

TWEET SENTIMENT ANALYSIS

PEMBUATAN API DENGAN PEMODELAN MACHINE LEARNING
DAN DEEP LEARNING DAN ANALISA DATA

Kelompok 3

1. Yaya Hidayana
2. Ghifari Pangripta N
3. Alex Apriandi
4. Fional Khalik

Tweet

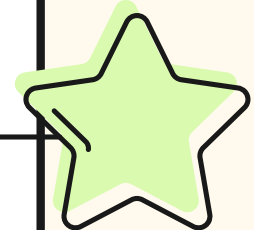
ooo

Neutral



Negative

Positive
Feedback



LATAR BELAKANG

1

INTERNET DAN SOSIAL MEDIA

Pertumbuhan internet di Indonesia sangat pesat dalam beberapa tahun terakhir. Menurut data survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), penetrasi pengguna internet di Indonesia telah mencapai 78,19 persen pada 2023 atau menembus 215.626.156 jiwa dari total populasi yang sebesar 275.773.901 jiwa.

Laporan **We Are Social** menunjukkan, jumlah pengguna aktif media sosial di Indonesia sebanyak 167 juta jiwa pada Januari 2023, setara dengan 60,4% dari populasi di dalam negeri.

2

APA ITU SENTIMEN?

Sentimen adalah pendapat atau pandangan mengenai sesuatu. Sementara itu, analisis sentimen itu sendiri adalah proses menganalisis teks digital untuk menentukan apakah nada emosional dari pesan tersebut positif, negatif, atau netral, termasuk teks di media sosial (Twitter).

3

MANFAAT SENTIMEN ANALISIS

Penerapan analisis sentimen tidak terbatas pada media sosial. Banyak perusahaan menggunakan analisis sentimen untuk meningkatkan produk atau layanan mereka berdasarkan ulasan khusus pelanggan. Untuk itu, kami mencoba menganalisis sentimen pada data tweet yang sudah ada dan membuat sistem (API) dengan metode machine learning menggunakan Neural Network (NN) deep learning Long-Short Term Memory (LSTM) guna mendapatkan perbandingan.

LATAR BELAKANG



4

RUMUSAN MASALAH

- Bagaimana gambaran nilai sentiment dari tweet para pengguna twitter?
- Bagaimana menentukan model dengan performa yang baik?
- Bagaimana melakukan cleansing atas data dan membangun mesin/API?

5

TUJUAN

- Mendapatkan nilai sentiment dari tweet para pengguna twitter.
- Mendapatkan model dengan performa terbaik untuk memprediksi sentimen.
- Membuat mesin/API yang dapat mengklasifikasikan sentimen dari data yang ada.

METODE PENELITIAN





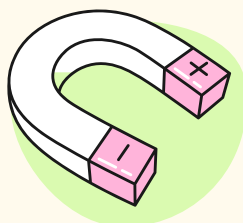
DATA

2 jenis data sekunder yang kami gunakan untuk menyelesaikan proyek ini, yaitu:

1. Daya untuk membuat model (`train_preprocess.tsv.txt`), yang terdiri dari **11.000 baris** dan **2 kolom**.
2. Data untuk melakukan klasifikasi atau menguji model, yang mana file utama (`data.csv`) terdiri dari **13.169 baris** dan **15 kolom**, kolom yang akan diuji adalah “tweet” dan beberapa data pendukung lainnya:

 `abusive.csv` 

 `new_kamusalay.csv` 



ANALISA DATA

- Memanipulasi data dalam bentuk DataFrame menggunakan **pandas**.
- Memvisualisasi data statistik dalam bentuk grafik dengan **seaborn**.
- Memvisualisasi data statistik dalam bentuk grafik dengan **matplotlib**.
- Menghitung operasi numerik pada data dengan **numpy**.
- Menggambarkan kata-kata yang sering muncul dalam teks dengan **wordcloud**.



MODELING DATA

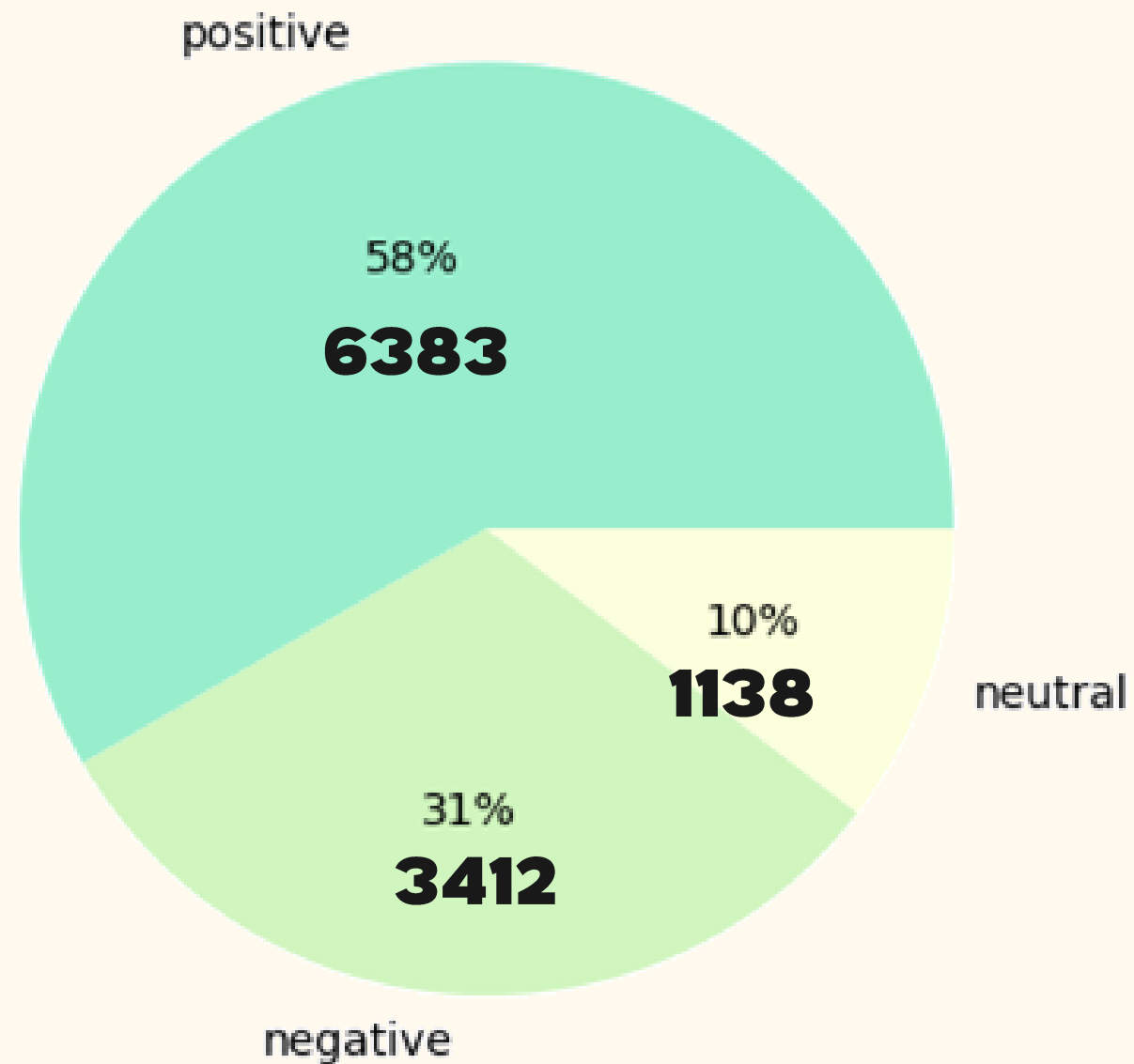
- Pembersihan/ Cleansing dengan **Regular Expression (Regex)**, seperti: link, retweet, baris baru (`/n`), double slash (`//`), double space, username, hashtag, rt, emoticon, menghilangkan semua symbol selain angka dan huruf, filter kata alay menggantinya dengan kata baku, menghapus kata-kata kasar.
- Lematisasi dan stop words dengan **Sastrawi**
- Membangun model:

	Feature Extraction	Splitting Dataset	Training	Evaluation
LSTM	tf-idf, Tensorflow (Tokenizer, pad_sequences)	sklearn (train_test_split)	TensorFlow (keras)	sklearn (classification_report)
NN	BoW, sklearn (CountVectorizer)	sklearn (train_test_split)	MLPClassifier	sklearn (classification_report)

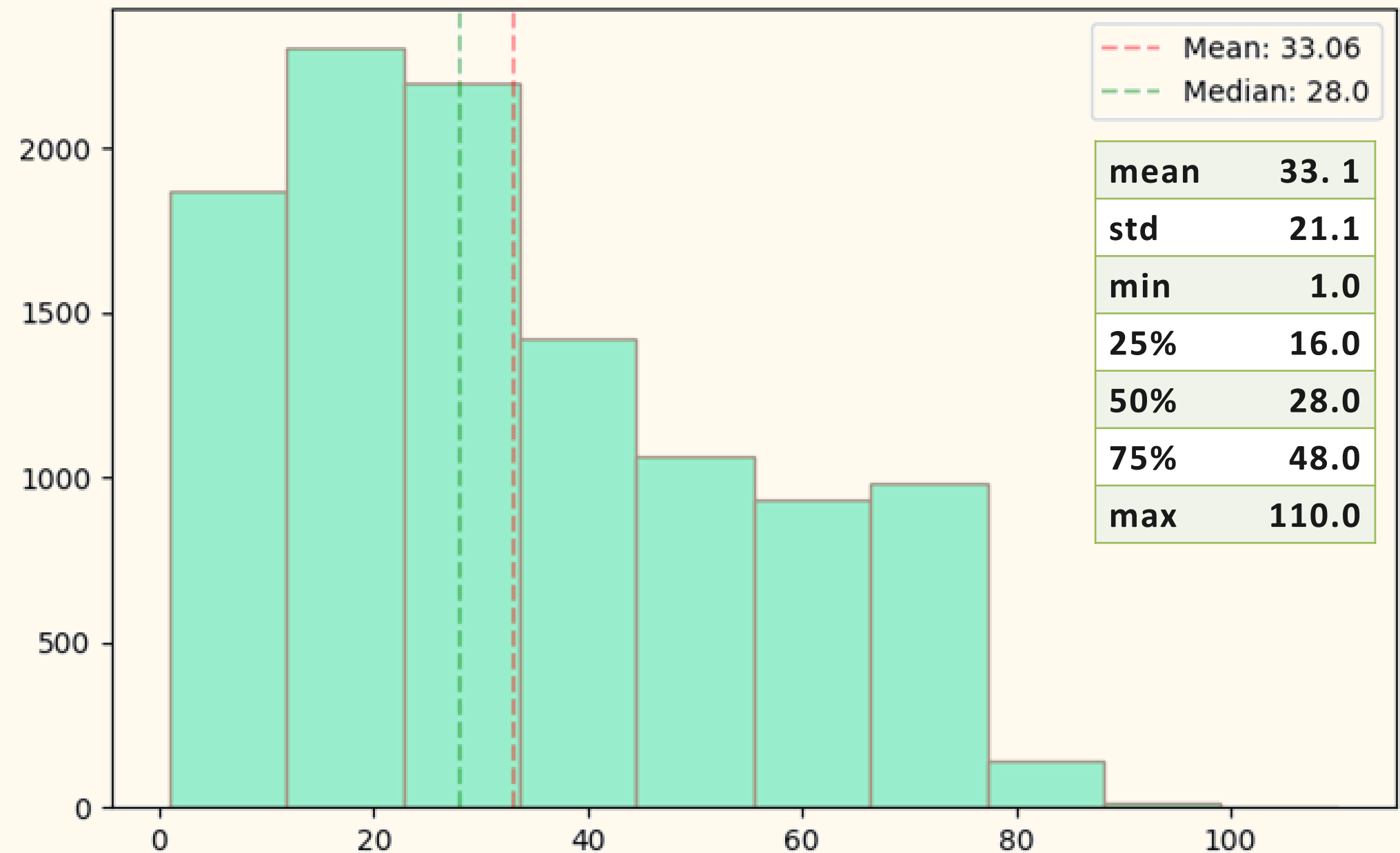
- Server API (backend dan frontend) dibuat dengan Flask dan **Swagger UI**
- Penyimpanan data menggunakan **SQLite** (SQLite3)

VISUALISASI - Exploratory Data Analysis (EDA) – DATA MODEL

DISTRIBUTION OF LABELS



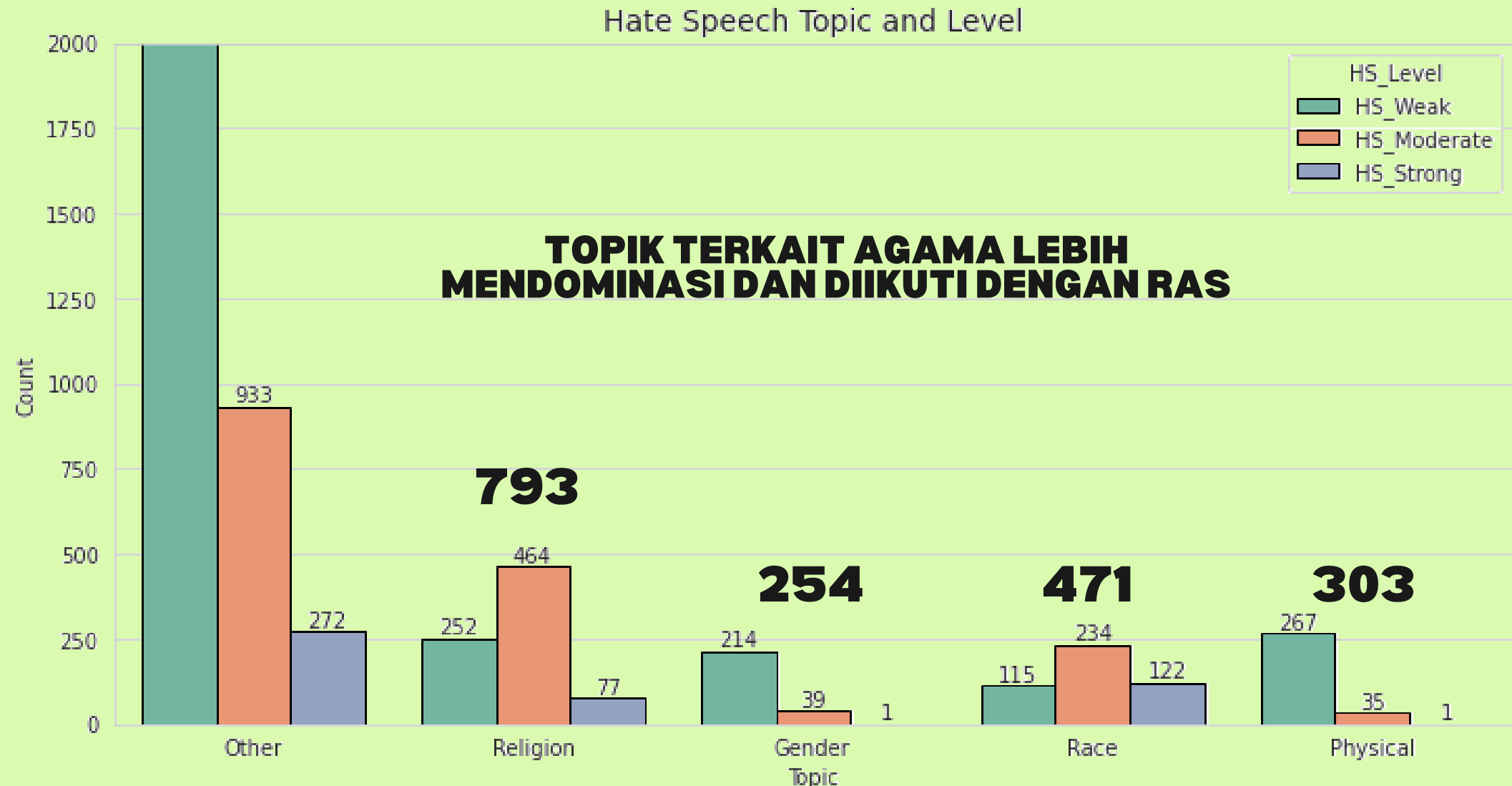
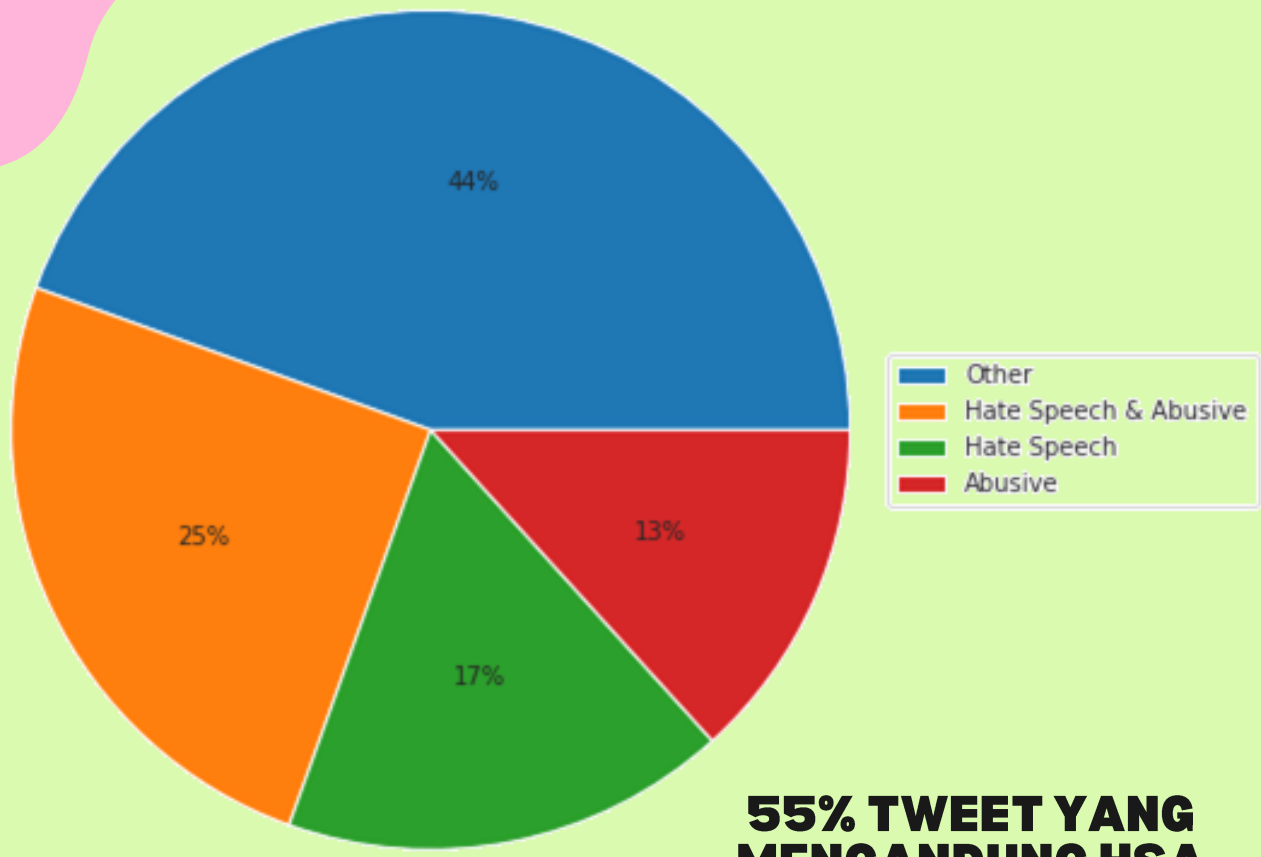
WORD TOTAL DISTRIBUTION



Ket:

Dataset semula terdiri dari 11.000 baris, namun terdapat 67 data duplikat, setelah dikurangi menjadi 10.933 baris.

VISUALISASI



Berikut ini banyak kata yang muncul dikelompokkan berdasarkan topik dari religion, gender, race, dan physical.

RELIGION



GENDER



RACE



PHYSICAL



VISUALISASI - Model Training & Evaluation (LSTM)

Hyperparameter

Input Layer	64	Embed_dim	100	Epoch	10	Dropout	0.5	Early Stopping	on	Mode	min
Output Layer	3	Learning Rate	0.0001	Batch Size	10	Activation	Softmax	Monitor	Val_loss	Verbose	1

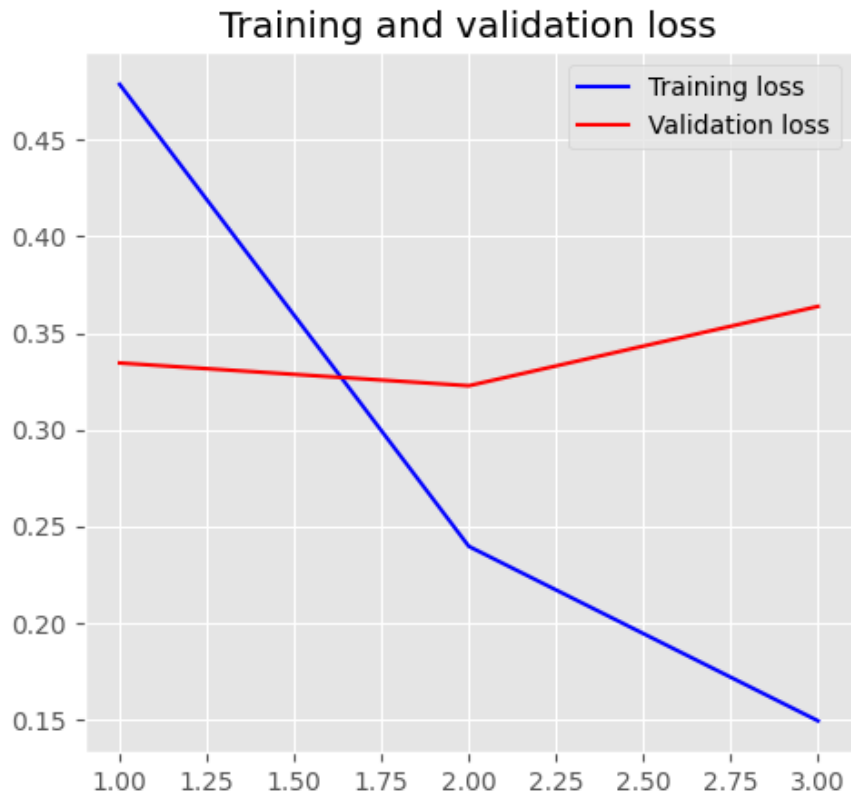
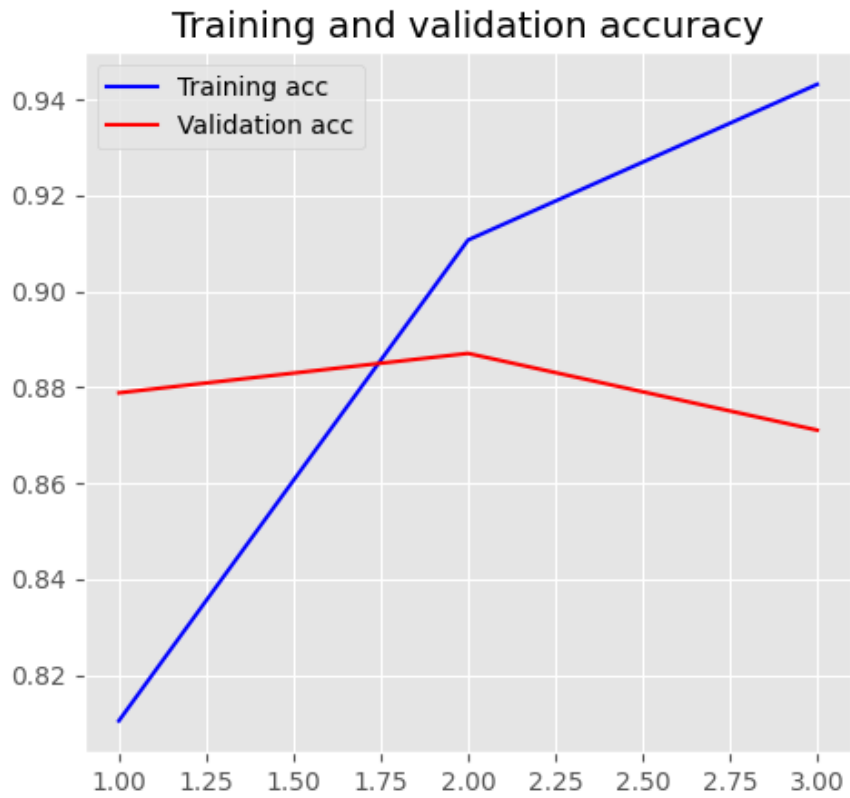
confusion matrix

69/69 [=====] - 2s 20ms/step

Testing selesai

	precision	recall	f1-score	support
0	0.87	0.76	0.81	681
1	0.76	0.76	0.76	235
2	0.88	0.94	0.91	1271
accuracy			0.87	2187
macro avg	0.84	0.82	0.83	2187
weighted avg	0.87	0.87	0.86	2187

Visualization



VISUALISASI - Model Training & Evaluation (NN)

Confusion Matrix

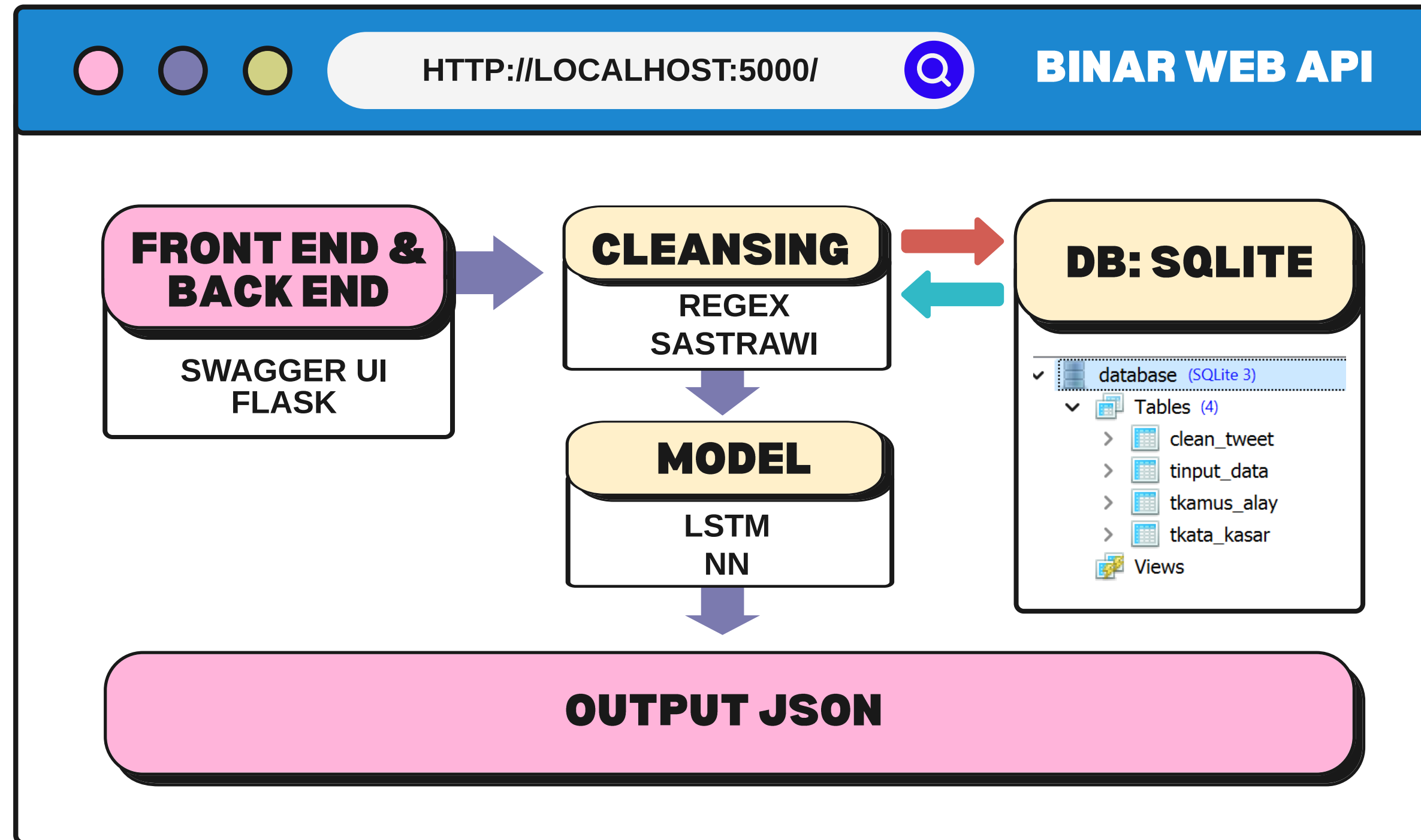
	precision	recall	f1-score	support
negative	0.75	0.81	0.78	664
neutral	0.89	0.56	0.69	232
positive	0.89	0.91	0.90	1304
accuracy			0.84	2200
macro avg	0.84	0.76	0.79	2200
weighted avg	0.85	0.84	0.84	2200

VISUALISASI - Alur Proses Kerja Web API

USER



1. INPUT TEXT
2. UPLOAD FILE CSV



VISUALISASI - Tampilan LSTM

Tampilan Frontend Web API

API Documentation BINAR PLATINUM CHALLENGE 1.0.0 // BETA

[Base URL: 127.0.0.1:5000]
[/docs.json](#)

API Documentation for Text Processing

Form Uji Model LSTM

POST

/input_dataLSTM

POST

/upload_dataLSTM

Form Uji Model NN

POST

/input_dataNN

POST

/upload_dataNN

OutPut

```
{
  "input": "aku merasa sedih sekali hari ini",
  "output": "aku rasa sedih sekali hari",
  "sentiment": "negative"
}
```

Response headers

```
connection: close
content-length: 106
content-type: application/json
date: Tue, 18 Jul 2023 12:48:41 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

OutPut

Response body

```
{
  "output": [
    {
      "new_tweet": "saat semua cowok usaha lacak perhati kamu lantas remeh perhati saya kasih khusus kamu dasar kamu cowok",
      "sentiment": "negative"
    },
    {
      "new_tweet": "siapa telat beri tau kamu gaul cigax jifla cal sama siapa licew",
      "sentiment": "negative"
    },
    {
      "new_tweet": "41 kadang aku pikir aku tetap percaya tuhan padahal aku selalu jatuh kali kali kadang aku rasa tuhan tinggal ak",
      "sentiment": "positive"
    },
    {
      "new_tweet": "aku aku ku tau mata lihat mana aku",
      "sentiment": "positive"
    },
    {
      "new_tweet": "kaum lihat dongok awal tambah haha",
      "sentiment": "negative"
    },
    {
      "new_tweet": "dan kawan kawan",
      "sentiment": "neutral"
    }
  ]
}
```

Response headers

```
connection: close
content-length: 11619
content-type: application/json
date: Sun, 16 Jul 2023 06:37:43 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

VISUALISASI - Tampilan NN

Tampilan Frontend Web API

API Documentation BINAR PLATINUM CHALLENGE 1.0.0 // BETA

[Base URL: 127.0.0.1:5000]
[/docs.json](#)

API Documentation for Text Processing

Form Uji Model LSTM

POST

/input_dataLSTM

POST

/upload_dataLSTM

Form Uji Model NN

POST

/input_dataNN

POST

/upload_dataNN

OutPut

Response body

```
{
  "input": "aku merasa sedih sekali hari ini",
  "output": "aku rasa sedih sekali hari",
  "sentiment": "positive"
}
```

Response headers

```
connection: close
content-length: 106
content-type: application/json
date: Tue, 18 Jul 2023 12:53:25 GMT
```

OutPut

Response body

```
{
  "sentiment_results": [
    {
      "cleaned_tweet": "saat semua cowok usaha lacak perhati kamu lantas remeh perhati saya kasih khusus kamu dasar kamu cowok",
      "sentiment": "neutral"
    },
    {
      "cleaned_tweet": "siapa telat beri tau kamu gaul cigax jifla cal sama siapa licew",
      "sentiment": "negative"
    },
    {
      "cleaned_tweet": "41 kadang aku pikir aku tetap percaya tuhan padahal aku selalu jatuh kali kali kadang aku rasa tuhan pilih jadi kristen aku anak ter",
      "sentiment": "positive"
    },
    {
      "cleaned_tweet": "aku aku ku tau mata lihat mana aku",
      "sentiment": "positive"
    },
    {
      "cleaned_tweet": "kaum lihat dongok awal tambah haha",
      "sentiment": "negative"
    },
    {
      "cleaned_tweet": "dan kawan kawan",
      "sentiment": "negative"
    }
  ]
}
```

Response headers



KESIMPULAN

- Sentimen analisis memberikan wawasan yang berharga bagi pengguna. Khusus untuk perusahaan dan organisasi, dapat menjadi acuan dalam mengambil keputusan yang lebih baik, merespons dengan cepat terhadap perubahan pasar atau opini masyarakat, serta meningkatkan interaksi dan hubungan dengan pelanggan.
- Model LSTM memiliki performa yang lebih baik dibandingkan dengan model NN, meskipun model cenderung grafik overfitting, tetapi hasil sentiment cukup baik.
- API yang dibuat setiap model memiliki 2 endpoint (untuk memproses teks dan data file csv) yang dapat menampilkan label positif, negatif, atau netral berdasarkan hasil dari proses modelnya.



SARAN

- Diperlukan percobaan lebih lanjut sehingga dapat menghasilkan nilai akurasi model yang baik dengan hasil grafik evaluasi yang mendekati goodfit.
- Dapat dikembangkan menjadi program otomatis yang dintergrasikan dengan twitter API, sehingga pengguna mendapatkan warning terhadap penyebaran tweet yang bernada negative.

