BAB III
METODE PENELITIAN

3.1 Prosedur Pengumpulan Data

Pengumpulan data dalam penelitan ini dilakukan melalui dua jenis sumber data, yaitu data primer dan data sekunder,

1. Data Primer

   Data primer dalam penelitian ini diperoleh secara langsung dari sumber utama, yaitu melalui *data crawling* dari media sosial twitter. Proses pengumpulan data dilakukan dengan menggunakn Twitter *API* melalui *tools* Twitter Harvest untuk mengunduh *tweet* yang mengandung opini masyarakat mengenai layanan Indihome. Pengambilan data dilakukan dalam periode tertentu untuk mendapatkan sampel yang representatif yaitu sebanyak 2000 data yang diambil pada rentang waktu 2021 hingga 2024 dimana setiap tahun berjumlah 500 data.
2. Data Sekunder

   Data sekunder diperoleh dari berbagai sumber yang mendukung penelitian ini, seperti jurnal ilmiah, buku, artikel, laporan penelitian terkait, dan studi sebelumnya yang membahas sentimen pengguna terhadap suatu produk atau layanan.

3.2 Analisa Kebutuhan

Dalam penelitian ini, analisa kebutuhan dilakukan untuk menentukan sumber daya yang diperlukan dalam proses pengolahan data dan penerapan model. Kebutuhan penelitan ini dibagi menjadi dua aspek utama diantaranya sebagai berikut :

1. Kebutuhan perangkat keras, yaitu Laptop ASUS X541L dengan spesifikasi :

- Prosesor : Intel Core i3-4005U
- RAM : 8 GB
- Penyimpanan : SSD 256 GB

2. Kebutuhan perangkat lunak :
    - Microsoft Windows 11
    - Google Chrome
    - Google Collab
    - Rapid Miner Studio

3.3 Teknik Analisis Data

Dalam penelitian ini, analisis data dilakukan dengan menggunakan metode *SEMMA* (*Sample, Explore, Modify, Model, Asses*) yang terdiri dari lima tahapan utama untuk memastikan pengolahan data berjalan secara sistematis dan optimal. Berikut merupakan tahapan-tahapan dalam metode *SEMMA* yang diterapkan dalam penelitian ini :

1. *Sample*

   Data yang digunakan dalam penelitian ini diperoleh melalui dua sumber utama yaitu penelitian terdahulu yang relevan dan data yang dikumpulkan dari Twitter menggunakan teknik *crawling*. Data yang diambil mencakup berbagai opini pengguna terkait layanan Indihome melalui *tweet* pada Twitter. Data yang dikumpulkan kemudian disimpan dalam format csv untuk diproses lebih lanjut.

2. *Explore*

   Tahap ini bertujuan untuk memahami karakteristik data dengan melakukan eksplorasi terhadap pola, distribusi, serta korelasi antar atribut dalam dataset. Atribut atau fitur yang tidak relevan atau tidak memiliki nilai yang signifikan akan diseleksi dah dihapus agar tidak mengganggu proses analisis. Lalu dilakukan pengecekan terhadap data yang hilang (*missing values*) atau data yang tidak konsisten.

3. *Modify*

   Pada tahap ini, data yang masih dalam bentuk tidak terstruktur diproses agar menjadi terstruktur sehingga dapat digunakan dan dapat dimplementasikan dalam pemodelan. Proses *preproessing* meliputi beberapa langkah, yaitu : *cleaning*, *tokenization*, *transform cases*, *stopword removal*, dan *filtering*. Selanjutnya melakukan visualisasi kata yang sering muncul dalam bentuk *wordcloud*.

4. *Model*

   Pada tahap ini data diberi label sesuai dengan kategorinya untuk mengidentifikasi opini sebagai positif atau negatif. Pelabelan dilakukan secara manual untuk sebagian data saja sebagai data latih. Data yang telah diproses, digunakan sebagai dataset unutk membangun model klasifikasi sentimen untuk melabeli kategori pada data uji secara otomatis.

5. *Asses*

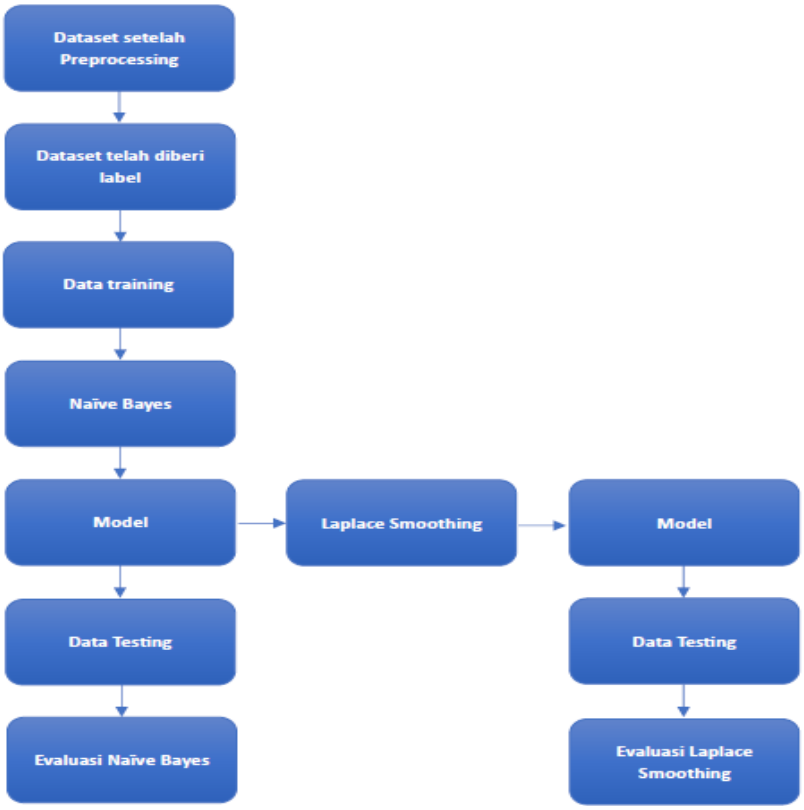
   Setelah model terbentuk, tahap ini bertujuan untuk mengukur kinerja model yang telah dibangun. Evaluasi dilakukan dengan berdasarkan nilai pada *Confussuion Matrix* yang memberikan informasi perbandingan hasil klasifikasi yang dilakukan model dengan hasil klasifikasi sebenranya yang terdiri dari akurasi, presisi, dan recall

3.4 Model yang diusulkan

Dalam penelitian ini, proses klasifikasi dilakukan menggunakan algoritma *Naïve Bayes* untuk mengidentifikasi sentimen positif dan negatif dari *tweet* dan penerapan metode *Laplace Smoothing* yang telah melewati tahap pengolahan data. Proses klasifikasi ini mencakup beberapa langkah, yaitu *Training*, *Learning*, dan *Testing*. Klasifikasi *Naïve Bayes* dibangun oleh data pelatihan untuk memperkirakan probabilitas dari setiap kategori yang terdapat pada ciri dokumen yang diuji. Sistem akan dilatih dengan menggunakan data baru (data

latih dan data uji) dan selanjutnya diberi tugas untuk menebak nilai fungsi target dari data tersebut.

Dalam proses klasifikasi sentimen, data yang telah melalui tahap preprocessing akan diberikan bobot menggunakan metode *TF-IDF*. Saat data latih diproses, sistem akan melakukan proses ekstraksi fitur kata dan selanjutnya model data akan disimpan. Model yang telah dilatih kemudian diuji menggunakan data uji guna mengukur tingkat akurasi klasifikasi yang dilakukan. Selanjutnya, model tersebut diproses kembali dengan menerapkan metode *Laplace Smoothing* untuk megatasi probabilitas nol pada kata yang tida muncul di suatu kelas. Model hasil *Laplace Smoothing* juga diuji menggunakn data uji yang sam untuk memperoleh nilai akurasi kedua. Hasil akhir dari proses ini adalah perbandingan hasil prediksi baik pada model *Nzaïve Bayes* biasa maupun *model Naïve Bayes* dengan *Laplace Smoothing.*



Gambar 3. 1 Flowchart Naïve Bayes dan Laplace Smoothing

3.5 Uji Model

Untuk mengetahui performa dari model, maka harus dilakukan proses pengujian terhadap model yang dibuat. Hasik klasifikasi akan divisualisasikan dalam bentuk *Confussion Matrix*, yaitu salah satu metode yang dapat digunakan untuk mengukur kinerja suatu model klasifikasi. Pada dasarnya *confussion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi seharusnya.

Pengujian model dilakukan terlebih dahulu dengan membagi data menjadi dua bagian, yaitu data latih dan data uji dengan cara *split validation*. Cara *split validation* dilakukan dengan jumlah data yang dijadikan sebagai pengujian diambil dari data training sebesar 30%. Setelah pengujian model dilakukan maka akan tampil seberapa besar performa metode yang dilakukan.

Alasan penggunaan rasio 30% data latih dan 70% data uji pada percobaan kedua didasarkan pada hasil evaluasi yang menunjukkan akurasi lebih tinggi dibanding percobaan pertama. Pada percobaan pertama dengan rasio 50%-50%, jumlah data latih yang lebih besar berpotensi menyebabkan overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data latih sehingga kinerja menurun saat diuji dengan data baru, yang terlihat dari akurasi yang hanya mencapai 62.50%. Selain itu, porsi data uji yang lebih besar pada percobaan kedua juga meningkatkan peluang distribusi kelas positif dan negatif yang lebih seimbang. Ukuran data uji yang besar ini turut mengurangi penyimpangan acak dan membuat evaluasi lebih handal, sehingga akurasi 70,12% yang diperoleh peada percobaan kedua dianggap lebih reliabel disbanding percobaan pertama.