengan menghitung nilai accuracy

Perhitungan untuk metode confusion matrix yang digunakan yaitu nilai skor Accuracy, Untuk mengukur seberapa sering model klasifikasi membuat prediksi yang benar (Khatib Sulaiman et al., 2023).

1. Akurasi

Akurasi dapat dijelaskan sebagai seberapa dekat nilai prediksi dengan nilai sebenarnya. Semakin tinggi *akurasi*, semakin baik proses klasifikasi tersebut.

Rumus untuk menghitung akurasi dalam Persamaan (2.6).

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Total\ Prediksi} \quad (2.6)$$

Pada rumus di atas, cara menentukan akurasi dari sebuah data dapat dilihat dengan menghitung jumlah prediksi yang benar dan membaginya dengan jumlah total prediksi yang dilakukan.

2.3.10. Jadwal Penelitian

Penelitian ini akan diawali dengan penentuan judul, identifikasi masalah, studi literatur, rancangan metode, pemilihan studi kasus, dan penyusunan proposal. Setelah tahap pra-penelitian selesai, penelitian akan melanjutkan ke pengumpulan data menggunakan metode *crawling* di *Google Play Store Aplikasi DANA*. Data yang terkumpul akan dilabeli oleh ahli Bahasa (*expert*) dan menggunakan model *lexicon* diproses sebelum pembobotan kata dengan metode TF-IDF. Selanjutnya data akan dibagi menjadi set pelatihan dan pengujian sebelum dilakukan klasifikasi menggunakan algoritma K – Nearest Neighbor (KNN) dan Naive Bayes. Tahap akhir adalah evaluasi dan analisis hasil untuk mengevaluasi efektivitas metode yang digunakan. Setelah selesai, akan dilakukan penyusunan laporan dan presentasi seminar untuk memperkenalkan hasil penelitian. Penelitian akan dilaksanakan sesuai jadwal terperinci pada Tabel 2.1 Jadwal Penelitian.

Tabel 2.1 Jadwal Penelitian

No	Kegiatan	Bulan/2024					
		Feb	Mar	Apr	Mei	Juni	Juli
Tahap Pra Penelitian							
1	Menentukan Judul						
2	Identifikasi Masalah						
3	Studi Literatur						
4	Rancangan Metode						
5	Pemilihan Studi Kasus						
6	Menyusun Proposal						
7	Review Desk Simpel						
Tahap Penelitian							
1	Pengumpulan Data (Crawling)						
2	Labelling Data						
3	Pre-Processing Data						
4	Pembobotan Kata (TF-IDF)						
5	Split Data						
6	Klasifikasi KNN dan Naive Bayes						
7	Evaluasi dan Analisis Hasil						
Tahap Akhir Penelitian							
1	Penyusunan Laporan						

No	Kegiatan	Bulan/2024					
		Feb	Mar	Apr	Mei	Juni	Juli
2	Seminar Hasil						