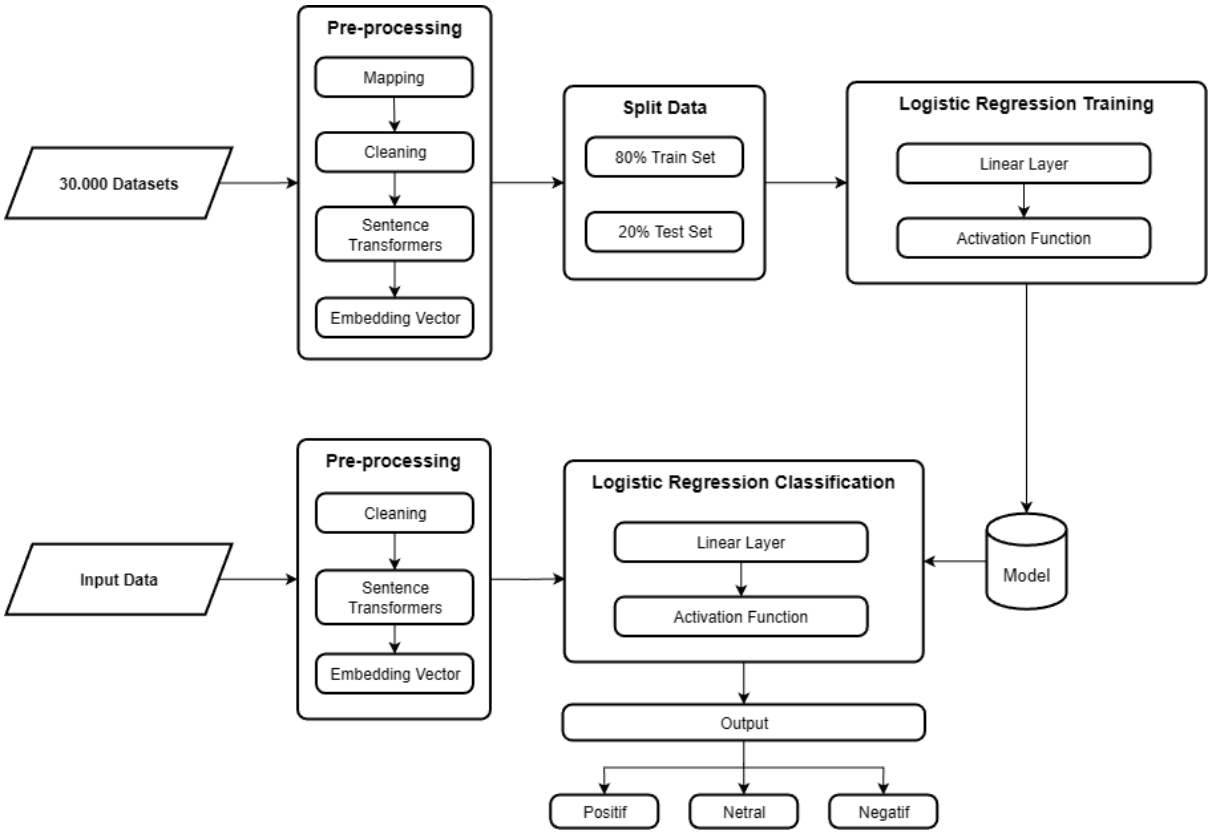
Pipeline Analisis Sentimen Ulasan Aplikasi Gojek – Dimas Dwi Aji



Pipeline analisis sentimen terdiri dari dua tahap utama, yaitu tahap pelatihan model (training) dan tahap inferensi (classification). Kedua tahap tersebut memiliki alur preprocessing yang hampir sama untuk menjaga konsistensi distribusi data. Berikut penjelasan lengkap tiap bagian.

1. Tahap Pelatihan Model

A. Input Dataset

Sistem menerima dataset ulasan sebanyak 30.000 data yang telah dilabel berupa rating aplikasi (skala 1-5).

B. Pre-processing

Tahap preprocessing bertujuan menyiapkan data teks sebelum masuk ke model. Proses ini terdiri dari beberapa langkah:

- Mapping, mengubah skor rating menjadi representasi kategori untuk kebutuhan model.
- Cleaning, proses pembersihan teks dengan menghapus karakter yang tidak diperlukan, seperti tanda baca, angka, emoji, URL, mention, dan karakter spesial. Selain itu pada tahap ini juga dilakukan normalisasi teks (lowercase, hapus spasi berlebih).
- Sentence Transformers, teks yang telah dibersihkan dimasukkan ke model Sentence Transformer untuk menghasilkan representasi vektor (embedding) yang lebih bermakna secara semantik.

- Embedding Vektor, setiap teks diubah menjadi vektor berdimensi yang menjadi input untuk model Logistic Regression.

C. Split Data

Data embedding kemudian dibagi menjadi dua bagian:

- 80% sebagai data latih (Train Set).
- 20% sebagai data uji (Test Set).

Pembagian ini digunakan untuk memastikan proses pelatihan dapat dievaluasi dengan benar.

D. Logistic Regression Training

Data latih diproses menggunakan model Logistic Regression yang terdiri dari dua komponen utama:

- Linear Layer, lapisan linier menghitung nilai prediksi berdasarkan bobot dan bias.
- Activation Function, fungsi aktivasi softmax digunakan untuk menghasilkan probabilitas pada tiga kelas sentimen.

Model hasil pelatihan kemudian disimpan untuk digunakan pada tahap inferensi.

2. Tahap Inferensi Model

A. Input Data

User memasukkan teks baru yang sentimennya ingin diprediksi.

B. Pre-processing

Tahap preprocessing di inferensi sama dengan proses saat training agar distribusi data tetap konsisten. Langkah-langkahnya:

- Cleaning, teks dibersihkan menggunakan aturan yang sama seperti tahap pelatihan.
- Sentence Transformers, teks yang telah dibersihkan dimasukkan ke model Sentence Transformer yang sama untuk menghasilkan vector embedding.
- Embedding Vektor, hasil embedding menjadi input bagi model klasifikasi.

C. Logistic Regression Classification

Embedding vector diproses oleh model yang telah disimpan:

- Linear Layer, embedding diproyeksikan ke tiga kelas melalui bobot model.
- Activation Function, fungsi aktivasi softmax menghasilkan probabilitas sentimen.

D. Output

Model menghasilkan prediksi salah satu dari tiga kelas sentimen:

- Positif
- Netral
- Negatif