BAB 3
METODOLOGI PENELITIAN

3.1 Metode Penilitian

Dalam melakukan penelitian implementasi Naive Bayes berikut ini adalah metodolgi penilitian yang diterapkan:

3.1.1 Studi Literatur

Penyusunan literatur adalah langkah pertama yang melibatkan pembelajaran teori-teori terkait, termasuk *text preprocessing*, *Term Frequency-Inverse Document Frequency* (TF-IDF), *Confusion Matrix*, dan Naive Bayes.

3.1.2 Pengumpulan Data

Pada tahap ini dilakukan pengumpulan data dari tweet dari pengguna aplikasi X menggunakan kata kunci 'bensin oplosan', 'pertamax oplosan', dan 'pertamax dicampur pertalite'. Proses pengumpulan dilakukan menggunakan *tweet harvest*. *tweet harvest* dapat mengambil data *tweets* berdasarkan kata kunci yang diberikan. Setelah selesai melakukan pengambilan data *tweets*, data tersebut disimpan kedalam bentuk CSV.

3.1.3 Pengolahan Data

Pada tahap ini merupakan tahap dimana data *tweets* yang telah dikumpulkan diolah agar laya untuk di uji. Terdapat berbagai proses yang diimplementasikan dalam pengolahan data, yang pertama merupakan menerjemahkan atau *translate* karena terdapat *tweets* yang menggunakan bahasa Inggris dan harus diubah ke dalam bahasa Indonesia. Tahap yang kedua merupakan *case folding*, yang dimana pada tahap ini dilakukan untuk mengubah semua huruf kapital menjadi huruf kecil, lalu pada tahap ketiga adalah membuang *tweets* yang duplikat, tahap ini bertujuan untuk menghindari potensial spam yang dapat memengaruhi akurasi. Tahap keempat merupakan *Tokenizing*, tahap ini melakukan pemisahan kalimat sesuai dengan kata pembentuknya. Tahap kelima adalah *Remove Stop Words*, tahap ini melakukan penghapusan kata yang tidak relevan dalam sebuah kalimat. Tahap

keenam merupakan Normalisasi, yang dimana digunakan untuk merubah kata-kata yang tidak baku menjadi baku dalam sebuah kalimat. Tahap ketujuh merupakan *Stemming*, yang dimana kata yang mempunyai kata imbuhan kemuadian diubah menjadi kata dasarnya.

Setelah pengolahan data selesai, berikutnya adalah proses labeling. Proses ini memberikan label sentimen (positif atau negatif) untuk setiap tweet. Pada penelitian ini menggunakan Lexicon untuk proses labeling. Bobot yang diberikan pada setiap kata dalam *lexicon* adalah antara negatif 5 sampai dengan positif 5. Untuk perhitungan setiap tweet, jika nilai total bobot di bawah nol adalah sentimen negatif, sebaliknya nilai total bobot di atas nol adalah sentimen positif. Tweet dengan total bobot nol akan dikeluarkan dari dataset.

3.1.4 Pengujian dan Evaluasi

Setelah model menyelesaikan tahap labeling dan data dipisahkan menjadi data *test* dan data *train*, langkah berikutnya adalah pengujian model. Data uji digunakan untuk mengevaluasi kinerja model yang telah dilatih. Pengujian dilakukan dengan mengukur metrik evaluasi standar seperti akurasi, presisi, *recall*, dan *F1-score* berdasarkan hasil klasifikasi dari model Naive Bayes. Hasil pengujian ini dievaluasi secara komprehensif guna menetapkan kinerja dan tingkat akurasi model menggunakan metode *Confusion Matrix*. Evaluasi ini juga mencakup analisis perbandingan performa antara varian Multinomial Naive Bayes dan Complement Naive Bayes.

3.1.5 Penulisan Laporan

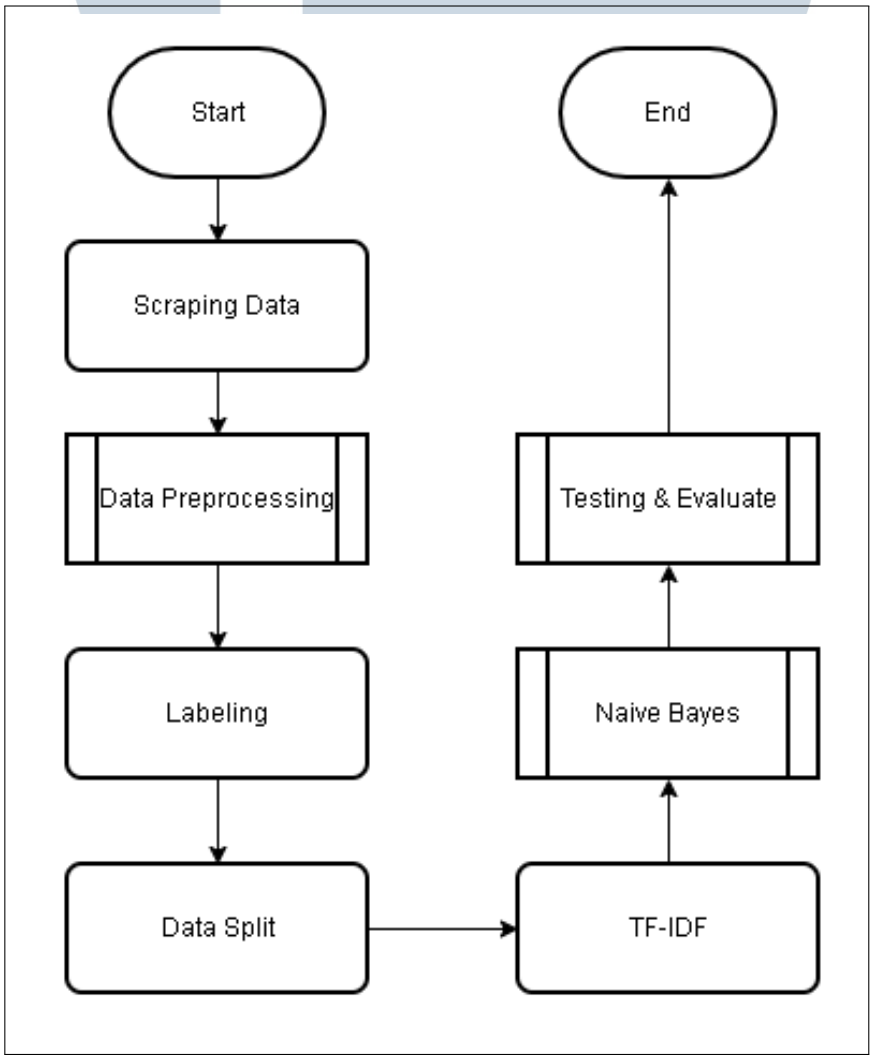
Tahap terakhir adalah penulisan laporan. Tahap ini berisi pendahuluan sampai dengan kesimpulan dan saran. Dengan adanya penulisan laporan ini, membuktikan bahwa penelitian telah dilakukan dan diselesaikan.

3.2 Perancangan Sistem

Pada bagian ini membahas alur sistem yang dibuat. Penggambaran alur sistem dalam penelitian ini dilakukan melalui sebuah *flowchart*.

3.2.1 Gambaran Umum Sistem

Gamabr 3.1 Ini adalah uraian tentang langkah-langkah sistem yang digunakan dalam penelitian. Penelitian dimulai dengan melakukan scraping tweet data menggunakan *Tweet Harvest*. Setelah pengambilan data, langkah berikutnya adalah data preprocessing untuk menyiapkan data agar dapat diberi label. Labeling dilakukan dengan memanfaatkan Lexicon, yang berfungsi untuk memberi bobot atau nilai pada setiap kata dalam kalimat tertentu. Kemudian, data tersebut dibagi menjadi data pelatihan dan pengujian, di mana TF-IDF diterapkan dan data disimpan sebagai model. Model ini kemudian diklasifikasi menggunakan naive bayes untuk memperoleh hasil akhir berupa nilai akurasi, presisi, recall, dan f1-score.



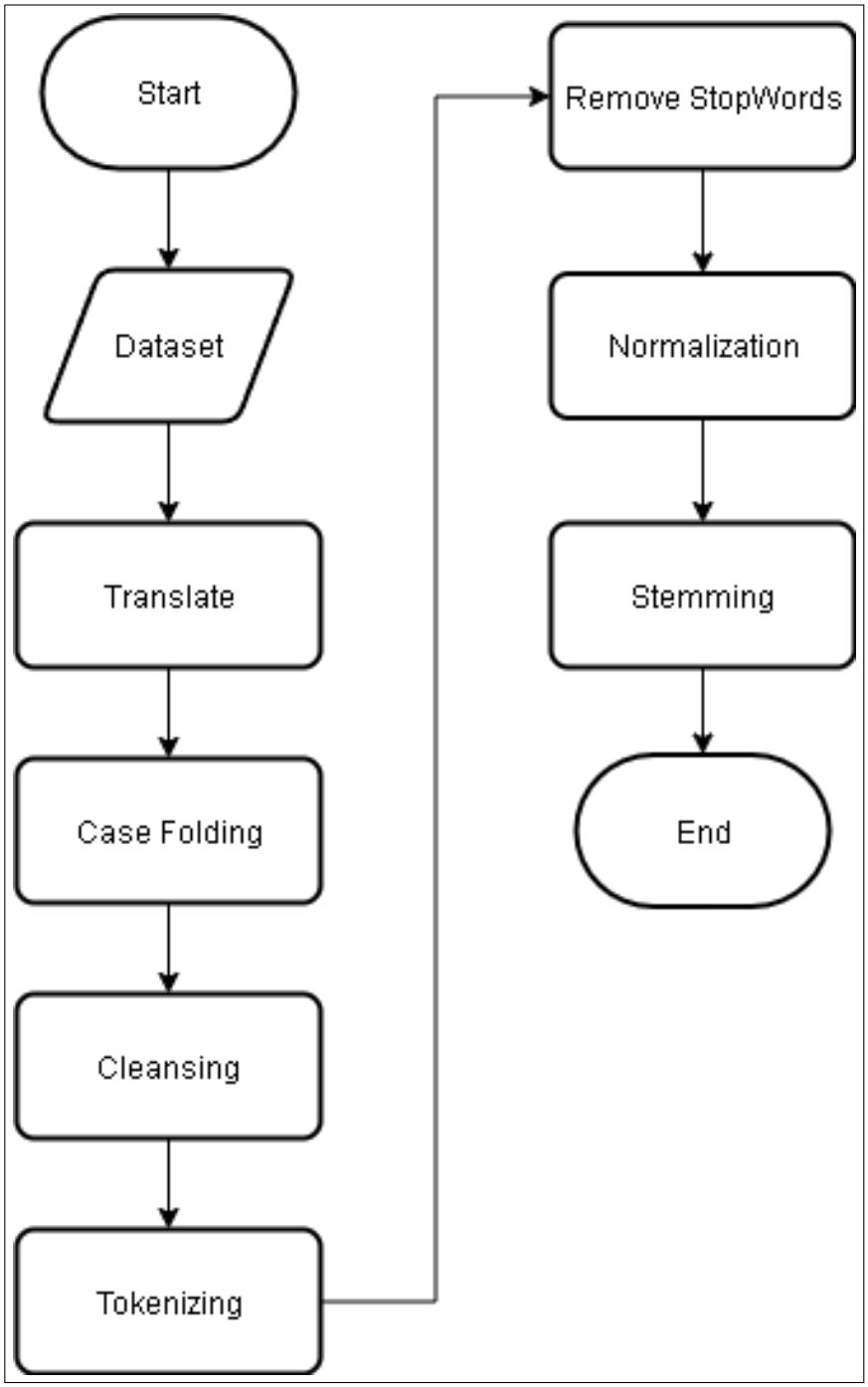
Gambar 3.1. Gamabran Umum

3.2.2 Data Mining

Penelitian analisis sentimen ini mengumpulkan data utama dengan memanfaatkan sebuah *tool* bernama *Tweet Harvest*. Alat ini memfasilitasi pengambilan tweet dari Twitter berdasarkan kata kunci spesifik. Fokus penelitian diarahkan pada topik pencampuran pertamax dan pertalite di Indonesia. Oleh karena itu, kata kunci yang diterapkan dalam proses text mining adalah "Pertamax dicampur Pertalite". Penggunaan kata kunci dalam bahasa Indonesia dipilih untuk memastikan keterkaitan data dengan konteks bahasa dan lokasi di Indonesia.

3.2.3 Data Pre-Processing

Proses data preprocessing terdiri dari beberapa tahapan penting yang dilakukan sebelum analisis data dimulai. Tahap pertama adalah penerjemahan data ke dalam bahasa Indonesia. Langkah ini diperlukan karena sejumlah tweet yang dikumpulkan menggunakan bahasa asing, terutama bahasa Inggris. Selanjutnya, dilakukan case folding, yaitu proses konversi seluruh huruf dalam data menjadi huruf kecil untuk menyamakan format penulisan. Tahap ketiga adalah cleansing, yakni pembersihan data dari unsur-unsur non-teks seperti hashtag, mention, simbol, dan konten spam. Tujuan dari tahap ini adalah untuk menyisakan kata-kata yang relevan dan diperlukan dalam proses pemberian bobot. Tahap keempat adalah tokenizing, di mana setiap kalimat diuraikan menjadi kata-kata terpisah menggunakan pustaka Natural Language Toolkit (NLTK). Setelah itu, dilakukan proses penghapusan stop words dengan bantuan pustaka Sastrawi. Tahap ini bertujuan untuk mengeliminasi kata-kata yang dianggap tidak memberikan kontribusi makna yang signifikan. Tahap selanjutnya adalah normalization, yaitu proses mengubah kata-kata tidak baku menjadi bentuk baku sesuai kaidah bahasa Indonesia. Tahap terakhir adalah stemming, juga menggunakan pustaka Sastrawi, yang berfungsi untuk mengubah kata berimbuhan menjadi bentuk dasarnya.



Gambar 3.2. Flowchart Data Preprocessing

3.2.4 Labeling

Proses labeling merupakan tahapan penting yang dilakukan sebelum pengujian model dilakukan. Terdapat dua pendekatan utama dalam proses labeling, yaitu melalui pemanfaatan pustaka (library) khusus atau secara manual dengan

melibatkan bantuan pihak lain untuk memberikan label pada data. Dalam penelitian ini, metode yang digunakan adalah dengan menerapkan pustaka bernama Lexicon. Setiap kata yang terdapat dalam lexicon diberi nilai atau bobot tertentu yang berada dalam rentang minus 5 hingga positif 5.

Namun, karena penelitian ini hanya memfokuskan pada dua kategori sentimen, yaitu positif dan negatif, maka dilakukan proses agregasi terhadap bobot kata dalam setiap kalimat. Kalimat yang memiliki jumlah total bobot kurang dari nol dikategorikan sebagai sentimen negatif, sedangkan yang melebihi nol dikategorikan sebagai sentimen positif. Sementara itu, kalimat yang memiliki jumlah bobot sama dengan nol tidak digunakan dalam analisis lebih lanjut dan dikeluarkan dari dataset.

3.2.5 Data Splitting

Data kemudian dibagi menjadi dua bagian, yaitu data latih (*train*) dan data uji (*test*). Model dilatih menggunakan data latih, lalu dievaluasi menggunakan data uji. Sebanyak 20% dari keseluruhan data digunakan sebagai data uji untuk menguji kinerja model Multinomial Naive Bayes dan Complement Naive Bayes yang digunakan dalam penelitian ini.

3.2.6 TF-IDF

Setelah tahapan data preprocessing selesai dilakukan, langkah selanjutnya yang diperlukan sebelum proses klasifikasi dengan algoritma Naive Bayes adalah pemberian bobot pada setiap kata dalam data. Proses pembobotan ini diawali dengan perhitungan nilai Term Frequency (TF) dan Inverse Document Frequency (IDF). Selanjutnya, kedua nilai tersebut dikombinasikan untuk membentuk nilai TF-IDF sebagai representasi numerik dari tingkat kepentingan suatu kata dalam keseluruhan dokumen.

3.2.7 Penerapan Model Naive Bayes

Pada tahap ini, data yang telah dipisahkan dan pembobotan dengan metode TF-IDF digunakan pada tahap Train Data. Selanjutnya, Train Data tersebut digunakan dalam Modeling using Naïve Bayes untuk melatih model klasifikasi. Model yang telah dibentuk akan masuk ke tahap Train Model, di mana dua algoritma Naïve Bayes digunakan, yaitu Complement Naïve Bayes dan Multinomial Naïve Bayes. Setelah proses pelatihan selesai, diperoleh Trained Model yang siap dievaluasi.

Evaluasi model dilakukan menggunakan beberapa metrik performa, yaitu akurasi, presisi, *recall*, dan *F1-score* untuk menilai kualitas klasifikasi dari masing-masing model.

3.2.8 Testing dan Evaluasi

Setelah proses pelatihan model selesai dilakukan, tahap berikutnya adalah melakukan pengujian terhadap model menggunakan data uji yang telah disiapkan sebelumnya. Data uji ini bersifat identik dan representatif terhadap data yang digunakan dalam pelatihan, guna memastikan bahwa model dapat melakukan generalisasi dengan baik. Untuk mengukur performa model dalam mengklasifikasikan sentimen secara akurat, digunakan metode evaluasi *confusion matrix* yang mampu memberikan gambaran mendetail mengenai jumlah prediksi yang benar dan salah untuk setiap kelas sentimen. Berdasarkan *confusion matrix* tersebut, diperoleh laporan evaluasi yang memuat metrik-metrik penting, yaitu nilai akurasi, presisi, *recall*, dan *F1-score* untuk masing-masing kelas sentimen, baik positif maupun negatif. Metrik-metrik ini menjadi acuan utama dalam menilai seberapa baik model dalam mengenali dan membedakan opini pengguna secara objektif serta konsisten.

3.3 Dokumentasi

Tahap terakhir dalam penelitian ini adalah proses dokumentasi. Pada tahap ini, peneliti menyusun laporan penelitian berdasarkan seluruh rangkaian kegiatan yang telah dilaksanakan, dimulai dari perumusan latar belakang hingga penarikan kesimpulan. Penulisan laporan dilakukan menggunakan Overleaf, sesuai dengan ketentuan dan persyaratan penulisan ilmiah yang ditetapkan oleh Universitas Multimedia Nusantara (UMN).