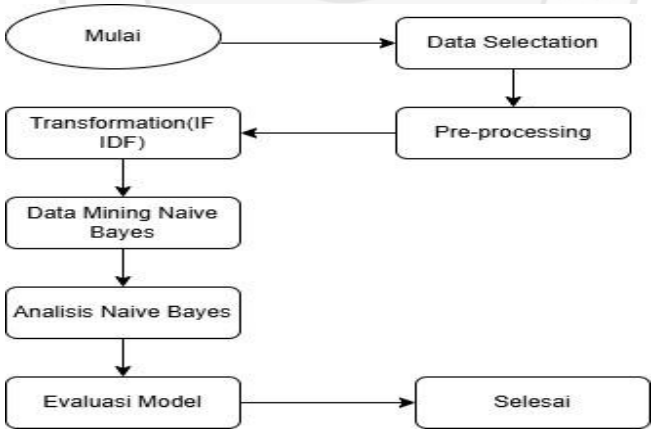
BAB III

METODE PENELITIAN

Pada bab ini membahas tentang metode yang digunakan dalalm penelitian, kerangka penelitian dan hal yang menyangkut dengan atau tata cara dalam melakukan penelitian. Naïve bayes adalah Metode Algoritma Naïve Bayes menganalisis setiap ulasan berdasarkan kata-kata yang ada di dalamnya untuk menentukan apakah sentimen ulasan tersebut tergolong positif, negatif, atau netral. Algoritma ini bekerja dengan menghitung probabilitas kemunculan kata-kata tertentu dalam masing-masing kategori sentimen, sehingga dapat memprediksi apakah ulasan cenderung positif atau negatif.

3.1 Kerangka Penelitian

Setiap langkah-langkah apa yang dilakukan oleh peneliti di jelaskan metode apa yang diguanakan pada peneliti ini. Kerangka penelitian yang digunakan yaitu. Dapat dilihat kerangka di bawah ini.



Gambar 3.1 Kerangka Penelitian

1. Data Selactation

Data Selection atau yang disebut sebagai seleksi data, adalah langkah untuk memilih data yang penting dari kumpulan data yang lebih besar agar analisis atau pengolahan data menjadi lebih baik. Dalam langkah ini, kita mengambil subset data yang sangat terkait dengan tujuan analisis, sehingga hanya data yang relevan yang digunakan. Hal ini mengurangi beban komputasi, membuat analisis menjadi lebih efisien, serta memastikan hasil yang lebih tepat.

2. Pre-Processing

Pre-processing atau prapemrosesan adalah langkah awal dalam pengolahan data yang bertujuan untuk membersihkan dan menyederhanakan data, sehingga lebih siap untuk dianalisis atau digunakan dalam model pembelajaran mesin. Dalam konteks data teks, prapemrosesan biasanya melibatkan mengubah semua huruf menjadi huruf kecil untuk menyamakan bentuk kata, memecah teks menjadi kata-kata (tokenisasi), menghapus kata-kata umum yang tidak relevan (stopword), serta mengubah kata-kata menjadi bentuk dasarnya melalui proses stemming atau lemmatization. Proses ini juga biasanya menghilangkan tanda baca atau simbol lain yang tidak diperlukan.

3. Tranformation (TF-IDF)

TF-IDF, atau *Term Frequency-Inverse Document Frequency*, TF-IDF, adalah suatu cara untuk mengevaluasi signifikansi suatu kata dalam satu dokumen jika dibandingkan dengan dokumen lainnya dalam koleksi teks. Metode ini menghitung dua aspek: seberapa sering kata itu muncul dalam dokumen tertentu (TF) dan seberapa jarang kata itu ditemukan dalam

koleksi dokumen secara keseluruhan (IDF). Ketika sebuah kata muncul sering dalam satu dokumen namun jarang di dokumen lain, nilai TF-IDF untuk kata tersebut menjadi tinggi. Pendekatan ini sangat bermanfaat untuk menunjukkan kata-kata yang penting, karena kata-kata umum seperti "dan" atau "di" memiliki nilai rendah, sedangkan kata-kata yang lebih unik dan spesifik akan mendapati nilai yang lebih tinggi. TF-IDF banyak diterapkan dalam pemrosesan bahasa alami dan pencarian teks dengan tujuan menemukan informasi yang signifikan di antara sejumlah besar data teks.

4. Data Mining Navie Bayes

Naive Bayes merupakan sebuah pendekatan klasifikasi dalam pengolahan data yang mengandalkan Teorema Bayes dengan keyakinan bahwa setiap fitur dalam dataset bersifat mandiri. Walaupun keyakinan ini jarang terjadi, metode Naive Bayes tetap berhasil dan mudah, sering kali digunakan dalam tugas seperti pengkategorian teks dan penilaian sentimen. Proses ini menghitung kemungkinan kelas berdasarkan fitur yang ada, kemudian memilih kelas dengan kemungkinan tertinggi. Naive Bayes banyak diminati karena cepat, simpel untuk diterapkan, dan berkinerja baik meski ada data yang hilang, namun memiliki keterbatasan jika fitur-fitur saling berkorelasi dengan kuat. Naive Bayes adalah teknik klasifikasi yang menggunakan prinsip probabilitas yang berasal dari Teorema Bayes. Persamaan dasar dari algoritma ini adalah: Rumus dasar algoritma ini adalah:

$$P(C|X)= \frac{P(X|C)\cdot P(C)}{P(X)} \qquad (1)$$

Dengan asumsi independensi antar fitur, rumus tersebut dapat disederhanakan menjadi:

$$P(C|X)xP(C)\cdot\prod_{i=1}^{n} P(xi|C) \quad (2)$$

Keterangan:

- *P(C|X)*: Probabilitas ulasan termasuk dalam kategori sentimen *C* berdasarkan data *X*.
- *P(X|C)*: Probabilitas munculnya kata-kata dalam ulasan (*X*) pada kategori sentimen *C*.
- *P(C)*: Probabilitas prior dari kategori sentimen *C*.
- *P(X):* Probabilitas data ulasan (*X*).

5. Analisis Naïve Bayes

Analisis Naïve Bayes merupakan teknik yang memanfaatkan prinsip probabilitas untuk mengklasifikasikan atau memprediksi informasi yang ada. Teknik ini berlandaskan pada Teorema Bayes dengan anggapan bahwa setiap fitur berdiri sendiri. Karena kesederhanaan dan efisiensinya, Naïve Bayes banyak digunakan di berbagai area, termasuk pengenalan teks dan analisis sentimen.

6. Evaluasi Model

Evaluasi model adalah proses langkah untuk mengukur seberapa baik sebuah model dalam memprediksi atau mengklasifikasikan saat menjalankan tugas tertentu. Tujuan dari evaluasi ini adalah untuk memastikan bahwa model dapat berfungsi dengan baik, tidak hanya pada data yang digunakan untuk melatih tetapi juga pada data baru yang belum pernah dihadapi sebelumnya. Dalam evaluasi model Naïve Bayes, fokusnya adalah untuk menilai kemampuan model dalam memprediksi data baru berdasarkan hasil pelatihan yang sudah dilakukan. Proses ini mencakup pengukuran kinerja

Model dengan berbagai metrik yang sesuai.

3.2 Metode Pengumpulan Data

Data dikumpulkan dengan menggunakan teknik web scraping dari Google Play Store dengan menggunakan gogglecollab. Data yang diambil meliputi username, rating bintang, waktu ulasan, dan konten ulasan. Data ini kemudian disimpan dalam format file di drive untuk memudahkan proses analisis lebih lanjut. Data yang digunakan dalam penelitian ini adalah data ulasan pengguna aplikasi Classroom yang diambil dari Google Play Store. Data tersebut terdiri dari berbagai atribut seperti nama pengguna, rating (bintang), tanggal ulasan, dan isi ulasan yang diberikan. Teknik pengumpulan data dilakukan menggunakan web scraping dengan bantuan pustaka goolecollab.

Proses pengumpulan data dilakukan melalui beberapa tahap sebagai berikut:

a. Menentukan Aplikasi dan Atribut Data: Peneliti memilih aplikasi Classroom di GooglePlay Store sebagai objek penelitian. Atribut yang diambil meliputi nama pengguna, rating, tanggal, dan isi ulasan, yang diperlukan untuk proses analisis sentimen.

b. Pengaturan Bahasa dan Negara: Data ulasan yang dikumpulkan difokuskan pada pengguna di Indonesia. Pengaturan bahasa ditentukan dalam bahasa Indonesia (lang=id) dan negara Indonesia (country=id), untuk memastikan relevansi data dengankonteks penelitian.

c. Mengambil Data Berdasarkan Rating: Ulasan dikumpulkan berdasarkan skor rating (1hingga 5 bintang) untuk memastikan variasi sentimen. Data diambil secaraproporsional agar mencakup ulasan dengan berbagai tingkat kepuasan.

d. Penyimpanan Data drive: Data yang telah terkumpul kemudian disimpan dalam drive untuk memudahkan proses analisis. Format ini memungkinkan data diakses dandianalisis lebih lanjut menggunakan alat analisis data seperti googlecollab atau perangkat lunak statistik lainnya. Data, lalu memilih kelas dengan probabilitas tertinggi. Naive Bayes populer karena cepat, mudah diterapkan, dan tetap bekerja baik meskipun ada data yang hilang,namun kurang optimal jika antar-fitur memiliki korelasi kuat.