BINAR ACADEMY

DataScienceProjects

Cleaning Comment Data FROM X Platform In Indonesia with Regular Expressions

Rema Bagos Pudyastowo
Binar Academy

The purpose of this research is to eliminate abusive words in comments on the X platform.

namesurname @namesurname • 13 Dec
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

namesurname @namesurname • 25 Nov
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim.

Linkedin : Rema Bagos Pudyastowo

Email : remabagospudyastowo@gmail.com

INDONESIA

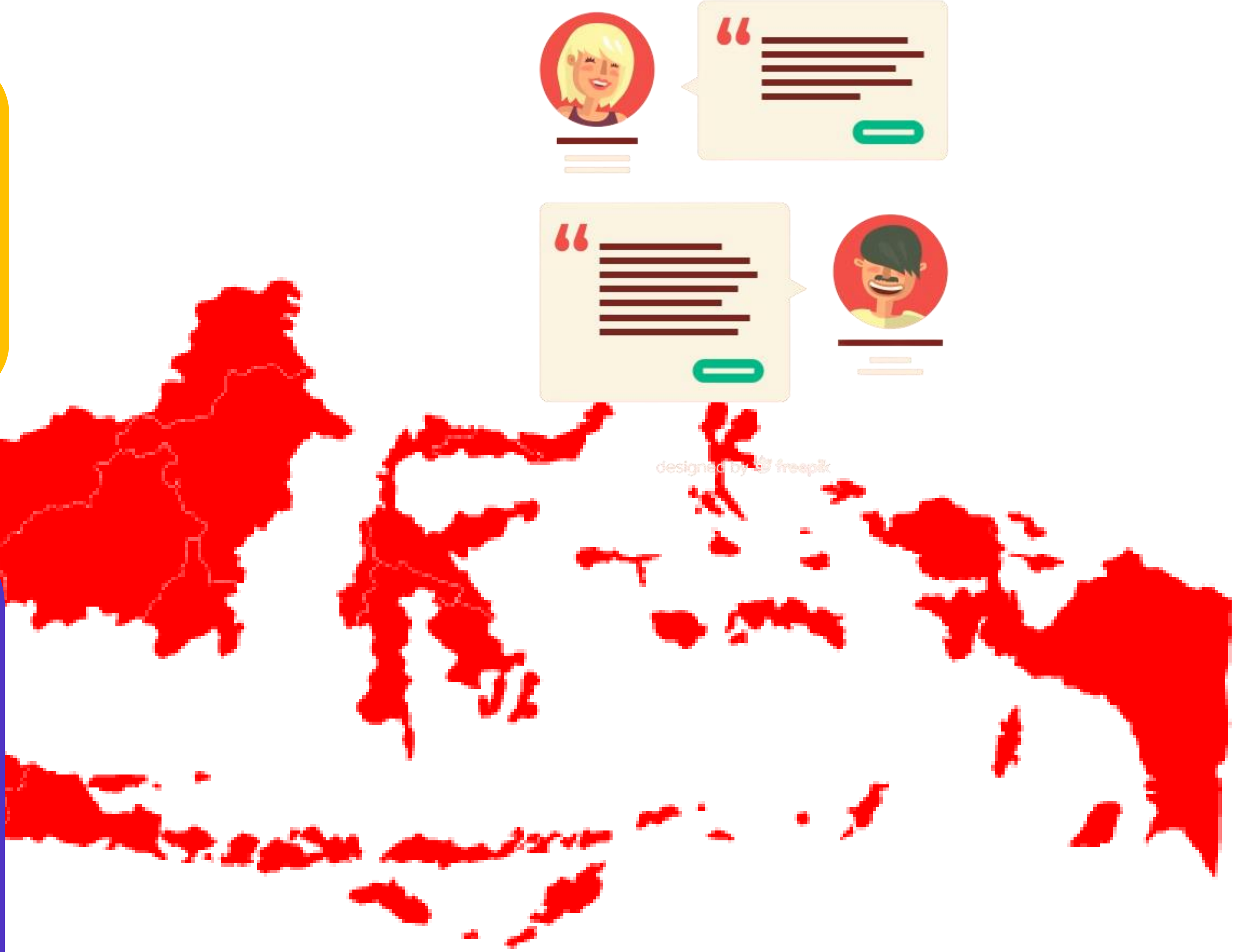
1

Indonesia is the country with the 5th most users of platform X (formerly: Twitter) in the world after the United Kingdom.

Source : https://www.kominfo.go.id/

2

Platform X users in Indonesia are users with high activity, especially in terms of commenting can trigger the emergence of hateful words, harsh words and other unwanted things.



Hate Speech In SOCIAL MEDIA

Hate speech is a direct or indirect speech toward a person or group containing hatred based on something (Komnas HAM, 2015)

Factors that are often used as bases of hatred include ethnicity, religion, disability, gender, and sexual orientation.

Hate speech spreading is a very dangerous action which can have some negative effects such as discrimination, social conflict, and even human genocide (Komnas HAM, 2015)

Hate Speech

- In everyday life, especially in social media, the hate speech spreading is often accompanied with abusive language (Davidson et al., 2017).

- Hate speech that contains abusive words/phrases often accelerates the occurrence of social conflict because of the use of the abusive words/phrases that triggers emotions.

- The use of abusive language in social media still can lead to conflict because of misunderstandings among netizens (Yenala et al., 2017).

- Moreover, children could be exposed to language inappropriate
- for their age from those abusive language scattered in their social media (Chen et al., 2012).

What can we do?

Remove Abusive Text!

The hate speech and abusive language on social media must be detected dan removed to avoid conflict between citizens and children learning the hate speech and inappropriate language from the social media they use.

PURPOSE FOR THIS RESEARCH

- Identifying abusive words in twitter comment data

- Removing abusive words from twitter comment data

About the data

Data.csv

Data containing Twitter comments containing abusive words.

New_kamusalay.csv

Contains a dictionary of 'alay' words and their root replacements.

Abusive.csv

Contains harsh words that match the comment data 'data.csv'

Source Data : kaggle

METHODS FOR THIS RESEARCH

Hasskommentare im Netz. Steuerungsstrategien für Redaktionen
#
!
Counter Speech (Gegenrede)
Argumentatives Sprechen

Descriptive Analytics

Analyze how much percentage of abusive words are in each comment by analyzing the relationship of each data label.

Here’s The Steps

1 Problem Understanding

2 Data Preparation

3 Data Analysis

4 Conclusion

5 Making API

6 Deploy the text cleansing project

TOOLS

python

Flask
web development,
one drop at a time

Swagger™

Additional Library :

pandas

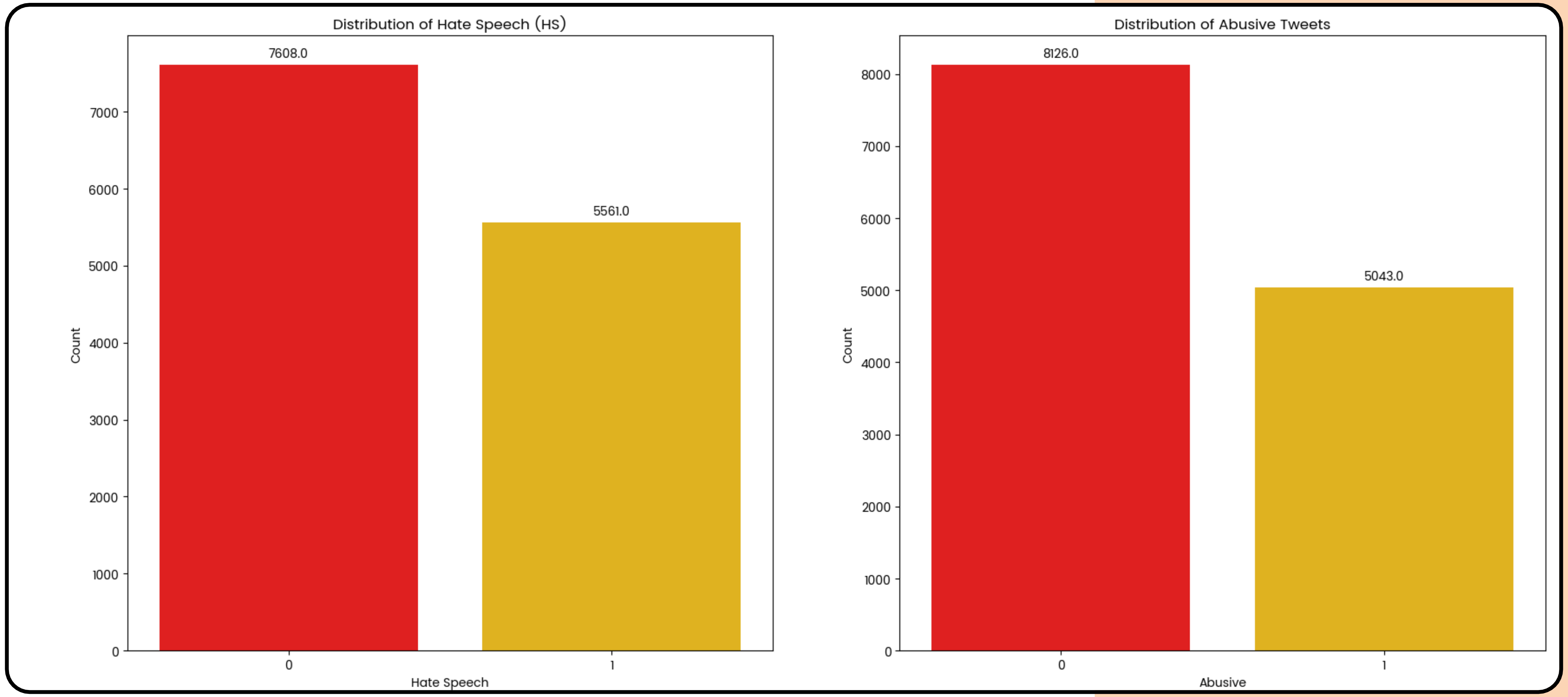
NumPy

RegEx

Data Visualization

About the Tweet Data (data.csv)

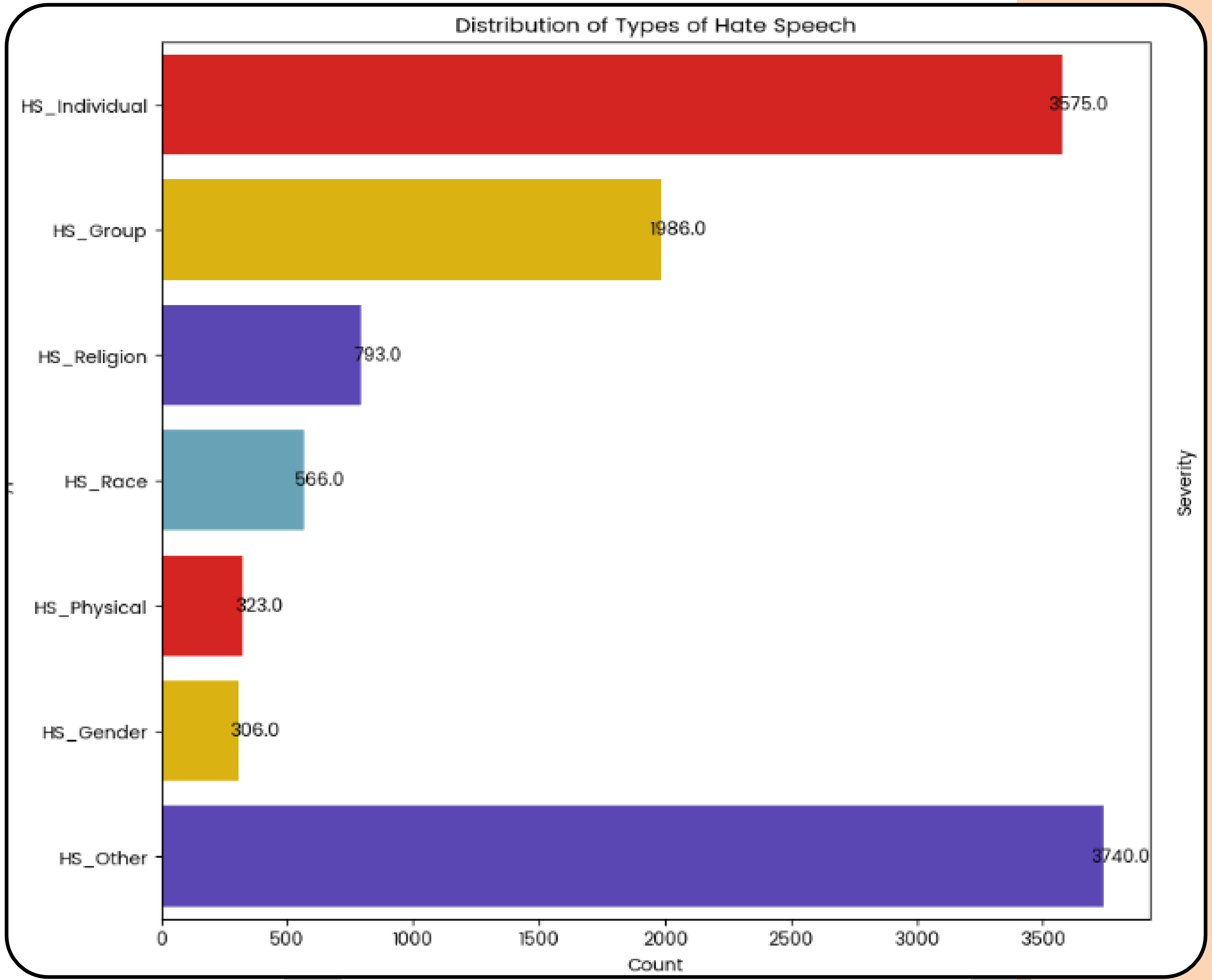
More



- **About 7608 comments did not fall under hateful comments and 5561 fell under hateful comments.**
- **42% of user comments fall into the category of hatred**
- **While overall comments containing abusive words are 5043 or about 38% of all comments in the twitter comment data (data.csv).**

Data Visualization

About the Tweet Data (data.csv)



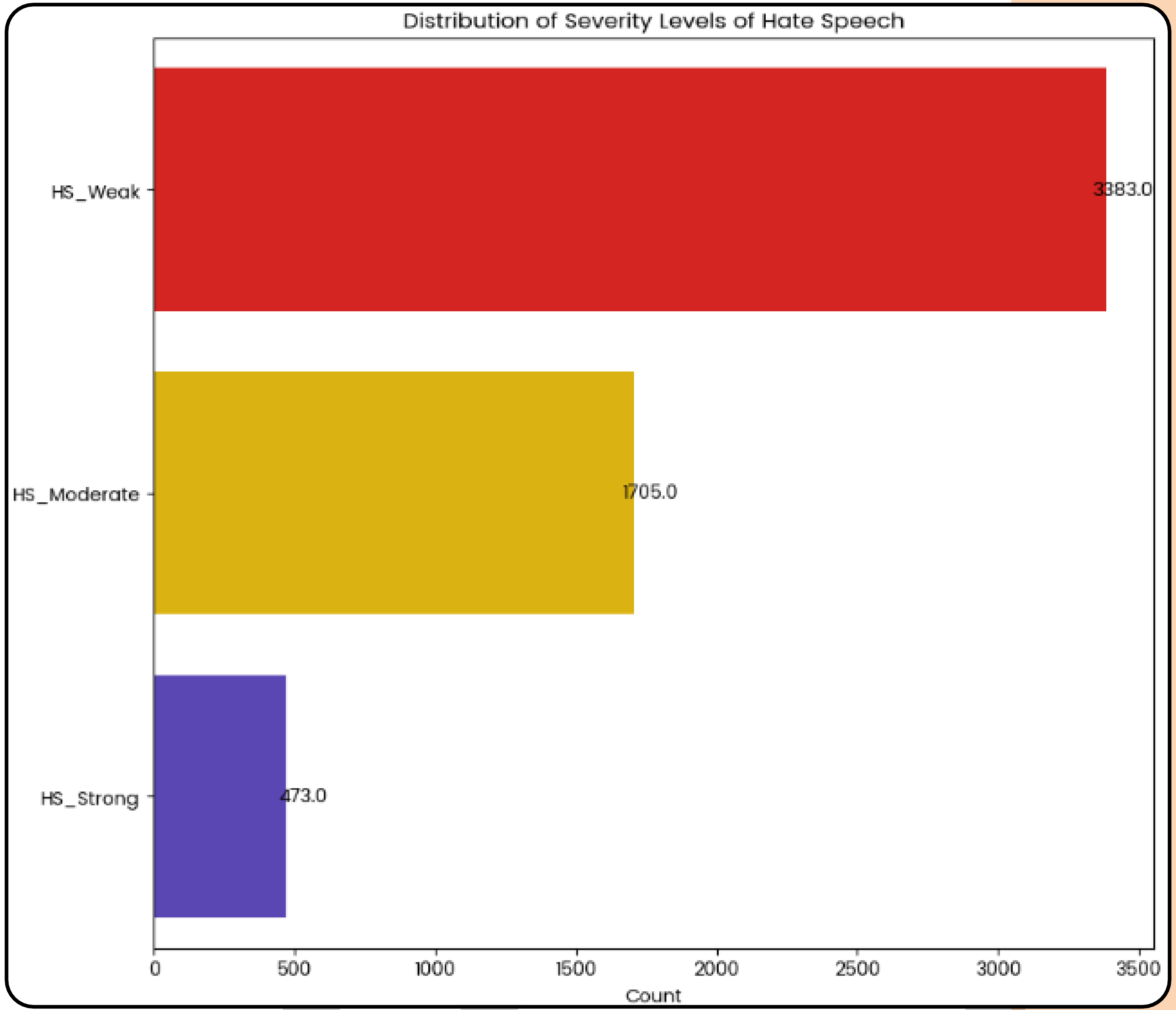
More

- Hate speech according to its object is divided into several types, namely HS_individual which leads to an individual, HS_group which leads to a certain group of people, HS_religion which leads to a certain religion, HS_race which leads to a certain tribe or race, HS_Physical which leads to a person's physical condition, HS_Gender which leads to male or female gender, and HS_other is hate speech directed to other things.

- Hate speech that leads to individuals is a type of hate speech that quite dominates the user's tweet comment data in Indonesia and hate speech that leads to gender has the smallest value. this shows that users in Indonesia more often do hate speech that leads to individuals.

Data Visualization

About the Tweet Data (data.csv)

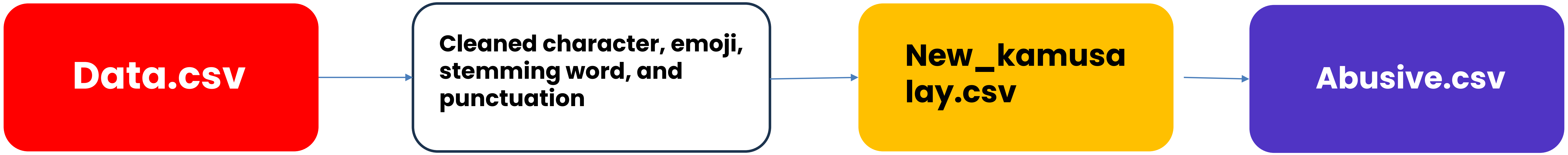


More

- Hate speech types from twitter comment data are broadly divided into 3 parts, namely mild hate speech, medium or moderate class, and heavy or strong class.

- Twitter comments are dominated by mild hate speech, which is around 60%, moderate class is 30% and heavy or strong class is 8%.

Data Cleaning

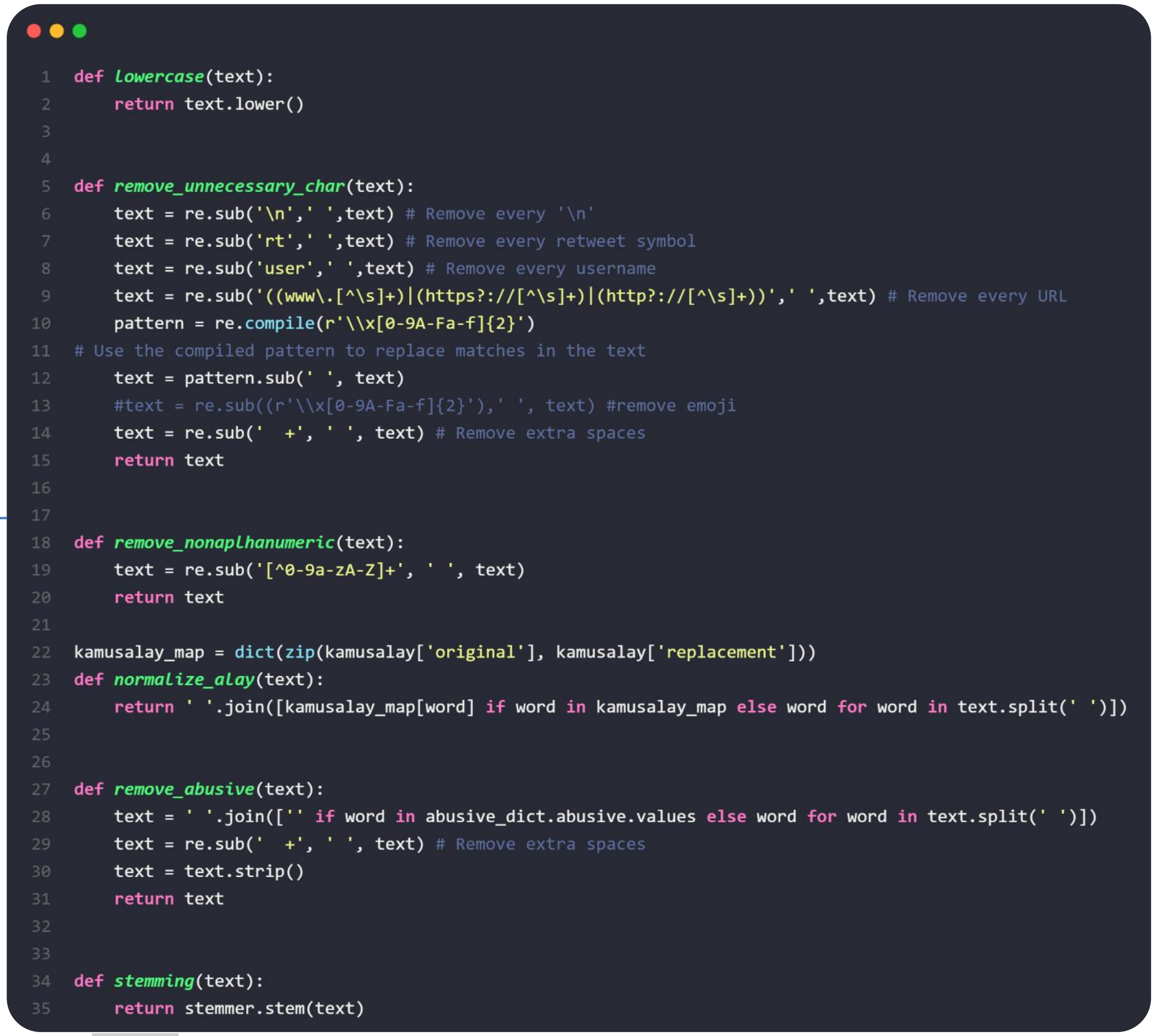


Data.csv

Cleaned character, emoji, stemming word, and punctuation

New_kamusalay.csv

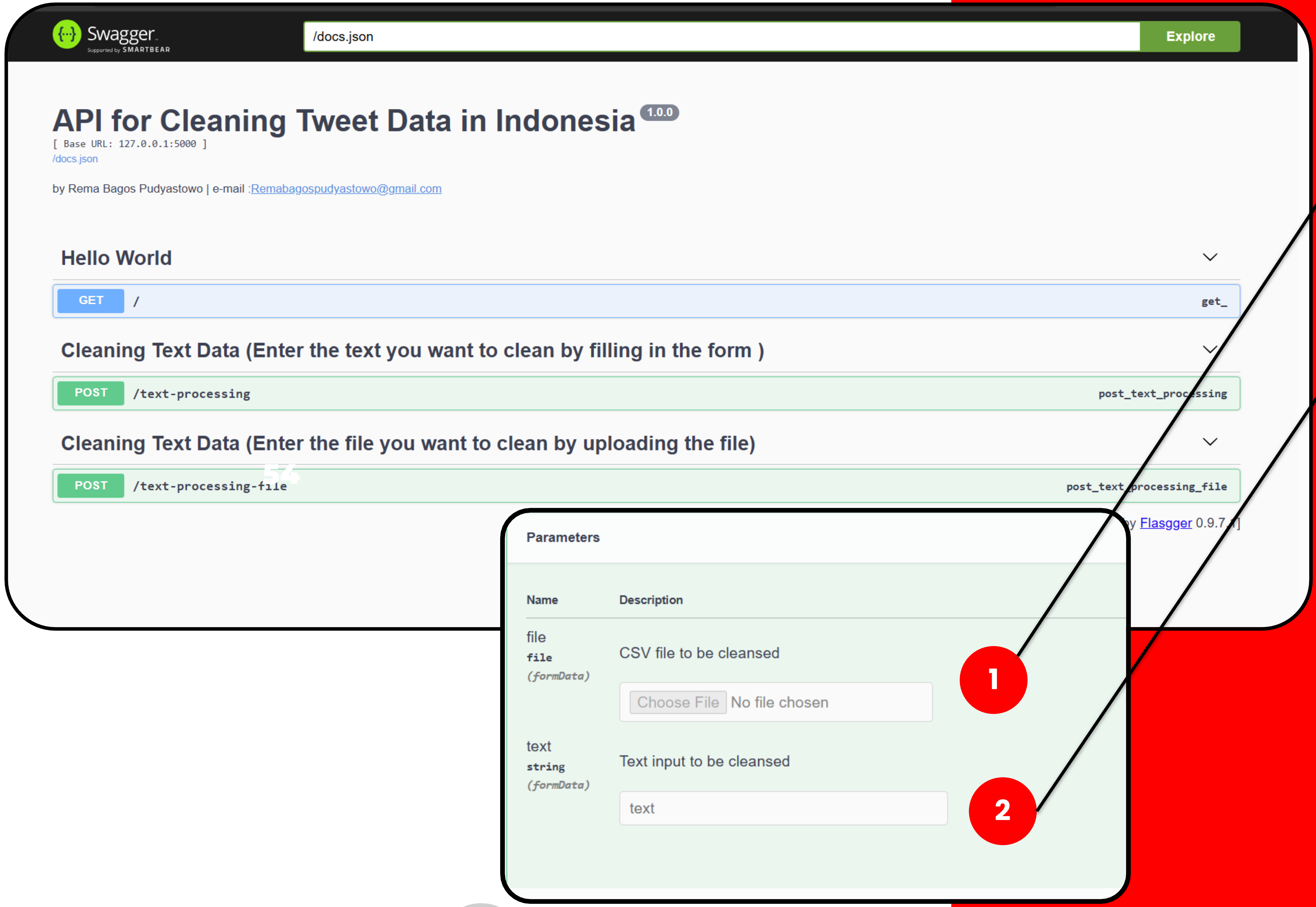
Abusive.csv



```python
def lowercase(text):
    return text.lower()


def remove_unnecessary_char(text):
    text = re.sub('\n',' ',text) # Remove every '\n'
    text = re.sub('rt',' ',text) # Remove every retweet symbol
    text = re.sub('user',' ',text) # Remove every username
    text = re.sub('((www\.[^\s]+)|(https?://[^\s]+)|(http?://[^\s]+))',' ',text) # Remove every URL
    pattern = re.compile(r'\\x[0-9A-Fa-f]{2}')
# Use the compiled pattern to replace matches in the text
    text = pattern.sub(' ', text)
    #text = re.sub((r'\\x[0-9A-Fa-f]{2}'),' ', text) #remove emoji
    text = re.sub('  +', ' ', text) # Remove extra spaces
    return text


def remove_nonaplhanumeric(text):
    text = re.sub('[^0-9a-zA-Z]+', ' ', text)
    return text

kamusalay_map = dict(zip(kamusalay['original'], kamusalay['replacement']))
def normalize_alay(text):
    return ' '.join([kamusalay_map[word] if word in kamusalay_map else word for word in text.split(' ')])


def remove_abusive(text):
    text = ' '.join(['' if word in abusive_dict.abusive.values else word for word in text.split(' ')])
    text = re.sub('  +', ' ', text) # Remove extra spaces
    text = text.strip()
    return text


def stemming(text):
    return stemmer.stem(text)
```

Cleaned data

API



API for Cleaning Tweet Data in Indonesia 1.0.0

[Base URL: 127.0.0.1:5000]
/docs.json

by Rema Bagos Pudyastowo | e-mail :Remabagospudyastowo@gmail.com

Hello World

GET / get_

Cleaning Text Data (Enter the text you want to clean by filling in the form)

POST /text-processing post_text_processing

Cleaning Text Data (Enter the file you want to clean by uploading the file)

POST /text-processing-file post_text_processing_file

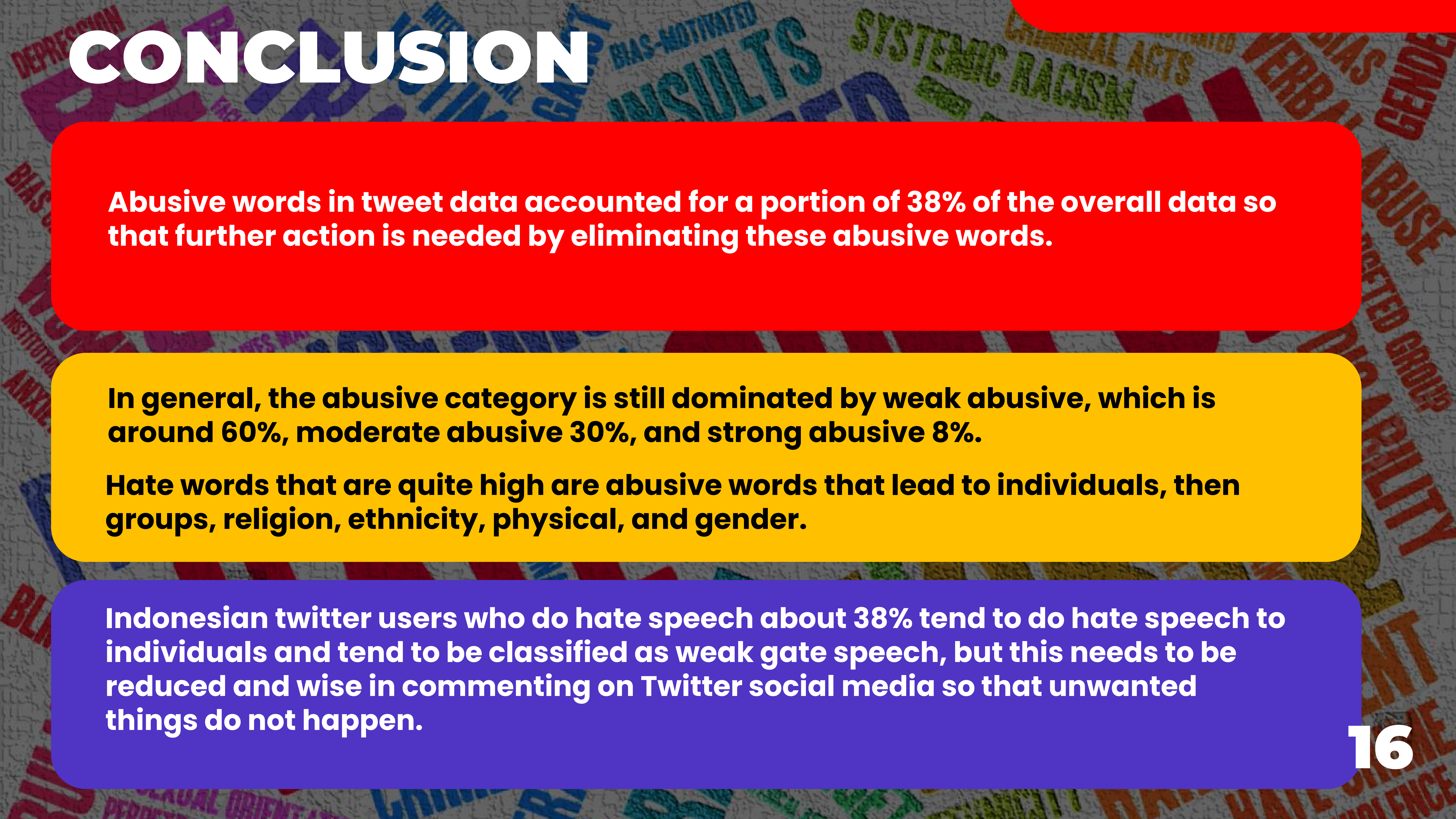
[Powered by Flasgger 0.9.7.1]

Parameters

Name	Description
file file *(formData)*	CSV file to be cleansed Choose File No file chosen
text string *(formData)*	Text input to be cleansed text

Feature

1. Cleaning text data through form
2. Cleaning text data through file upload (.csv)



CONCLUSION

Abusive words in tweet data accounted for a portion of 38% of the overall data so that further action is needed by eliminating these abusive words.

In general, the abusive category is still dominated by weak abusive, which is around 60%, moderate abusive 30%, and strong abusive 8%.

Hate words that are quite high are abusive words that lead to individuals, then groups, religion, ethnicity, physical, and gender.

Indonesian twitter users who do hate speech about 38% tend to do hate speech to individuals and tend to be classified as weak gate speech, but this needs to be reduced and wise in commenting on Twitter social media so that unwanted things do not happen.

BINAR ACADEMY

DataScienceProjects

THANKS FOR

Reading My Portfolio!

Rema Bagos Pudyastowo
Binar Academy

Linkedin : Rema Bagos Pudyastowo

Email : remabagospudyastowo@gmail.com

Github :

namesurname @namesurname • 13 Dec
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.
13 36 38

namesurname @namesurname • 25 Nov
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim.
25 12 15

HOW TO USE THIS API

1. Acces : http://127.0.0.1:5000/
2. Pick endpoint based on your purpose :
3. '/' to acces guide
4. '/text-processing to acces API that can cleansing your text data through filling the form
5. '/text-processing-file' to acces API that can cleansing your text data through upload file

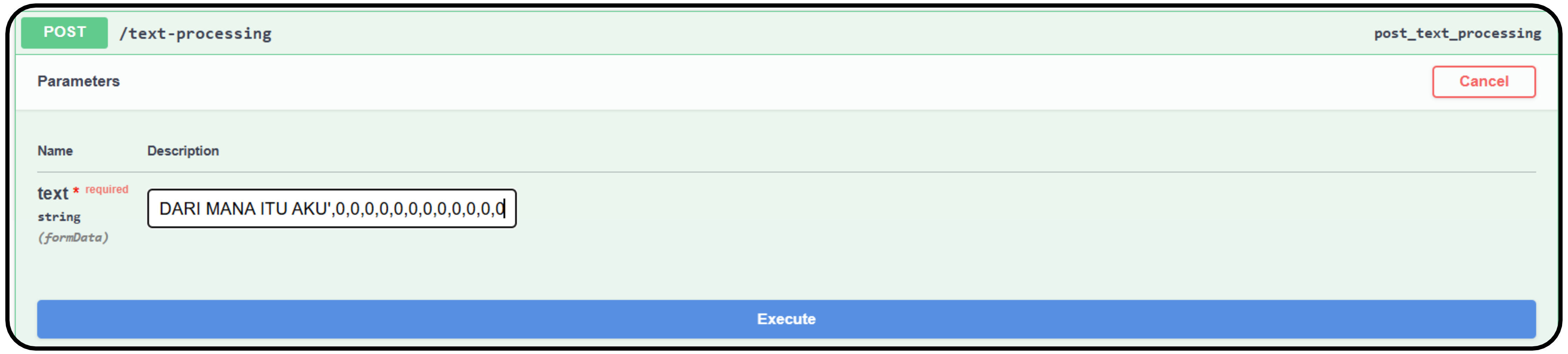
And

If you want to more beautiful and user friendly experience :

Acces this link:

http://127.0.0.1:5000/docs

Cleaning Data Through Form



POST /text-processing post_text_processing

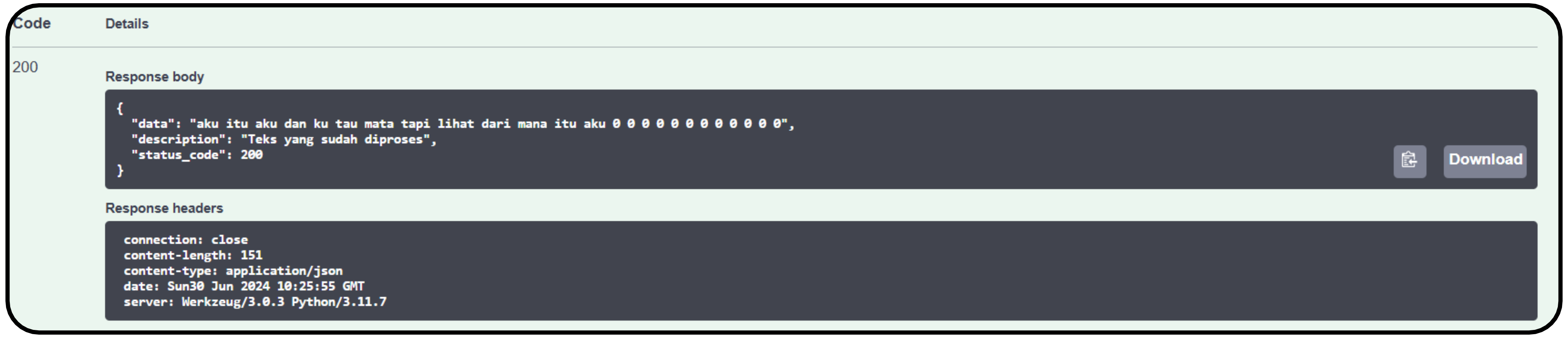
Parameters Cancel

Name	Description
text * required string *(formData)*	DARI MANA ITU AKU',0,0,0,0,0,0,0,0,0,0,0,0

Execute

Before Cleaning

Click Execute to cleaning text data



Code Details

200

Response body

```
{
  "data": "aku itu aku dan ku tau mata tapi lihat dari mana itu aku 0 0 0 0 0 0 0 0 0 0 0 0",
  "description": "Teks yang sudah diproses",
  "status_code": 200
}
```

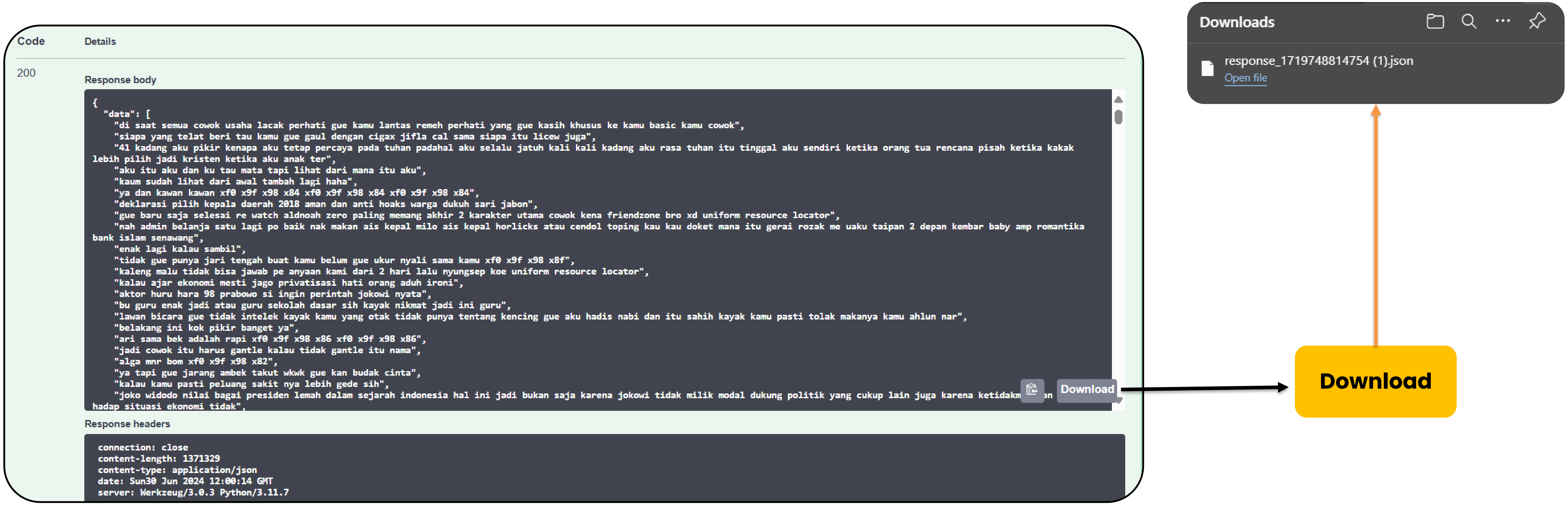
Download

Response headers

```
connection: close
content-length: 151
content-type: application/json
date: Sun30 Jun 2024 10:25:55 GMT
server: Werkzeug/3.0.3 Python/3.11.7
```

After Cleaning

Cleaning Data Through Upload File



```
{
  "data": [
    "di saat semua cowok usaha lacak perhati gue kamu lantas remeh perhati yang gue kasih khusus ke kamu basic kamu cowok",
    "siapa yang telat beri tau kamu gue gaul dengan cigax jifla cal sama siapa itu licew juga",
    "41 kadang aku pikir kenapa aku tetap percaya pada tuhan padahal aku selalu jatuh kali kali kadang aku rasa tuhan itu tinggal aku sendiri ketika orang tua rencana pisah ketika kakak
lebih pilih jadi kristen ketika aku anak ter",
    "aku itu aku dan ku tau mata tapi lihat dari mana itu aku",
    "kaum sudah lihat dari awal tambah lagi haha",
    "ya dan kawan kawan xf0 x9f x98 x84 xf0 x9f x98 x84 xf0 x9f x98 x84",
    "deklarasi pilih kepala daerah 2018 aman dan anti hoaks warga dukuh sari jabon",
    "gue baru saja selesai re watch aldnoah zero paling memang akhir 2 karakter utama cowok kena friendzone bro xd uniform resource locator",
    "nah admin belanja satu lagi po baik nak makan ais kepal milo ais kepal horlicks atau cendol toping kau kau doket mana itu gerai rozak me uaku taipan 2 depan kembar baby amp romantika
bank islam senawang",
    "enak lagi kalau sambil",
    "tidak gue punya jari tengah buat kamu belum gue ukur nyali sama kamu xf0 x9f x98 x8f",
    "kaleng malu tidak bisa jawab pe anyaan kami dari 2 hari lalu nyungsep koe uniform resource locator",
    "kalau ajar ekonomi mesti jago privatisasi hati orang aduh ironi",
    "aktor huru hara 98 prabowo si ingin perintah jokowi nyata",
    "bu guru enak jadi atau guru sekolah dasar sih kayak nikmat jadi ini guru",
    "lawan bicara gue tidak intelek kayak kamu yang otak tidak punya tentang kencing gue aku hadis nabi dan itu sahih kayak kamu pasti tolak makanya kamu ahlun nar",
    "belakang ini kok pikir banget ya",
    "ari sama bek adalah rapi xf0 x9f x98 x86 xf0 x9f x98 x86",
    "jadi cowok itu harus gantle kalau tidak gantle itu nama",
    "alga mnr bom xf0 x9f x98 x82",
    "ya tapi gue jarang ambek takut wkwk gue kan budak cinta",
    "kalau kamu pasti peluang sakit nya lebih gede sih",
    "joko widodo nilai bagai presiden lemah dalam sejarah indonesia hal ini jadi bukan saja karena jokowi tidak milik modal dukung politik yang cukup lain juga karena ketidakm   an
hadap situasi ekonomi tidak",
```

```
connection: close
content-length: 1371329
content-type: application/json
date: Sun30 Jun 2024 12:00:14 GMT
server: Werkzeug/3.0.3 Python/3.11.7
```

After cleaning data its over , you can download your cleaned data(.json) and store it into database with sqlite3