

ANALISIS SENTIMEN TWEET: MENGUNGKAP SPEKTRUM EMOSIONAL MEDIA SOSIAL

Sentiment API using LSTM and NN Model

BY:

KENTA EDMONDA

NINDITYA SALMA NUR AINI



PENDAHULUAN

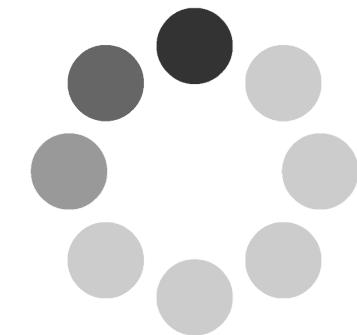
LATAR BELAKANG

Sentimen analisis, juga dikenal sebagai analisis opini, adalah teknik pengolahan bahasa alami yang digunakan untuk mengidentifikasi, mengekstraksi, dan memahami sentimen atau opini yang terkandung dalam teks atau data.

Sentimen ini dapat bersifat positif, negatif, atau netral. Tujuan utama dari sentimen analisis adalah untuk memahami bagaimana orang merasakan atau berpendapat tentang suatu subjek, topik, produk, layanan, atau entitas lain yang diungkapkan dalam teks.

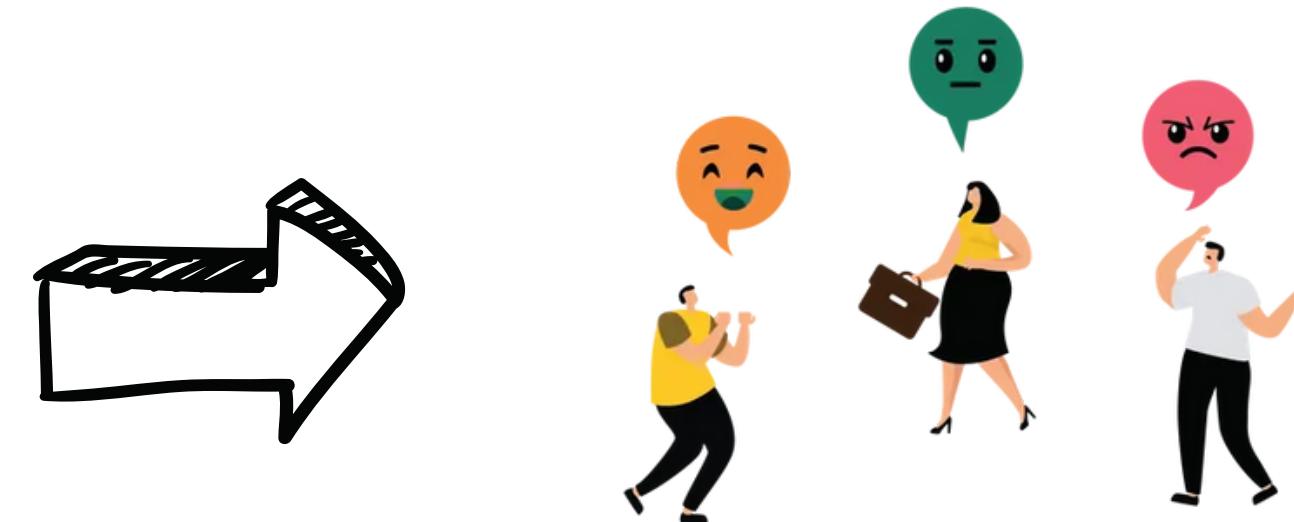


Kenta Edmonda @edmonda_kenta · 1s
Makanan ini layak diberi nilai 100!



TUJUAN

- Membuat API untuk melakukan *cleaning data* dan mengklasifikasi sentimen dari teks
- Integrasi API menggunakan Flask dan Swagger
- Mencari model dengan performa paling baik antara LSTM dan NN
- Terakhir, sekilasi tentang *Exploratory Data Analysis*



METODE PENELITIAN

DATA PREPARATION

- Dataset berisi 11000 baris dengan 2 kolom yang berisi kategori positive, neutral, dan negative
- Pada label Tweet diisi dengan nilai berupa kalimat yang dikumpulkan dari twitter
- Tidak ditemukan *missing values* pada dataset.
- Terdapat 67 data duplikat.
- Data duplikat diselesaikan dengan melakukan dropping duplicated data, guna mengatasi permasalahan pada modelling

	Tweet	Label
0	warung ini dimiliki oleh pengusaha pabrik tahu...	positive
1	mohon ulama lurus dan k212 mmbri hujah partai...	neutral
2	lokasi strategis di jalan sumatera bandung . t...	positive
3	betapa bahagia nya diri ini saat unboxing pake...	positive
4	duh . jadi mahasiswa jangan sompong dong . kas...	negative

df.shape
(11000, 2)

```
df.duplicated().sum()
```

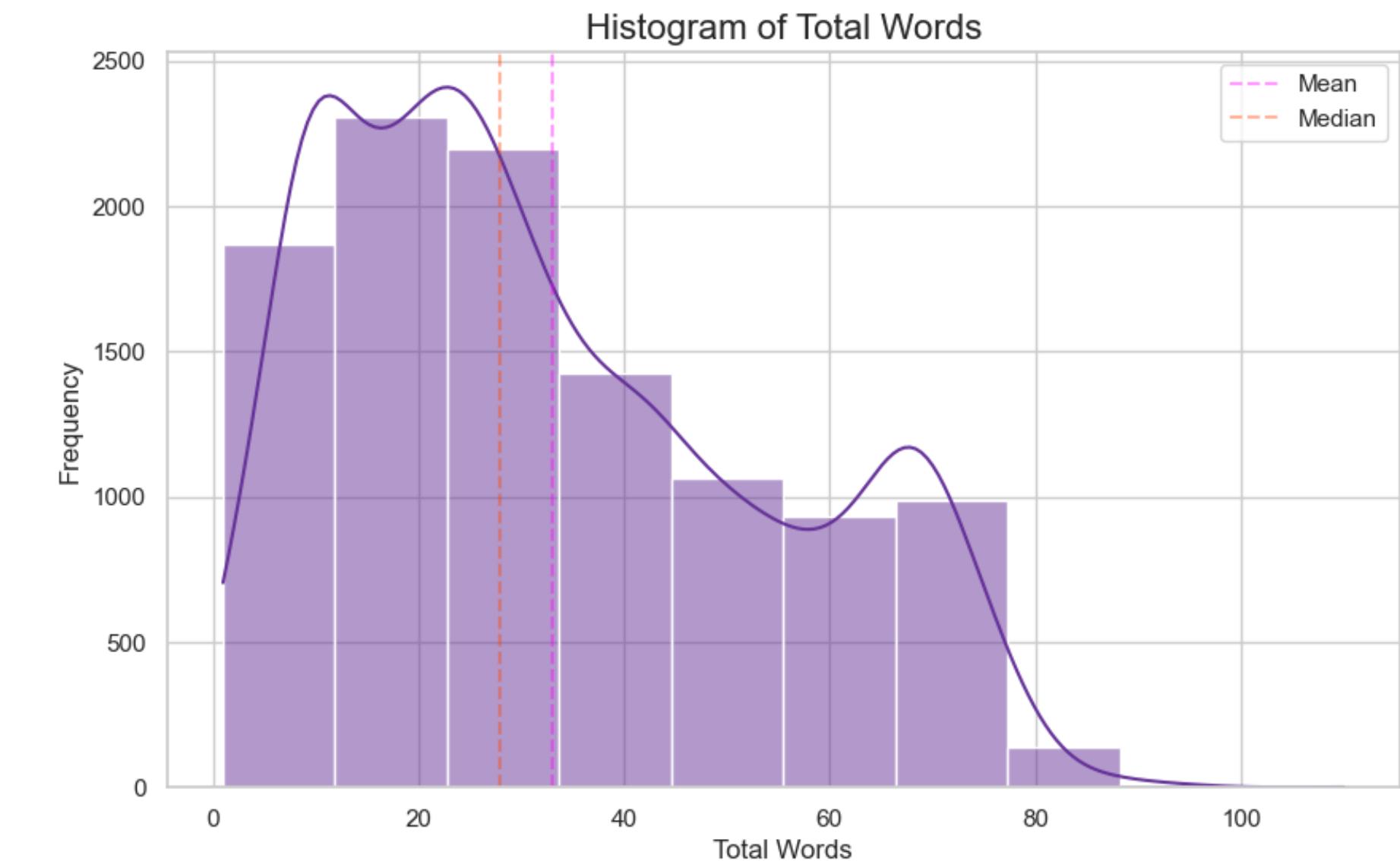
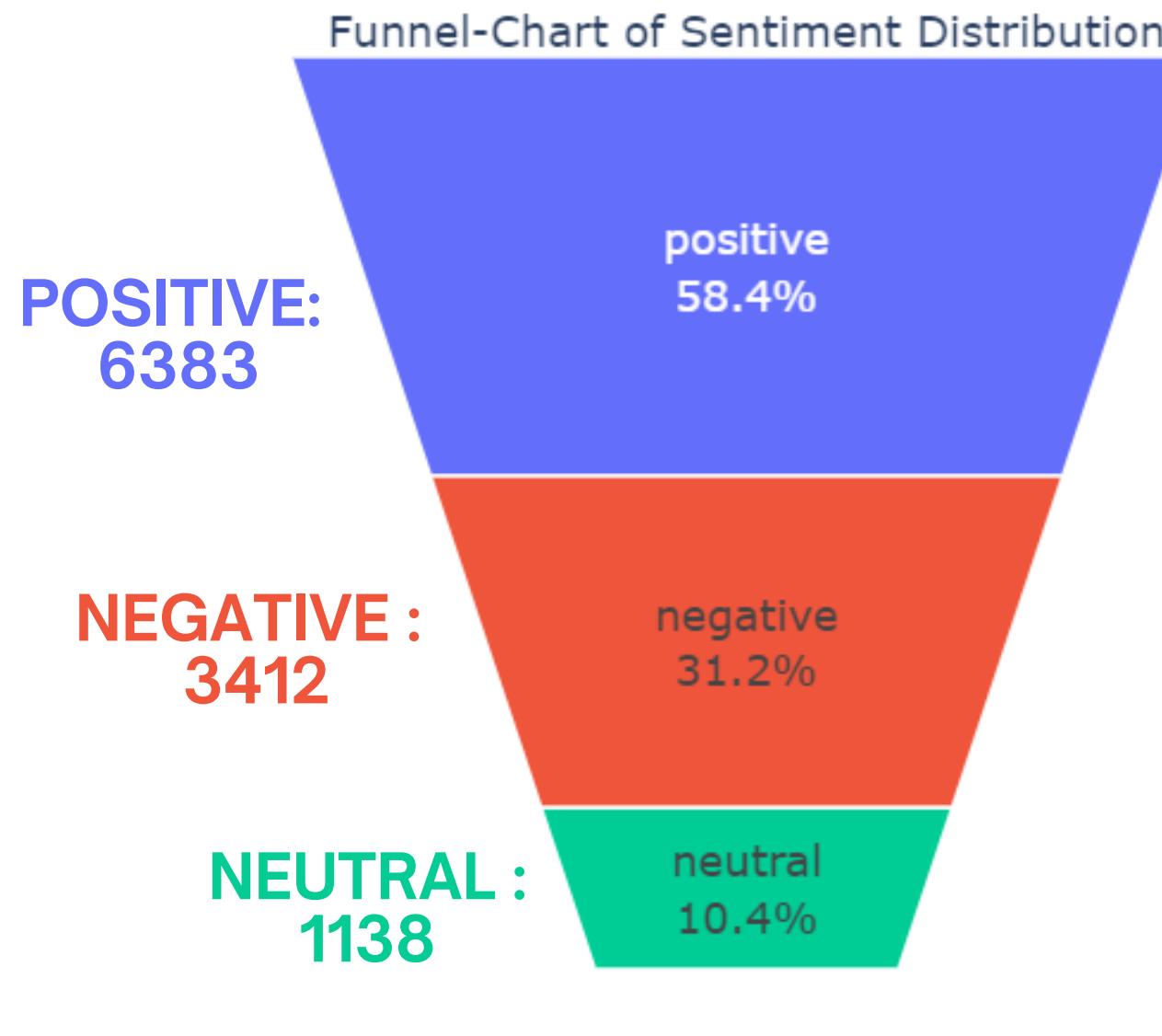
67

```
df.drop_duplicates(inplace=True)  
df.duplicated().sum()
```

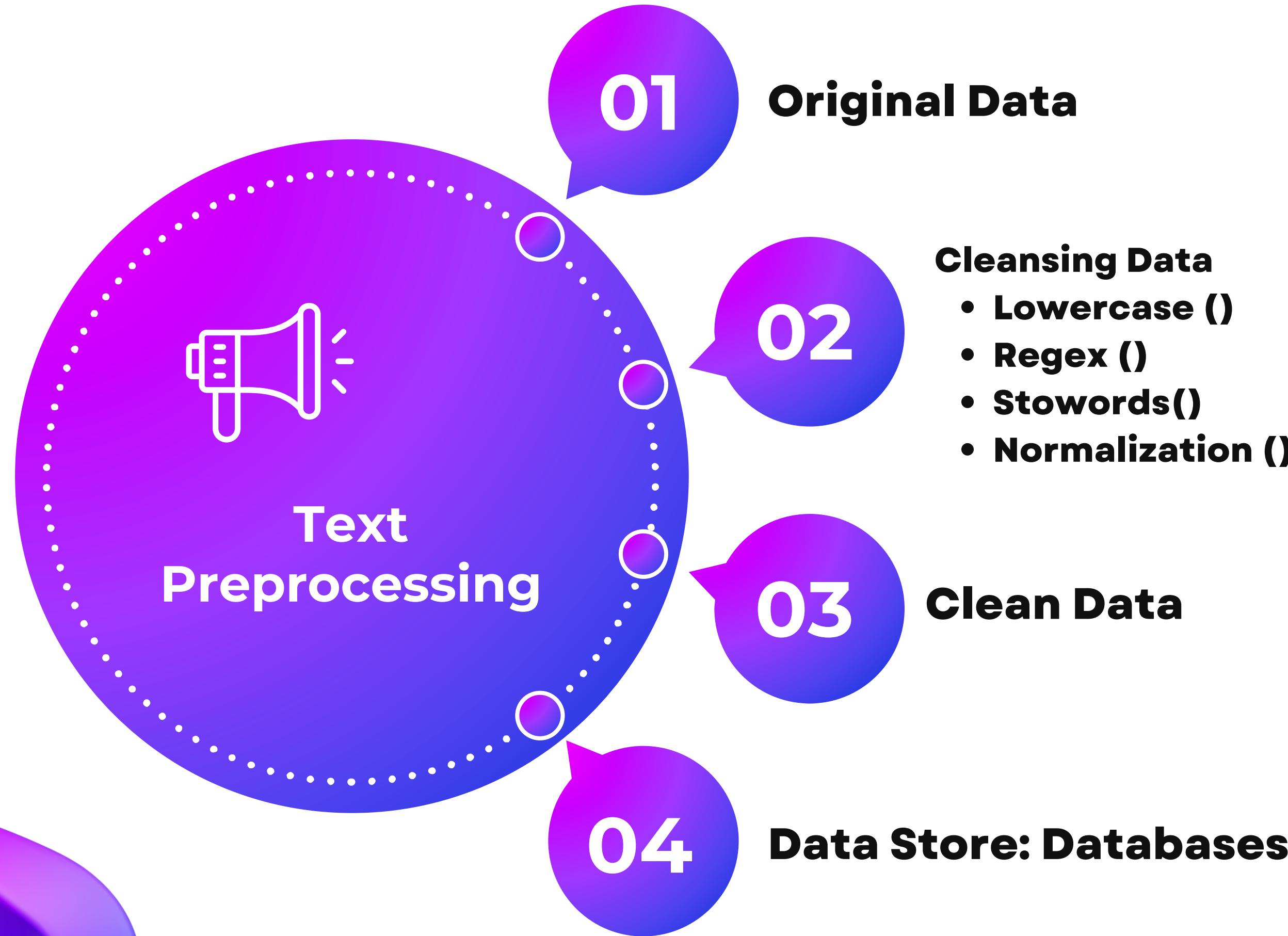
0

METODE PENELITIAN

EXPLORATORY DATA ANALYSIS



METODE PENELITIAN



METODE PENELITIAN

MACHINE LEARNING PREPARATION



**Handling
Imbalanced
Dataset**

Over Sampling



**Feature Label
Classification**

**Memisahkan
Label
X dan Y**



**Feature
Extraction**

- Tokenizer dan Pad Sequences for LSTM
- TFIDF or BoW for NN



**Modeling
(Train-Test
Split Data)**

- LSTM-Model
- NN - Model



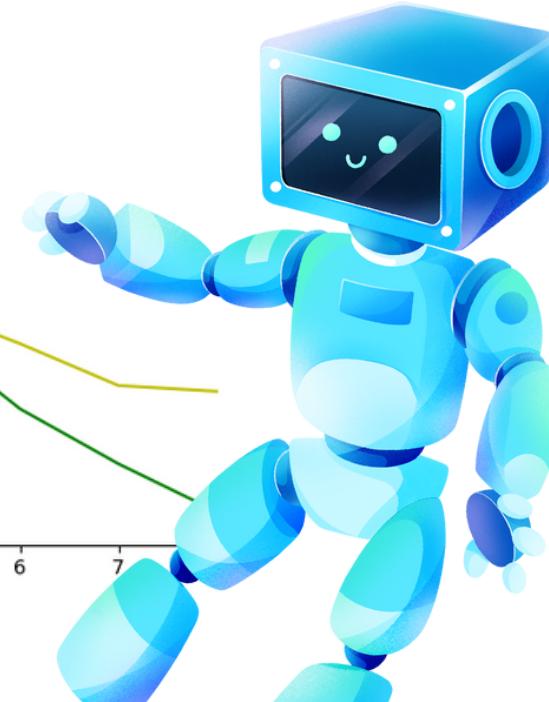
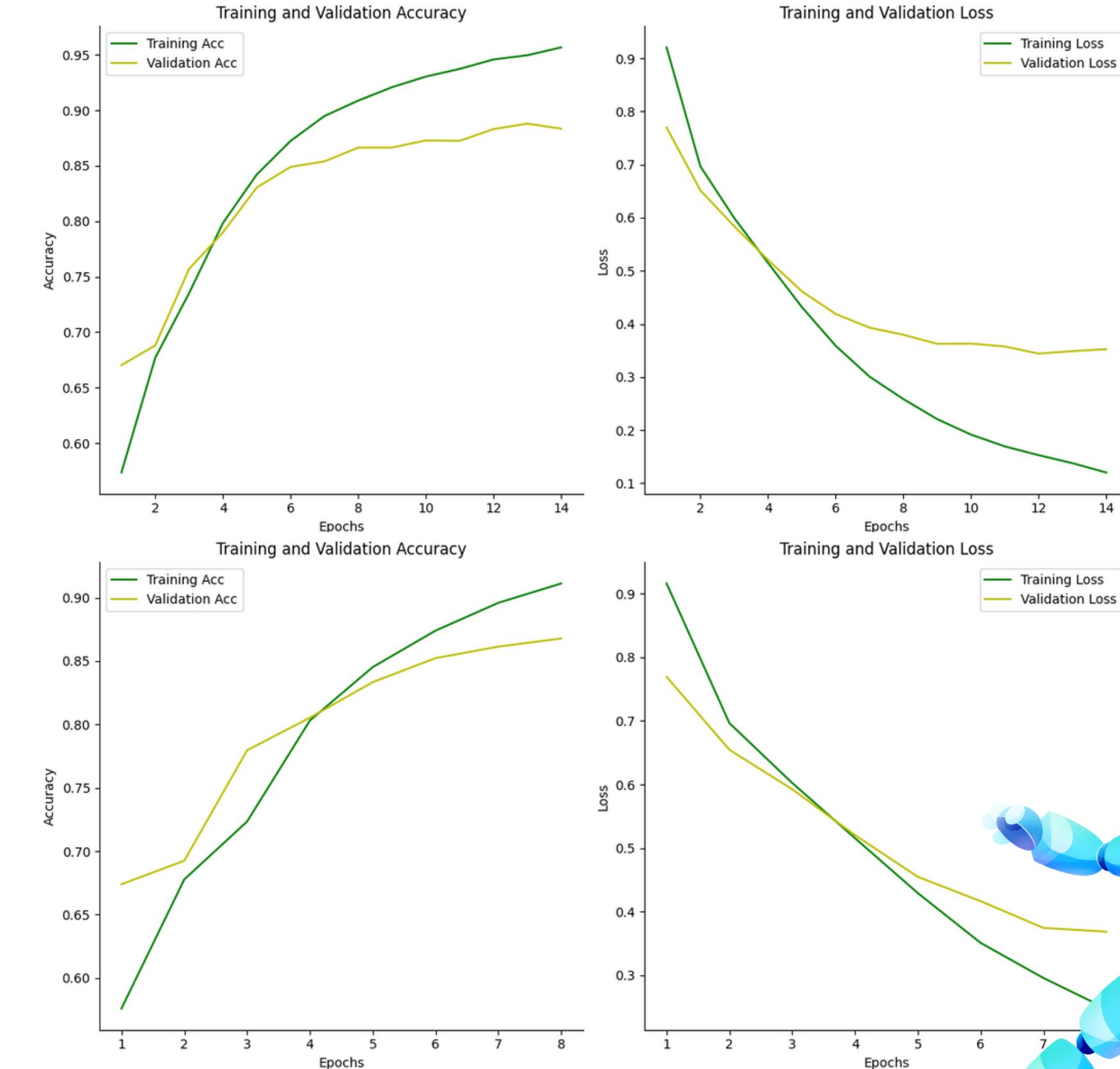
**Evaluation
Model**

- Confusion Matrix
- K-Fold Cross Validation

HASIL PENELITIAN

LSTM MODEL TRAINING & EVALUATION

PARAMETERS	
Input Layer	64
Output Layer	3
Activation	Softmax
Epoch Optimal	50
Dropout	0.2
Early Stopping	8
Monitor	Val_Loss
Mode	Min
Cross Validation	5



HASIL PENELITIAN

LSTM MODEL TRAINING & EVALUATION

Hasil Nilai Precision, Recall, F1-Score

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1040
1	0.97	0.95	0.96	958
2	0.96	0.93	0.94	1965
accuracy			0.93	3963
macro avg	0.92	0.93	0.93	3963
weighted avg	0.93	0.93	0.93	3963
=====				
Rata-rata Akurasi: 0.915				

HASIL PENELITIAN

NN MODEL TRAINING

Feature Extraction

```
Input Text: Dia memakan kue dengan lahap dan beringas jelek
Text Preprocessing: dia memakan kue dengan lahap dan beringas jelek
Extraction Feature:
(0, 8364)    0.523104538524453
(0, 7351)    0.523104538524453
(0, 7272)    0.3760260390845009
(0, 5925)    0.39780926453799464
(0, 3041)    0.33927842050202
(0, 2983)    0.15391011669079221
(0, 2853)    0.11948773949774336
```

```
Input Text: Dia memakan kue dengan lahap dan beringas jelek
Text Preprocessing: dia memakan kue dengan lahap dan beringas jelek
Extraction Feature:
(0, 5834)    1
(0, 7154)    1
(0, 7229)    1
(0, 8223)    1
|
```

Modeling with Best Parameters

```
Best parameters of Model NN-Bow: {
    'algoritma_activation': 'tanh',
    'algoritma_alpha': 0.01,
    'algoritma_early_stopping': True,
    'algoritma_hidden_layer_sizes': {10},
    'algoritma_learning_rate_init': 0.01
}
```

```
Best parameters of Model NN-TFIDF: {
    'algoritma_activation': 'relu',
    'algoritma_alpha': 0.1,
    'algoritma_early_stopping': True,
    'algoritma_hidden_layer_sizes': {10},
    'algoritma_learning_rate_init': 0.01
}
```

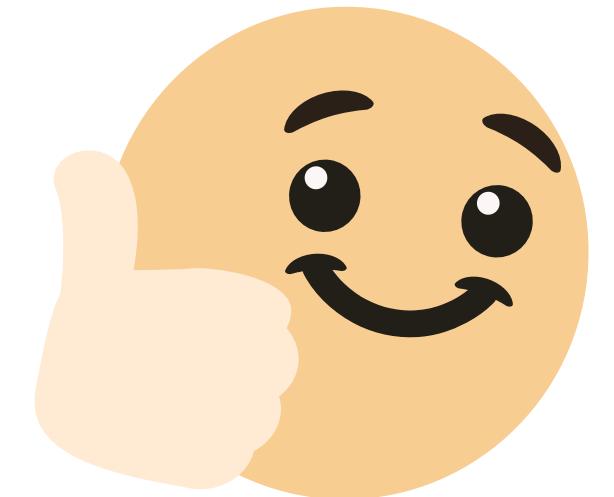
HASIL PENELITIAN

NN MODEL EVALUATION

Model Evaluasi	Label	Accuracy	
		TF-IDF	BoW
Data Train	Negative	0.97	0.96
	Neutral	0.98	0.97
	Positive	0.98	0.98
	Average	0.98	0.97
Data Test	Negative	0.81	0.82
	Neutral	0.78	0.79
	Positive	0.91	0.91
	Average	0.86	0.87
kfold = 5 Cross Validation	Average	0,864	0,874

	Nilai Delta (Data Train- Data Test)			
Label	Negative	Neutral	Positive	Average
TF-IDF	0,16	0,2	0,07	0,12
BoW	0,14	0,18	0,07	0,1

Nilai Delta < 0.25 != Overviting



HASIL PENELITIAN

NN MODEL - PREDICT API

Request URL

```
http://127.0.0.1:5000/NN_text
```

Server response

Code	Details
200	Response body

```
{  
    "data": {  
        "sentimen": "negative",  
        "text": "Selamat Pagi"  
    },  
    "description": "Hasil Prediksi NN Sentimen",  
    "status_code": 200  
}
```

Response headers

```
connection: close  
content-length: 144  
content-type: application/json  
date: Mon13 Nov 2023 13:54:46 GMT  
server: Werkzeug/2.3.7 Python/3.11.6
```

Request URL

```
http://127.0.0.1:5000/NN_file
```

Server response

Code	Details
200	Response body

```
{  
    "data": {  
        "sentimen": [  
            "negative",  
            "negative",  
            "negative"  
        ],  
        "tulisan": [  
            "hari ini cuaca nya indah tapi aku sedih",  
            "bahagia itu sederhana",  
            "berlari-lari di tengah keramaian"  
        ]  
    },  
    "description": "Hasil Prediksi NN Sentimen",  
    "status_code": 200  
}
```

Response headers

```
connection: close  
content-length: 311  
content-type: application/json  
date: Mon13 Nov 2023 13:56:01 GMT  
server: Werkzeug/2.3.7 Python/3.11.6
```

HASIL DAN KESIMPULAN

SENTIMENT ANALISIS LSTM DAN NN MODEL

- Dataset terdiri dari tweet bernada positif, netral, dan negatif dengan sebaran tweet berlabel positif sebanyak 6.383 (58.4%) tweet, berlabel negatif sebanyak 3.412 (31%) tweet, dan label netral sebanyak 1.138 (10.4%) tweet
- Model LSTM memiliki performa yang lebih baik dibandingkan model NN. Nilai akurasi tebesar dari model LSTM sebesar 0.915 sedangkan akurasi terbesar dari Model NN sebesar 0.874. Nilai akurasi ini didapat dari pengujian KFold Cross Validation K = 5 masing-masing model.
- API yang dibuat memiliki 2 endpoint untuk setiap model (untuk memproses teks dan file data) dan dapat memberikan label positif, negatif atau netral berdasarkan sentimen disediakan