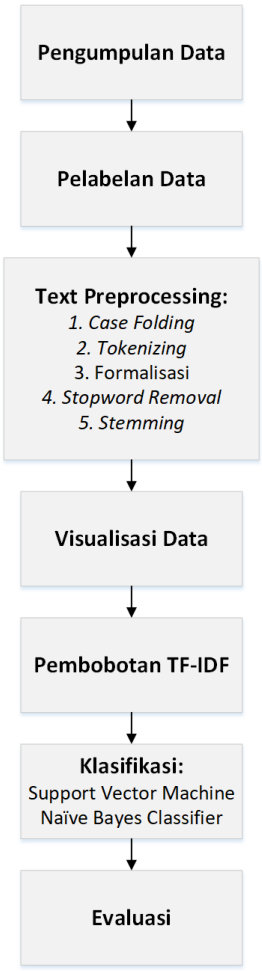
BAB III

ANALISA DAN PERANCANGAN

Penelitian ini menggunakan pendekatan kuantitatif dengan teknik eksperimen, di mana data ulasan pengguna akan dianalisis menggunakan algoritma Support Vector Machine dan Naïve Bayes Classifier. Gambar 3.1 memperlihatkan alur kerja dari penelitian ini.



Gambar 3.1 Alur Penelitian

3.1 Alat Dan Perangkat Lunak Yang Digunakan

A. Bahasa Pemrograman: Python

B. Library:

- Scikit-learn: Untuk implementasi SVM dan Naïve Bayes, serta evaluasi kinerja.
- NLTK/Spacy: Untuk pra-pemrosesan teks.
- Pandas dan NumPy: Untuk manipulasi data.

C. Google Colab/Jupyter Notebook: untuk pengembangan dan eksperimen

3.2 Alur Penelitian

3.2.1 Pengumpulan Data

Tahap pertama yang dilakukan adalah pengumpulan data. Pengumpulan data pada penelitian ini menggunakan teknik *scraping* data pada ulasan Aplikasi Shopee didalam layanan Google Play Store. *Scraping* data dilakukan dengan mengumpukan data ulasan pengguna aplikasi Shopee dari Google Play Store menggunakan bantuan dari library google-play-scraper dan IDE Google Collab. Data yang diperoleh sebanyak 5000 baris dan 11 kolom/variable. Variabel dari data yang didapatkan adalah reviewId, userName, userImage, content, score, thumbsUpCount, reviewCreatedVersion, at, repyContent, repliedAt, dan appVersion.

3.2.2 Pelabelan Data

Untuk melakukan analisis sentimen perlu dilakkukan pembuatan label (*labelling*). Pada tahap ini dataset diklasifikasikan menjadi tiga kategori yaitu,

negatif, netral, atau positif. Proses ini akan dilakukan dengan mengkategorikan jumlah bintang pada ulasan. Ulasan yang memiliki jumlah bintang 1 dan 2 dikategorikan sebagai sentimen “Negatif”. Ulasan yang memiliki jumlah bintang 3 dikategorikan sebagai sentimen “Netral”. Sedangkan ulasan yang memiliki jumlah bintang 4 dan 5 dikategorikan sebagai sentimen “Positif”. Selain melakukan *labelling*, variabel yang relevan untuk proses pemodelan akan dipilih terlebih dahulu

3.2.3 Text Preprocessing

Preprocessing merupakan proses awal dalam pengolahan data masukan sehingga akan menghasilkan data dengan format yang sesuai dan siap diproses pada tahap selanjutnya. Tujuan dilakukannya *preprocessing* agar data yang digunakan lebih presisi dan memudahkan dalam menjalankan analisis sentimen. Dibutuhkan lima tahapan dalam melakukan *text preprocessing* yaitu: *case folding*, *tokenizing*, formalisasi, *stopword removal*, dan *stemming*.

a. Case Folding

Tahap case folding adalah proses pembersihan teks dari kata yang tidak diperlukan. Kata yang dihilangkan ialah karakter, simbol, huruf tunggal, link URL, dan sebagainya. Tahap ini adalah proses mengubah teks menjadi bentuk standar sehingga data masukkan akan diubah menjadi huruf kecil. Tahap ini akan dibantu dengan bantuan library RegEx.

b. Tokenizing

Tokenizing merupakan proses menyortir dan membagi frase menjadi kata-kata individu yang dikenal sebagai tokenisasi. Token bisa berupa kata-kata, karakter, atau bagian-bagian kata.

c. Formalisasi

Pada tahap ini akan dilakukan penyeragaman kata yang memiliki makna yang sama, dapat diakibatkan dari penulisan yang salah, penyingkatan kata, ataupun Bahasa gaul. Kata-kata singkatan tersebut dimasukkan ke dalam suatu dokumen teks/text yang nanti akan dibaca oleh kode pemograman. Proses ini bertujuan untuk menyederhakan teks sehingga lebih mudah untuk dianalisis, diproses maupun dibandingkan dengan teks lainnya.

d. Stopword Removal

Stopword Removal berfungsi untuk menghapus kata yang tidak ada makna, seperti kata umum yang tidak memiliki nilai. Dalam tahap ini, kata yang tidak penting terhadap makna dokumen akan dihilangkan.

e. Stemming

Proses ini yaitu mengembalikan kata yang berimbuhan menjadi kata induk. Ini dilakukan untuk mengurangi kesalahan pada proses selanjutnya. Untuk stemming dalam bahasa indonesia menggunakan library python sastrawi. Sastrawi merupakan library yang dapat mengubah kata berimbuhan Bahasa indonesia menjadi kata dasar.

3.3 Visualisasi Data

Visualisasi data melibatkan representasi grafis dari informasi yang terdapat dalam teks atau data teks yang telah diproses. Ini membantu dalam pemahaman, analisis, dan komunikasi hasil dari algoritma pemrosesan bahasa alami. Pada tahap ini akan dilakukan visualisasi dalam bentuk WordCloud dan BarChart.

3.4 Pembobotan TF-IDF

Pada tahap ini dilakukan pembobotan kata dari hasil stemming dengan metode Term Inverse Document Frequency (TF-IDF). Metode TF-IDF digunakan untuk mengetahui seberapa sering suatu kata muncul di dalam dokumen. Tahap ini dibantu dengan library sklearn.

3.5 *Klasifikasi*

Klasifikasi ulasan pengguna dilakukan menggunakan algoritma Support Vector Machine dan Naïve Bayes Classifier yang akan dibantu dengan library Scikit-Learn. Proses klasifikasi menggunakan nilai data latih dan data uji sebesar 80% : 20%, dilakukan percobaan sebanyak 15 kali dan kernel akan menggunakan kernel rbf.

3.6 Evaluasi Kinerja

Model algoritma akan dievaluasi menggunakan confusion matrics dengan melakukan perhitungan accuracy, precission, recall, dan fl-score.

a. *Accuracy*: merupakan hasil pengujian berdasarkan tingkat pendekatan antara nilai aktual dengan nilai prediksi.
b. *Precision*: merupakan nilai kecocokan data antara data yang dimodelkan kedalam metode dengan data asli (data analisis sentimen) dalam pengklasifikasian.
c. *Recall*: merupakan nilai prediksi dalam pernyataan benar antara data yang dimodelkan kedalam metode dengan data asli.
d. F1-*Score*: merupakan nilai rata-rata antara precission dan recall yang dibobotkan.

Hasil dari pengujian ini akan dibandingkan untuk melihat nilai tingkat akurasi dan kinerja dari dua algoritma tersebut.