Indonesian Text Sentiment Analysis Menggunakan IndoBERT Transformer

Bintang Fajar Julio – Politeknik Negeri Jakarta

Table of contents

01

Dataset

Dataset Source

- https://huggingface.co/datasets/sepidmnorozy/Indonesian_sentiment

label

text

- Label 0: Sentimen Negatif
- Label 1: Sentimen Positif

Dataset Size

7926

Train Data

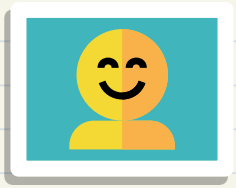
1132

Validation Data

2266

Test Data

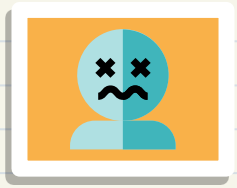
Dataset Value



Sebaran Data per Label

Label 0: 4005 data

Label 1: 7319 data



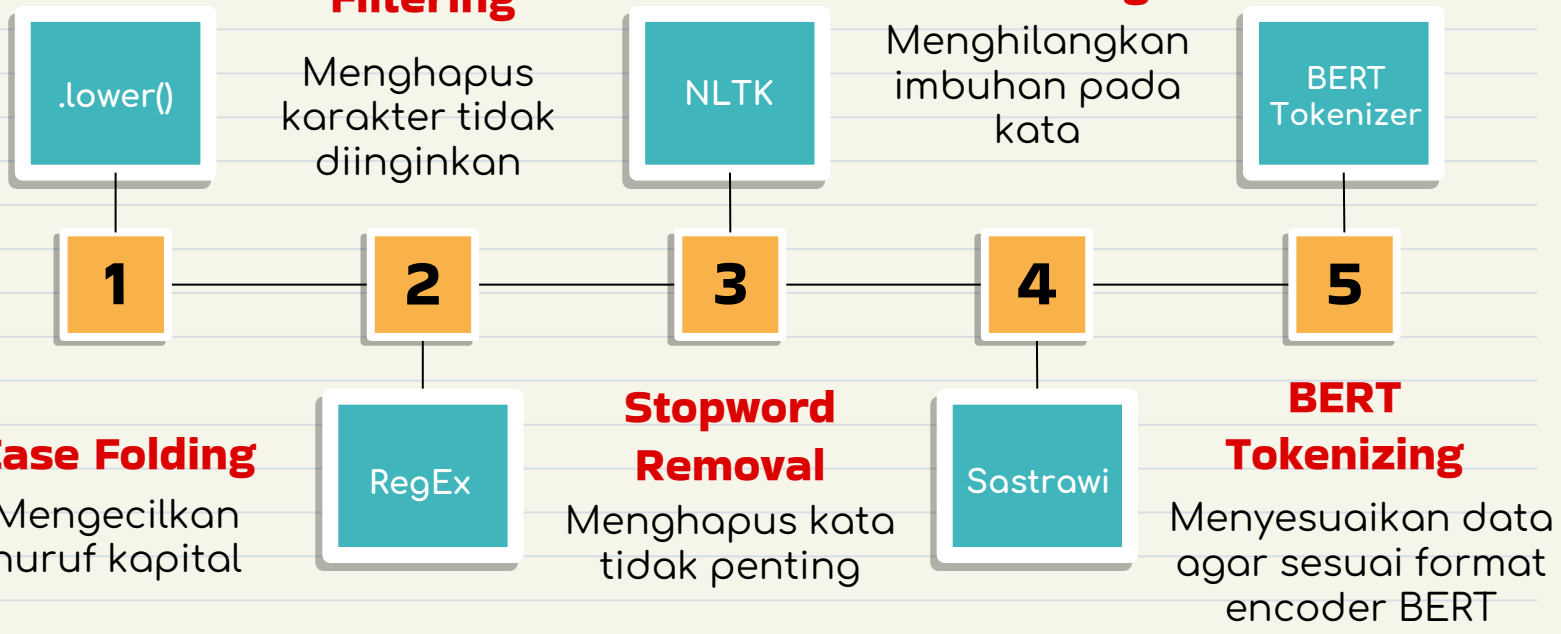
NaN

Dataset tidak memiliki data NaN

02

Preprocessing

Tahapan Preprocessing



BERT Tokenizing [Pretrained Tokenizer]

Memecah kalimat menjadi kata

Memecah isi teks pada kalimat agar menjadi per kata atau token

Menyelipkan token spesial

Token [CLS] sebagai tanda awal kalimat, [SEP] sebagai tanda akhir kalimat

Padding & Truncation

Mensimetriskan jumlah token pada baris sesuai panjang yang diinginkan (padding mengisi kekurangan token & truncation untuk memotong token berlebih)

Menyesuaikan format encoder

Format Encoder BERT:

- input_ids: Mengkonversi token menjadi id berdasarkan korpus dari pretrained tokenizer
- token_type_ids: Informasi id token untuk decoder (tidak digunakan)
- attention mask: Binary label untuk membedakan apakah token adalah id atau hanya padding

Konversi Hasil Preprocessing ke Tensor

Rubah label menjadi binary list

[Negatif, Positif]
Label 0 = [1, 0]
Label 1 = [0, 1]

Gabung menjadi Tensor Dataset

List dari input_ids, token_type_ids, attention mask, binary label dirubah menjadi tensor lalu disatukan menggunakan PyTorch

03

Modeling

Modeling Stage

Training Step

7926 Train Data

Validation Step

1132 Validation Data

Test Step

2266 Test Data

Modeling Monitor

Accuracy

nilai yang digunakan untuk menentukan tingkat keberhasilan model yang telah dibuat

Loss

ukuran kesalahan yang dibuat oleh model

Pytorch Trainer Setup

Setup	Value
Batch Size	32
Max Length	100
Max Epochs	12
Optimizer	Adam
Learning Rate	2e-5
Early Stopping Monitor	val_loss
Early Stopping Patience	3

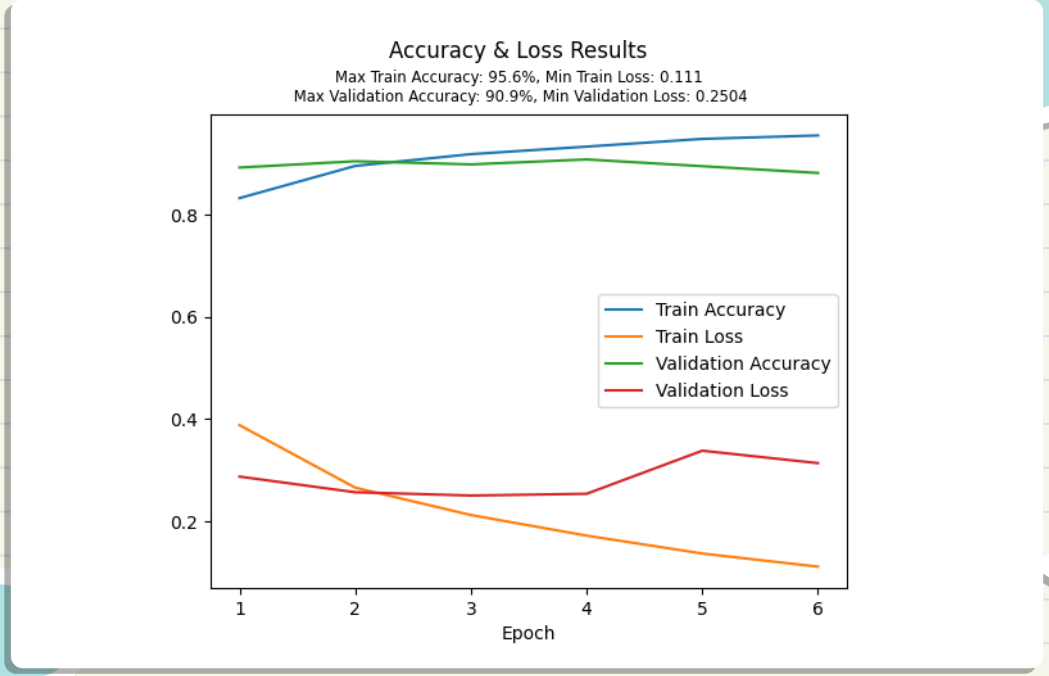
Model Pretrained IndoBERT Setup

Setup	Value
Embedding Size	768
Hidden Size	768
Dropout	0.3
Loss Function	BCEWithLogitsLoss
Activation Function	Tanh
Pretrained Model	indolem/indobert-base-uncased
Accuracy Report	sklearn

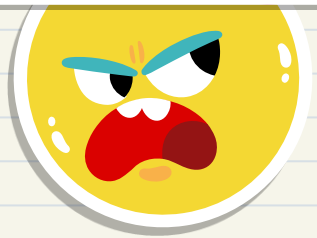
04

Result

Train & Validation Results

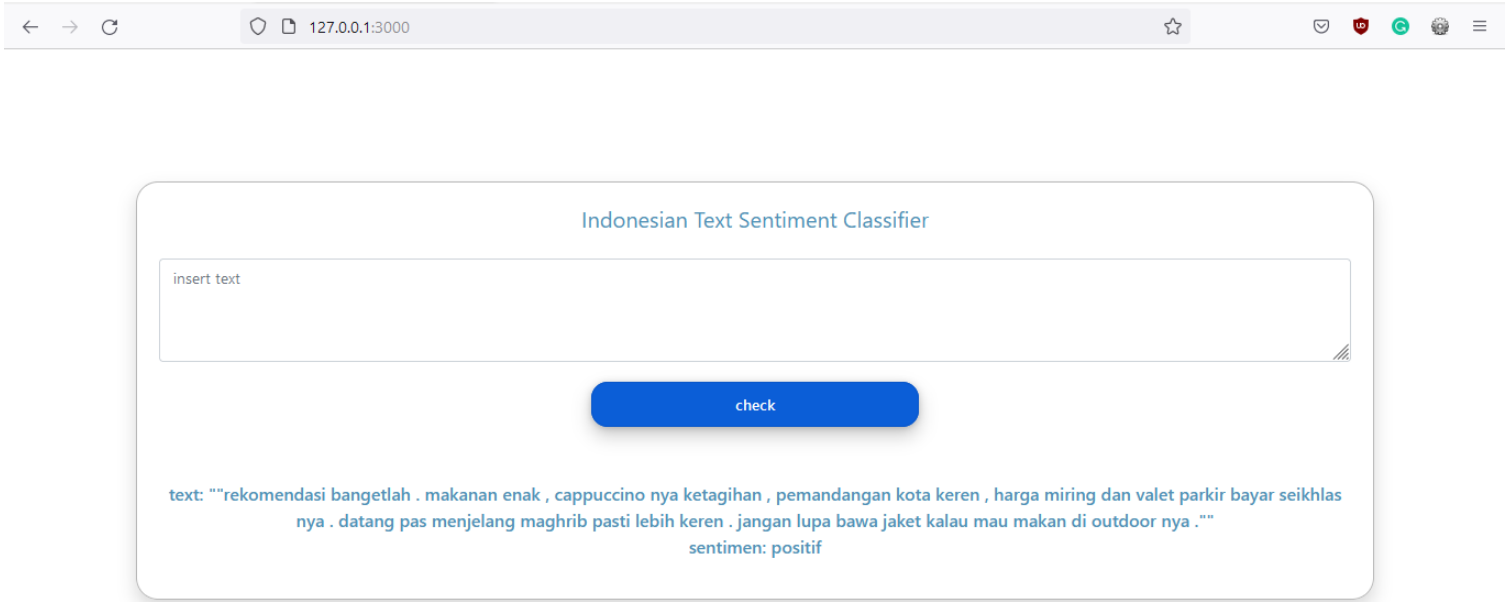


Test Step Result



0.260	89.37%
Test Loss	Test Accuracy

Flesk Web Deployment



Thanks!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**