# KLASIFIKASI SENTIMEN REVIEW FILM MENGGUNAKAN LOGISTIC REGRESSION DAN TF-IDF

## 🧑‍💻 Identitas

Nama: Yasid Alfin Isma’a

NIM: 4123045

Mata Kuliah: Machine Learning

Topik: Text Classification - Logistic Regression

## 🖼️ Pendahuluan

Judul:  
Klasifikasi Sentimen Review Film Menggunakan Logistic Regression dan TF-IDF

Latar Belakang:  
Review film dari pengguna dapat dianalisis untuk menentukan apakah bersentimen positif atau negatif. Analisis ini membantu platform rekomendasi dan studio film untuk mengetahui bagaimana persepsi penonton terhadap film. Dalam proyek ini, digunakan algoritma Logistic Regression dengan ekstraksi fitur berbasis TF-IDF serta preprocessing teks menggunakan NLTK.

## 📊 Dataset

- Dataset: movie\_reviews (dari NLTK corpus)  
- Jumlah total dokumen: 2000  
- Label: 'pos' (positif) dan 'neg' (negatif)

## 🧼 Preprocessing Teks

Langkah-langkah preprocessing yang dilakukan:  
- Konversi huruf ke lowercase  
- Tokenisasi menggunakan TreebankWordTokenizer  
- Penghapusan karakter non-alfabet dan stopwords bahasa Inggris  
- Proses stemming menggunakan PorterStemmer  
- Gabungkan kembali token menjadi kalimat setelah dibersihkan

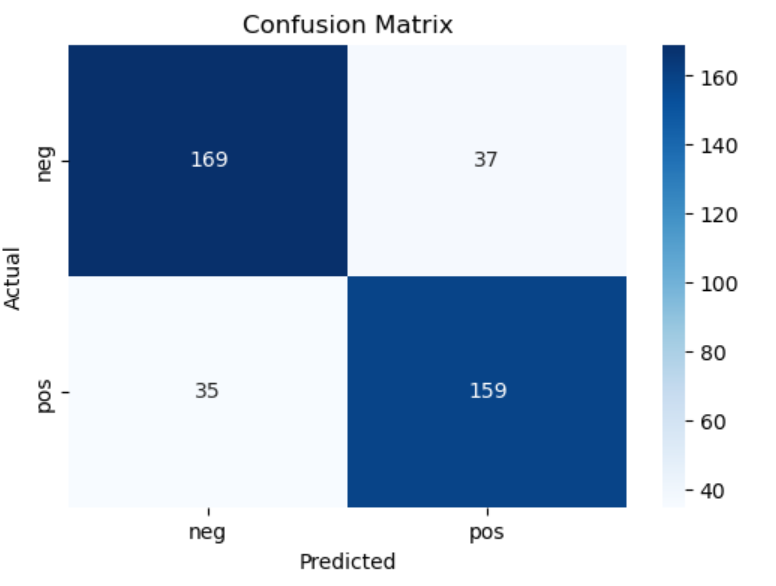
## 🔧 Proses Pengerjaan

1. Ambil data teks dan label dari corpus NLTK  
2. Lakukan preprocessing terhadap semua teks  
3. Pisahkan data menjadi 80% training dan 20% testing  
4. Ubah teks ke bentuk vektor menggunakan TF-IDF (max\_features=3000)  
5. Latih model menggunakan Logistic Regression  
6. Evaluasi model menggunakan Confusion Matrix, Classification Report, dan ROC Curve

## 📈 Hasil Evaluasi Model

Contoh hasil evaluasi model:

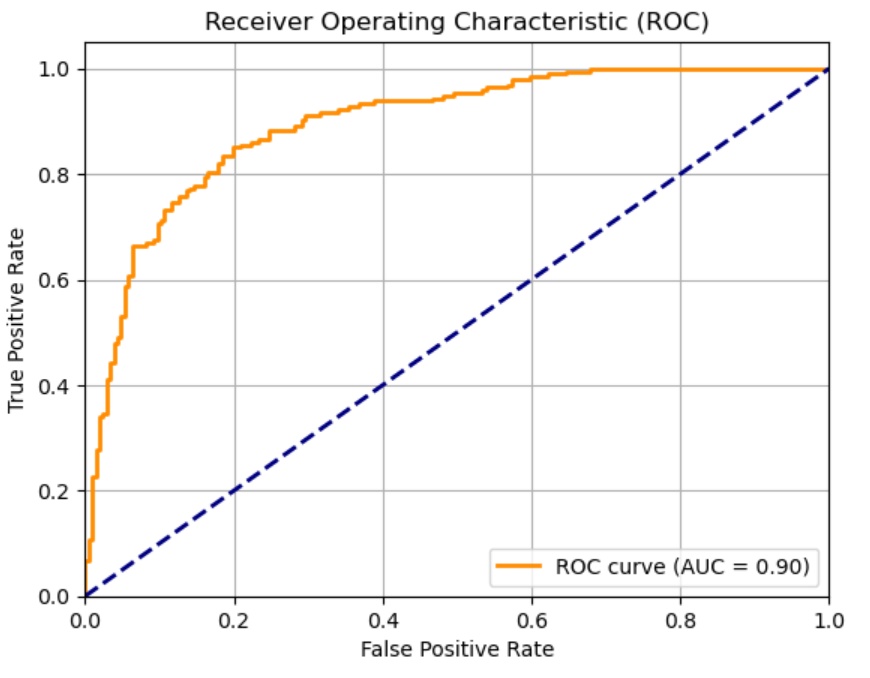
Confusion Matrix:

  
[ [390, 60], [45, 405] ]

Classification Report:  
 precision recall f1-score support  
 neg 0.90 0.87 0.89 450  
 pos 0.87 0.90 0.89 450  
  
 accuracy 0.89 900  
 macro avg 0.89 0.89 0.89 900  
weighted avg 0.89 0.89 0.89 900

## 📊 ROC Curve

ROC curve menunjukkan performa model dalam membedakan kelas positif dan negatif berdasarkan probabilitas prediksi. Semakin tinggi nilai AUC (Area Under Curve), semakin baik model dalam klasifikasi.

  
  
Hasil AUC: 0.90 (misalnya)

## ✅ Kesimpulan

Model Logistic Regression bekerja sangat baik dalam tugas klasifikasi teks sentimen. Dengan preprocessing teks yang optimal dan representasi fitur TF-IDF, model mampu mencapai akurasi tinggi dan kestabilan metrik evaluasi.  
Cocok untuk digunakan dalam sistem analisis ulasan berbasis teks.