BAB 3
METODE PENELITIAN

Penelitian ini merupakan analisis umpan balik pengguna untuk mengidentifikasi sentimen dan tren topik pada aplikasi Bima *Mobile*. Metode yang digunakan yaitu kombinasi algoritma SVM dan NMF. Berikut ini merupakan bahan, alat, serta jalan penelitian dalam melakukan analisis sentimen dan pemodelan topik dari umpan balik pengguna aplikasi Bima *Mobile*.

3.1 Bahan dan Alat Penelitian

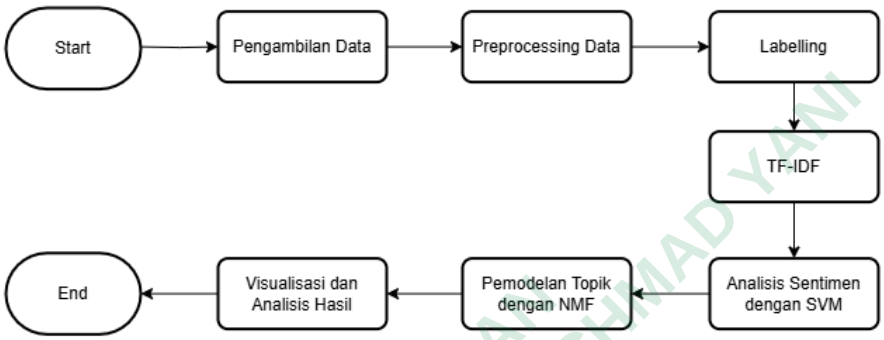
Bahan yang diperlukan untuk melakukan penelitian ini adalah data umpan balik pengguna aplikasi Bima *Mobile* yang didapatkan dari *Google Playstore*. Data yang digunakan mencakup seluruh umpan balik yang tersedia, dengan jumlah data yaitu 4294 data.

Alat yang dibutuhkan dalam penelitian ini adalah laptop dengan spesifikasi yang memadai untuk menjalankan sistem operasi, perangkat lunak pengembangan, serta memiliki konektivitas internet. Adapun sistem operasi dan *software* yang digunakan untuk menunjang penelitian, sebagai berikut:

1. Sistem Operasi *Windows 11 64-bit*
2. Bahasa Pemrograman *Python 3.11.11*
3. *Google Colaboratory*
4. *Google Spreadsheet*
5. *Library Python*:
    a. Google Play Scraper
    b. Pandas
    c. Numpy
    d. Matplotlib
    e. Scikit-learn
    f. NLTK
    g. Sastrawi
    1. Gensim

3.2 JALAN PENELITIAN

Penelitian ini memanfaatkan bahasa pemrograman *Python*, dan *platform Google Colaboratory* untuk melakukan pengambilan dan pengolahan data dengan memanfaatkan beberapa *library* yang dimiliki oleh bahasa pemrograman Python. Jalan penelitian yang dilakukan, ditunjukkan pada Gambar 3.1



Gambar 3.1 *Flowchart* Jalan Penelitian

Berikut adalah penjelasan dari setiap tahapan pada penelitian ini, yaitu:

1. *Scraping* atau Pengambilan Data

Data yang diambil merupakan data seluruh rating dan umpan balik dari pengguna Bima *Mobile* yang ada di *Google Play Store*, dengan ketentuan umpan baliknya tidak kosong (hanya rating). Pengambilan data dilakukan dengan menggunakan bantuan *Google-Play-Scraper*, kemudian disimpan dalam format csv.

2. *Preprocessing* Data

Preprocessing merupakan tahap pemrosesan data yang bertujuan untuk membersihkan, dan mengubah data mentah supaya dapat digunakan dalam analisis atau pemodelan. Berikut merupakan tahapan dalam *preprocessing* data:

a. Data *Cleaning*

Data cleaning merupakan proses membersihkan data dengan menghilangkan kata atau karakter yang tidak sesuai, seperti menghapus angka, tanda baca, spasi berlebih, *emoticon*, dan simbol-simbol lain.

b. *Case Folding*

Case folding yaitu salah satu tahapan *text preprocessing* dengan tujuan untuk mengubah semua huruf ke dalam huruf kecil.

c. *Tokenizing*

Tokenizing merupakan proses membagi teks menjadi unit-unit yang lebih kecil supaya lebih mudah diolah atau diinterpretasi dalam analisis teks.

d. Normalization

Normalisasi teks merupakan langkah untuk mengubah kata-kata tidak baku atau *slang* menjadi kata yang lebih standar. Normalisasi dapat dilakukan dengan menggunakan bantuan daftar *slangword* (kata tidak baku) yang tersedia di GitHub maupun melalui penyusunan daftar secara manual sesuai dengan konteks dan kebutuhan analisis (Dani, 2024).

e. *Stopword Removal*

Stopword removal merupakan proses penghapusan kata-kata umum yang sering muncul dalam teks, tetapi tidak memiliki nilai informasi yang signifikan. Contohnya seperti kata "dan", "di", "ke", dan sebagainya. Salah satu cara untuk melakukan *stopword removal* adalah dengan menggunakan *library* Sastrawi, yang menyediakan daftar kata umum dalam Bahasa Indonesia yang dapat dihapus secara otomatis dari teks (Aini, 2023).

f. *Stemming*

Stemming digunakan untuk menyederhanakan kata-kata ke dalam bentuk dasarnya agar dapat direduksi menjadi bentuk yang lebih seragam.

3. *Labelling*

Labelling adalah proses penandaan data teks dengan label sesuai dengan kategori sentimennya. Untuk mengurangi subjektivitas dalam penilaian, proses pelabelan data dikategorikan menjadi dua label, yaitu label "Positif" dan label "Negatif" (Hidayat & Sugiyono, 2023). Tahapan pelabelan dilakukan secara manual dengan menentukan secara pribadi suatu umpan balik termasuk ke dalam kelas positif atau negatif (Sujjadaa et al., 2023). Tujuan dari *labelling* yaitu untuk menghasilkan data yang dapat digunakan dalam proses *training* model.

4. TF-IDF

Sebelum proses *training*, hal yang perlu dilakukan yaitu ekstraksi fitur pada data teks. Ekstraksi fitur merupakan tahapan untuk mengubah data teks ke dalam vektor numerik yang digunakan untuk melatih model *machine learning*. Proses

ekstraksi fitur pada penelitian ini dilakukan menggunakan TF-IDF dengan mengimpor modul *TfidfVectorizer* dari *library Scikit-learn*. Tahapan ini dilakukan untuk mengubah kata yang telah melalui tahap *preprocessing* menjadi vektor yang merepresentasikan tingkat kepentingan kata dalam dokumen.

5. Analisis Sentimen dengan SVM

Setelah menghitung TF-IDF, langkah berikutnya adalah membangun model klasifikasi dengan data yang telah dibagi menjadi data latih atau *training data*, dan data uji atau *testing data*. Pembuatan model dilakukan dengan menggunakan algoritma SVM dengan kernel linear untuk memprediksi kategori sentimen positif atau negatif berdasarkan data yang telah diproses sebelumnya. Proses selanjutnya, yaitu melakukan evaluasi model SVM yang telah dibuat. Model dievaluasi untuk mengukur kinerja dan akurasinya dalam menganalisis umpan balik pengguna aplikasi Bima *Mobile*. Dalam penelitian ini, evaluasi dilakukan dengan *confusion matrix* untuk menunjukkan jumlah prediksi yang benar dan salah, serta mendukung perhitungan metrik evaluasi lainnya, seperti akurasi, presisi, *recall*, dan *F1-score*.

6. Pemodelan Topik dengan NMF

Setelah diklasifikasikan menjadi kategori sentimen positif dan negatif dengan model SVM, tahap selanjutnya adalah pemodelan topik dengan menggunakan NMF untuk mengidentifikasi topik utama dalam setiap kategori sentimen. Proses pemodelan topik dimulai dengan menentukan jumlah topik yang optimal untuk membangun model. Penentuan jumlah topik menjadi sangat penting, karena jumlah topik yang terlalu sedikit dapat mengurangi kemampuan model dalam menangkap variasi tema dalam dokumen, sementara jika jumlah topik yang dihasilkan cukup banyak dapat menyulitkan interpretasi hasil analisis (Kusumaningtyas et al., 2024).

Metode yang dipakai untuk menentukan jumlah topik optimal pada penelitian ini adalah *coherence score*. Teknik pengukuran *coherence* dilakukan menggunakan uji koherensi c_v. Suatu topik dianggap lebih koheren jika kata-kata yang dihasilkannya sering muncul bersama dalam dokumen yang sama. *Coherence score* yang tinggi menunjukkan bahwa kata-kata dalam suatu topik memiliki keterkaitan yang kuat, sedangkan *coherence score* yang rendah menandakan topik kurang bermakna atau tidak memiliki kesamaan konteks. Setelah diperoleh jumlah topik

optimal, langkah selanjutnya adalah menerapkan NMF pada setiap kategori untuk menghasilkan kata-kata kunci dari setiap kategori topik.

7. Visualisasi dan Analisis Hasil

Setelah melakukan pelatihan model sentimen dan pemodelan topik NMF, visualisasikan hasil distribusi sentimen serta kata-kata kunci dari topik yang dihasilkan. Distribusi sentimen divisualisasikan dengan menggunakan *bar chart*, dan kata-kata kunci hasil pemodelan topik divisualisasikan menggunakan *WordCloud*. Tahap ini dilakukan untuk memberikan gambaran visual terhadap hasil yang telah diperoleh dari pemodelan sentimen dan topik. Visualisasi distribusi sentimen menggunakan *bar chart* bertujuan untuk menunjukkan proporsi masing-masing kategori sentimen dalam data umpan balik. Sementara itu, visualisasi kata-kata kunci dari hasil pemodelan topik NMF menggunakan *WordCloud* bertujuan untuk mempermudah identifikasi kata-kata yang paling menonjol pada setiap topik yang terbentuk. Setelah visualisasi dilakukan, langkah selanjutnya adalah melakukan analisis terhadap hasil distribusi sentimen dan pemodelan topik guna memperoleh pemahaman yang mendalam serta mengidentifikasi tren topik dari umpan balik aplikasi Bima *Mobile*.