

Ringkasan Evaluasi Model

Rakagali Resda Krisandi Putra

PROYEK

SmartReceipt

SAMPEL

6.026 baris

PRODUK UNIK

1.419

KATEGORI

15 → 9 kanonikal

PEMBAGIAN DATA

BAGIAN

BARIS

PROPORSI

Data latih

4.474

74,2%

Data validasi

762

12,6%

Data uji

790

13,1%

MENGAPA DIPISAH PER PRODUK

Pembagian data dilakukan berbasis produk unik, bukan baris acak. Produk yang muncul di data uji dipastikan tidak pernah muncul di data latih. Tanpa aturan ini, akurasi akan terlihat tinggi tetapi menyesatkan, karena model sebenarnya hanya menghafal produk yang sudah dilihatnya.

KONFIGURASI MODEL

PARAMETER

NILAI

Vektor fitur

TF-IDF char_wb, n-gram 2–4

Pembobotan

sublinear TF, normalisasi L2

Pengklasifikasi

Multinomial Naive Bayes

Jalur tinjauan

Aktif saat selisih skor top-1 dan top-2 terlalu kecil

TEMUAN UTAMA

- Kesalahan terkonsentrasi pada kategori yang mirip, terutama Makanan dan Minuman.
- Teks OCR yang sangat pendek paling sering memicu jalur tinjauan, dan itu perilaku yang diinginkan.
- Koreksi pengguna menambah data latih dan memperbaiki kategori yang sering salah.

YANG AKAN DIPERBAIKI

Data uji sebaiknya berasal dari sumber berbeda sejak awal, bukan hanya dipisahkan dari data yang sama. Ini akan memberi gambaran yang lebih jujur tentang performa di dunia nyata.

Catatan kejujuran. Dokumen ini direkonstruksi dari arsip proyek untuk keperluan portofolio. Isinya merangkum keputusan dan angka yang benar-benar tercatat, bukan dokumen asli yang dibuat pada saat proyek berjalan.

