

## [ GOLD CHALLENGE ]

# ANALISIS DATA DAN PEMBUATAN API UNTUK *DATA CLEANSING*

### OBJEK PENELITIAN:

Data Twitter dengan Judul Multi-label Hate Speech and Abusive Language Detection in {I}ndonesian Twitter

Oleh :  
Nadiah Zulfa (Binarian Wave-8)

# PENDAHULUAN

## LATAR BELAKANG

**Hate speech** atau ujaran kebencian adalah suatu bentuk ekspresi yang dilakukan untuk menyebarkan rasa kebencian dan melakukan kekerasan serta diskriminasi terhadap seseorang atau sekelompok orang dengan berbagai alasan (Davidson, Warmley, Macy, & Weber, 2017). Kasus *hate speech* sangat sering kita jumpai di media sosial, salah satunya di **Twitter**.

Twitter memberikan sebuah kebebasan kepada penggunanya untuk mengekspresikan diri mereka melalui *tweet* (Kicauan) dengan batas maksimal 280 karakter. Adanya Batasan karakter membuat sebuah *tweet* mengalami penyingkatan kata, penggunaan bahasa yang tidak sesuai, ataupun terjadi kesalahan eja. Karena itu diperlukan proses untuk menormalisasi hal tersebut.

Objek penelitian ini ialah data saduran dari [Kaggle](#) dengan judul "Multi-label Hate Speech and Abusive Language Detection in {I}ndonesian Twitter" yang ditulis oleh Muhamaad Okky Ibrahim dan Indra Budi. Data ini mengklasifikasikan *Hate Speech* dengan berbagai label.

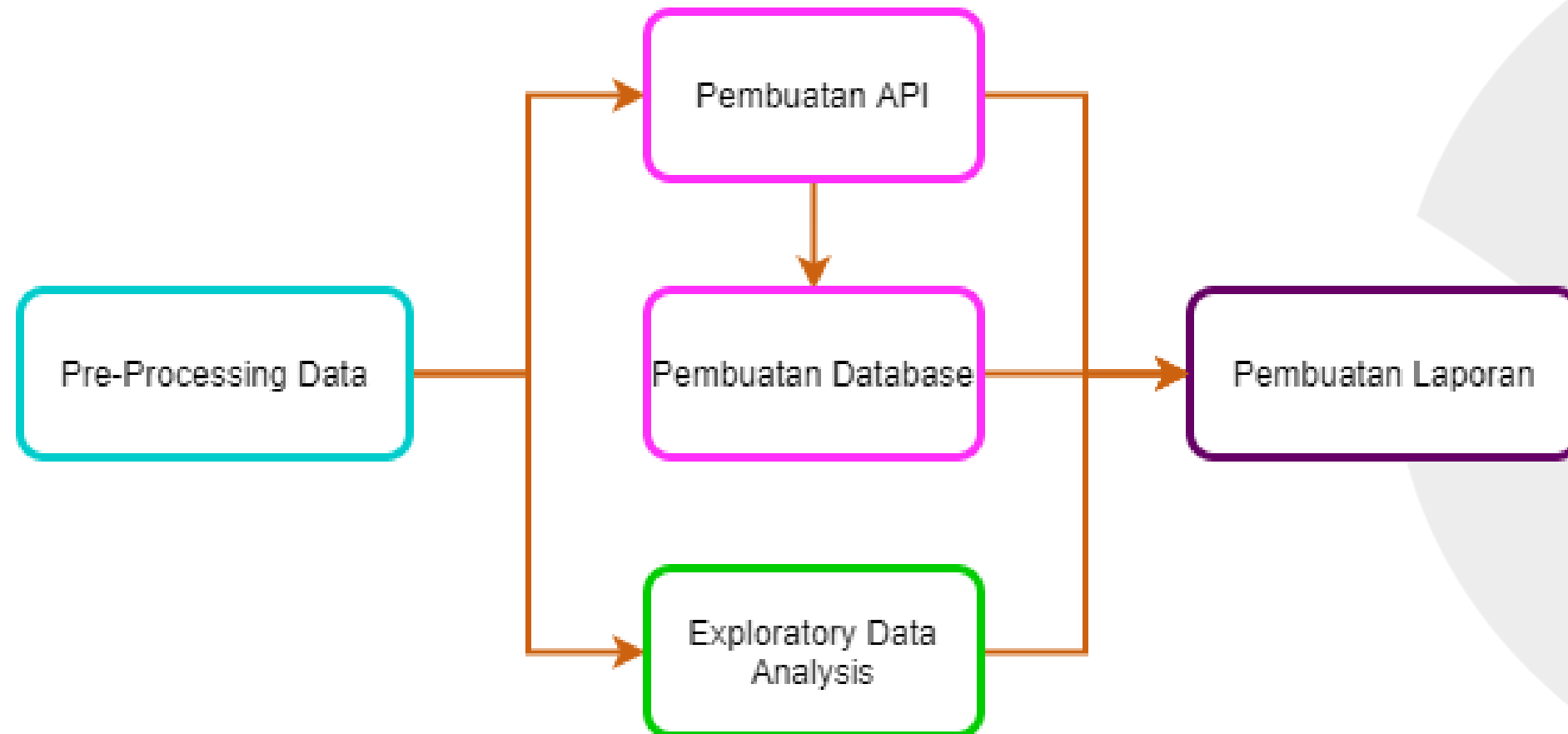
## RUMUSAN MASALAH

1. Bagaimana proses yang dilakukan untuk melakukan *data cleansing* ?
2. Bagaimana Hasil Analisis Deskriptif terhadap objek penelitian?
3. Bagaimana membuat API untuk memproses text/file cleansing dan menghasilkan output berupa teks/file yang sudah di-*cleansing*?

## TUJUAN

1. Melakukan berbagai proses yang diperlukan untuk melakukan *Data Cleansing*
2. Menjabarkan Hasil Analisis Deskriptif terhadap objek penelitian.
3. Membuat API untuk memproses text cleansing dan menghasilkan output berupa teks yang sudah di-*cleansing*.

# METODOLOGI PENELITIAN



# METODOLOGI PENELITIAN [2]

Menjelaskan apa saja yang dilakukan disetiap tahapan/proses dalam penelitian

## PRE-PROCESSING DATA

1. Drop Duplikat :  
menghilangkan data duplikat
2. Cleansing Data:
  - a) Lower-casing : Menjadikan semua huruf menjadi huruf kecil
  - b) Menghilangkan karakter selain alfa-numerik.
  - c) Menghilangkan URL.
  - d) Menghilangkan kata Retweet.
  - e) Menghilangkan kata RT
  - f) Menghilangkan spasi
  - g) Menormalisasi kata yang tidak baku

## PEMBUATAN API & DATABASE

1. Membuat API untuk cleansing data dengan masukan berupa teks
2. Membuat API untuk Cleansing data dengan masukan berupa file (csv)
3. Membuat database untuk menampung masukan sebelum dan sesudah dilakukan *cleansing*.

## EXPLORATORY DATA ANALYSIS

1. Menghitung Jumlah Karakter dan Jumlah Kata dari setiap Tweet
2. Menampilkan deskripsi data statistic (sebelum pre-processing Langkah ke-2)
3. Melihat komposisi tweet yang mengandung hate-speech dan yang tidak.
4. Melihat kata yang sering Muncul

## HASIL PENELITIAN



Hasil Pre-Processing Data



API dan Database



Exploratory Data Analysis



## HASIL PENELITIAN

### Hasil Pre-Processing Data

#### 1. Drop Duplikat

Menghilangkan data yang sama

Jumlah baris sebelum Drop Duplikat	: 13169
<b>Jumlah baris setelah Drop Duplikat</b>	<b>: 13044</b>
Jumlah Data Duplikat	: 125

#### 2. Cleansing Data

Menghilangkan karakter yang tidak penting dan menormalisasi kata tidak baku



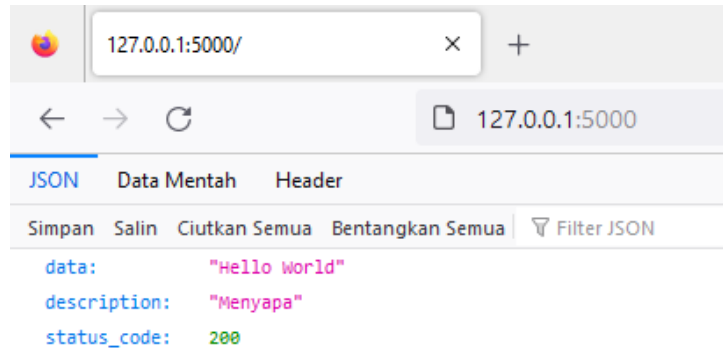
Hasil dari cleansing bisa dilihat langsung di file  
Preproses\_dan\_EDA\_GOLD\_Challenge\_NZ.ipynb



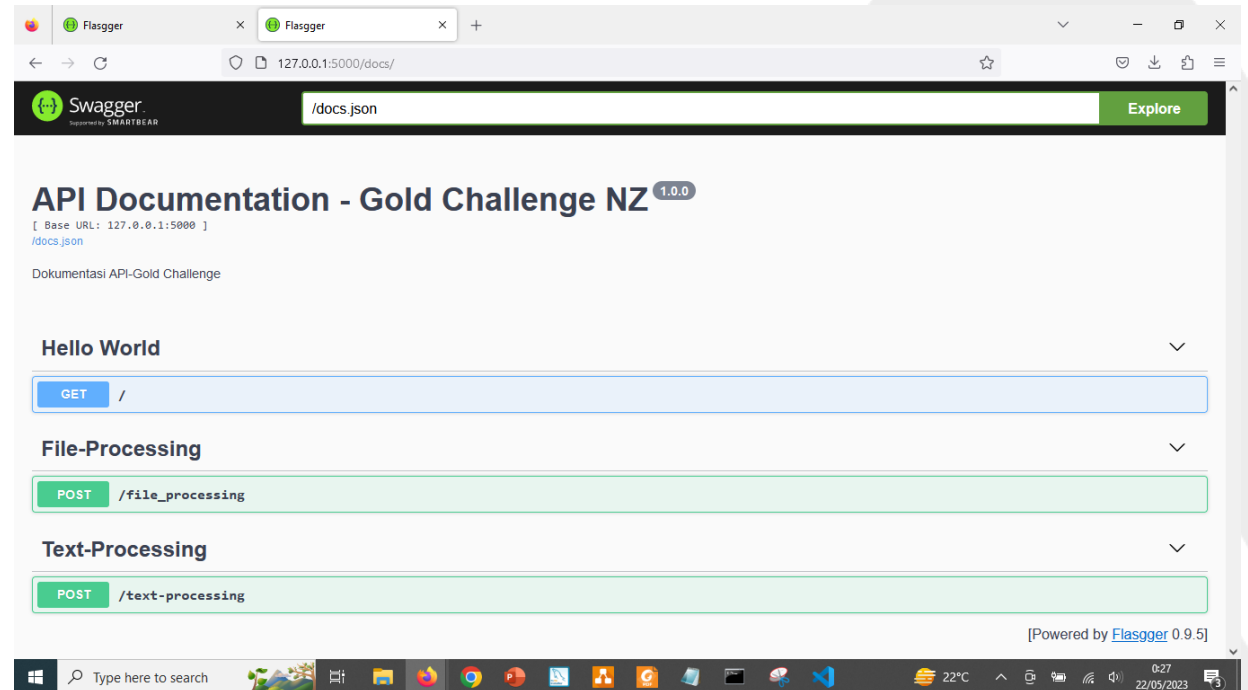
## HASIL PENELITIAN

### API dan Database (1)

1. Klik link local host <http://127.0.0.1:5000>



2. Klik link local host <http://127.0.0.1:5000/docs/>  
*Mohon tambahkan /docs untuk memunculkan swagger UI*





## HASIL PENELITIAN

### API dan Database (2)

## 3. Text-Processing

### 3.1 Percobaan 1

Masukan Teks : "di ujung jalan ada bencong gila tuh"

Keluaran Teks : "di ujung jalan ada \*\*\*disensor\*\*\* \*\*\*disensor\*\*\* itu "

### 3.2 Percobaan 2

Masukan Teks : "sebel banget masa adek gw dikatain bego "

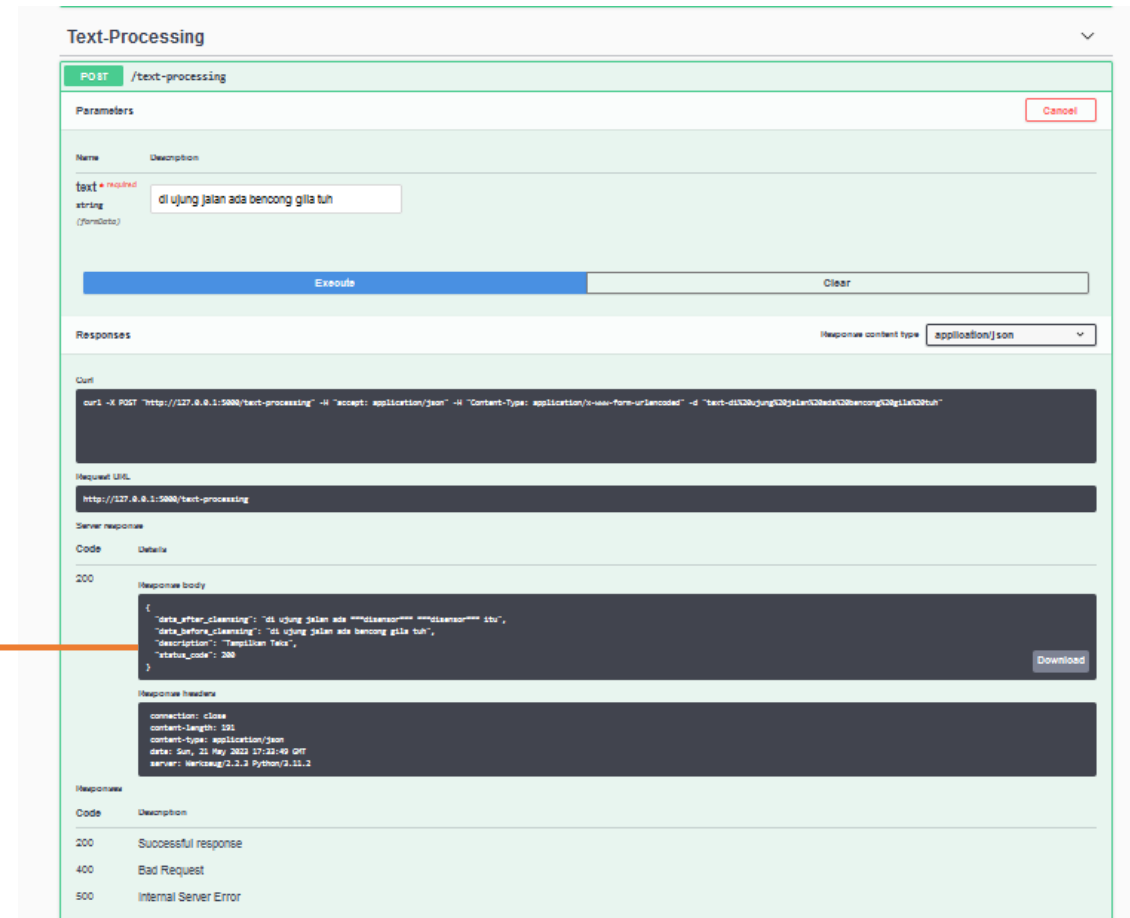
Keluaran Teks : "kesal banget masa adik gue diejek \*\*\*disensor\*\*\*"

#### Response body

```
{
  "data_after_cleansing": "kesal banget masa adik gue diejek ***disensor***",
  "data_before_cleansing": "sebel banget masa adek gw dikatain bego",
  "description": "Tampilkan Teks",
  "status_code": 200
}
```

#### Response body

```
{
  "data_after_cleansing": "di ujung jalan ada ***disensor*** ***disensor*** itu",
  "data_before_cleansing": "di ujung jalan ada bencong gila tuh",
  "description": "Tampilkan Teks",
  "status_code": 200
}
```



Text-Processing

POST /text-processing

Parameters

Name	Description
text	required string (formData)

di ujung jalan ada bencong gila tuh

Execute Clear

Responses

Response content type: application/json

Curl

```
curl -X POST 'http://127.0.0.1:5000/text-processing' -H 'accept: application/json' -H 'Content-Type: application/x-www-form-urlencoded' -d 'text=di%20ujung%20jalan%20ada%20bencong%20gila%20tuh'
```

Request URL

http://127.0.0.1:5000/text-processing

Server response

Code	Details								
200	<p>Response body</p> <pre>{   "data_after_cleansing": "di ujung jalan ada ***disensor*** ***disensor*** itu",   "data_before_cleansing": "di ujung jalan ada bencong gila tuh",   "description": "Tampilkan Teks",   "status_code": 200 }</pre> <p>Download</p> <p>Response headers</p> <pre>connection: close content-length: 181 content-type: application/json date: Sun, 20 May 2023 17:21:49 GMT server: Werkzeug/2.2.3 Python/3.11.2</pre> <p>Responses</p> <table border="1"><thead><tr><th>Code</th><th>Description</th></tr></thead><tbody><tr><td>200</td><td>Successful response</td></tr><tr><td>400</td><td>Bad Request</td></tr><tr><td>500</td><td>Internal Server Error</td></tr></tbody></table>	Code	Description	200	Successful response	400	Bad Request	500	Internal Server Error
Code	Description								
200	Successful response								
400	Bad Request								
500	Internal Server Error								





## HASIL PENELITIAN

### API dan Database (3)

## 4. File-Processing

### 4.1 Percobaan 1

Masukan File : data\_teroris.csv

Keluaran : Error

#### Error: INTERNAL SERVER ERROR

##### Response body

```
<!doctype html>
<html lang=en>
<title>500 Internal Server Error</title>
<h1>Internal Server Error</h1>
<p>The server encountered an internal error and was unable to complete your request. Either the server is overloaded or there is an error in the application.</p>
```

##### Response headers

```
connection: close
content-length: 265
content-type: text/html; charset=utf-8
date: Sun, 21 May 2023 16:11:00 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

File-Processing

POST /file\_processing

Parameters

Name Description

File \*required File to upload

file (FormData) Telusuri... Indonesian twee... out\_teroris.csv

Execute Clear

Responses

Response content type application/json

Curl

```
curl -X POST "http://127.0.0.1:5000/file_processing" -H "accept: application/json" -H "Content-Type: multipart/form-data" -F "file=@Indonesian_tweet_about_terroris.csv;type=application/vnd.ms-excel"
```

Request URL

http://127.0.0.1:5000/file\_processing

Server response

Code	Details
500	Error: INTERNAL SERVER ERROR

Response body

```
<!doctype html>
<html lang=en>
<title>500 Internal Server Error</title>
<h1>Internal Server Error</h1>
<p>The server encountered an internal error and was unable to complete your request. Either the server is overloaded or there is an error in the application.</p>
```

Response headers

```
connection: close
content-length: 265
content-type: text/html; charset=utf-8
date: Sun, 21 May 2023 16:11:00 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

Response

Code	Description
200	Successful response
400	Bad Request
500	Internal Server Error



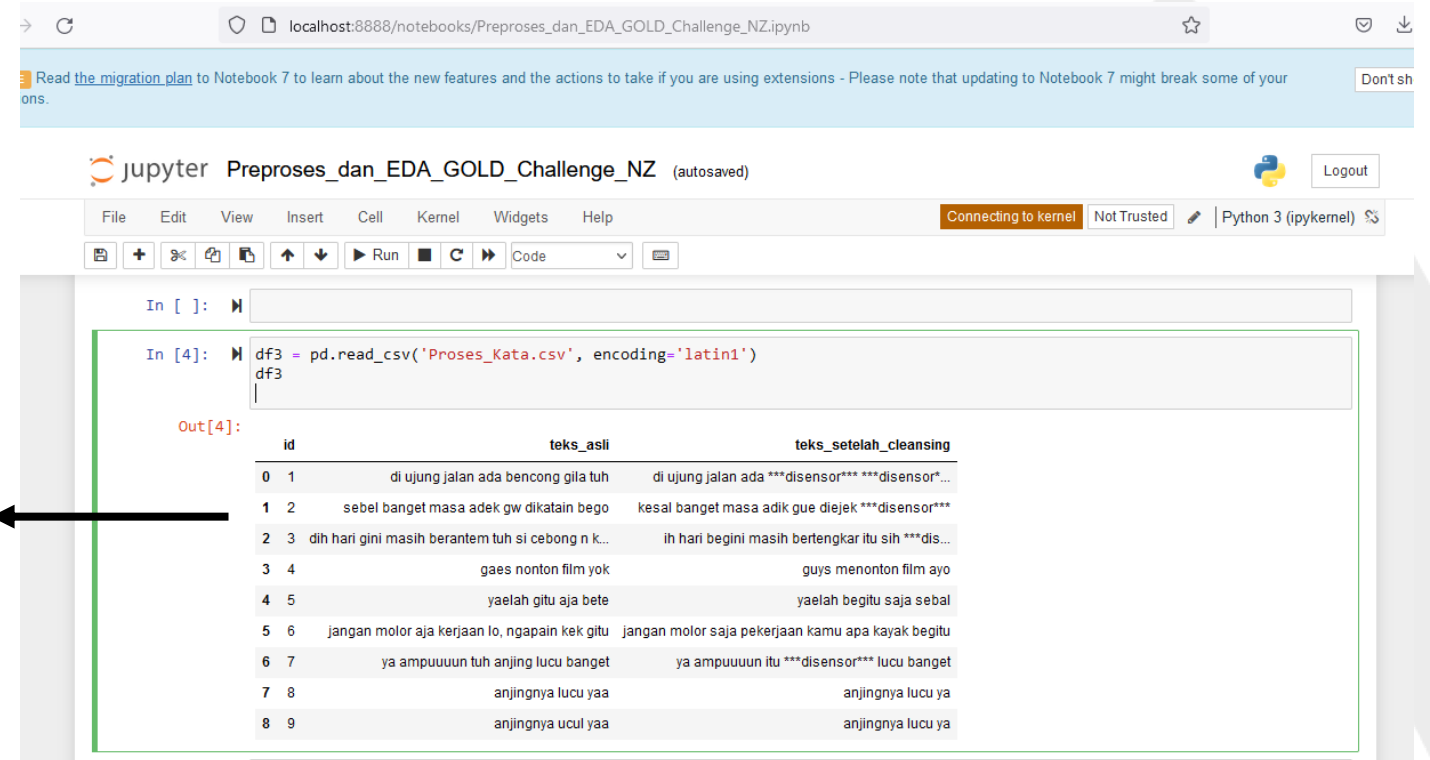
## HASIL PENELITIAN

### API dan Database (4)

## 5. Database

1. Coba berbagai teks di API
2. Convert file db to csv
3. Read file csv tersebut menggunakan pandas (di file Preproses\_dan\_EDA\_GOLD\_Challenge\_NZ.ipynb)

id	teks_asli	teks_setelah_cleansing
0 1	di ujung jalan ada bencong gila tuh	di ujung jalan ada ***disensor*** **disensor*...
1 2	sebel banget masa adek gw dikatain bego	kesal banget masa adik gue diejek ***disensor***
2 3	dih hari gini masih berantem tuh si cebong n k...	ih hari begini masih bertengkar itu sih ***dis...
3 4	gaes nonton film yok	guys menonton film ayo
4 5	yaelah gitu aja bete	yaelah begitu saja sebal
5 6	jangan molor aja kerjaan lo, ngapain kek gitu	jangan molor saja pekerjaan kamu apa kayak begitu
6 7	ya ampuuuun tuh anjing lucu banget	ya ampuuuun itu ***disensor*** lucu banget
7 8	anjingnya lucu yaa	anjingnya lucu ya
8 9	anjingnya ucul yaa	anjingnya lucu ya



The screenshot shows a Jupyter Notebook titled "Preproses\_dan\_EDA\_GOLD\_Challenge\_NZ" with the following code and output:

```
In [4]: df3 = pd.read_csv('Proses_Kata.csv', encoding='latin1')
df3
```

Out[4]:

id	teks_asli	teks_setelah_cleansing
0 1	di ujung jalan ada bencong gila tuh	di ujung jalan ada ***disensor*** **disensor*...
1 2	sebel banget masa adek gw dikatain bego	kesal banget masa adik gue diejek ***disensor***
2 3	dih hari gini masih berantem tuh si cebong n k...	ih hari begini masih bertengkar itu sih ***dis...
3 4	gaes nonton film yok	guys menonton film ayo
4 5	yaelah gitu aja bete	yaelah begitu saja sebal
5 6	jangan molor aja kerjaan lo, ngapain kek gitu	jangan molor saja pekerjaan kamu apa kayak begitu
6 7	ya ampuuuun tuh anjing lucu banget	ya ampuuuun itu ***disensor*** lucu banget
7 8	anjingnya lucu yaa	anjingnya lucu ya
8 9	anjingnya ucul yaa	anjingnya lucu ya

An arrow points from the output table in the Jupyter Notebook to the table on the left.



## HASIL PENELITIAN

3

### Exploratory Data Analysis



**BINAR**  
WAVE - 8

#### 1. Hitung Panjang Karakter dan Jumlah Kata\*

Menghitung Panjang Karakter dan Jumlah Kata serta menambahkan kolom untuk itu.

_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong	panjang_karakter	jumlah_kata
1	0	0	0	0	0	1	1	0	0	138	25
0	0	0	0	0	0	0	0	0	0	120	21
0	0	0	0	0	0	0	0	0	0	254	37

#### 2. Deskripsi Statistik\*

	Panjang_karakter	Jumlah_kata
Mean	114,20	17,28
Median	100	15
Range	557	51
Q1	59	9
Q2	100	15
Q3	152	23
Nilai Min	4	1
Nilai Max	561	52

\*Diolah sebelum dilakukan penghapusan karakter/kata

Kata yang paling muncul adalah kata “dan”, hal ini terjadi karena “dan” merupakan kata hubung.

## KESIMPULAN & SARAN

### KESIMPULAN

1. Proses yang dilakukan untuk melakukan *data cleansing* adalah:
  - ❖ Menghilangkan data duplikat
  - ❖ Menghilangkan karakter yang tidak perlu
  - ❖ Menormalisasi kata tidak baku
2. Analisi deskriptif menunjukkan bahwa:
  - ❖ Tweet yang tidak mengandung ujaran kebencian lebih banyak daripada tweet yang mengandung ujaran kebencian.
  - ❖ Kata yang paling sering muncul adalah “**dan**”.
3. API untuk cleansing data dengan masukan berupa teks berjalan dengan baik. Begitupun database berhasil dibuat. Namun, untuk masukan berupa file (csv) masih mengalami *error*.

### SARAN

1. Sepertinya perlu dilakukan proses penghilangan kata hubung pada tahap pre-processing, karena terlihat bahwa kata yang paling sering muncul adalah kata “dan”, “yang”, dan sejenisnya yang merupakan kata hubung.
2. Analisi deskriptif bisa digali dan diolah lebih eksploratif.
3. Fungsi untuk unggah dan membaca masukan berupa file perlu diulik Kembali supaya API dengan masukan berupa file bisa berjalan dengan baik.