



API UNTUK DATA CLEANSING & ANALISIS DATA TWEET "HATE SPEECH"

BY ALEX APRIANDI

PENDAHULUAN

LATAR BELAKANG

Hate speech atau ujaran kebencian adalah suatu bentuk ekspresi yang dilakukan untuk menyebarkan rasa kebencian dan melakukan tindakan kekerasan serta diskriminasi terhadap seseorang atau sekelompok orang karena berbagai alasan. Kasus *hate speech* sangat sering kita jumpai di media sosial, salah satunya di Twitter.

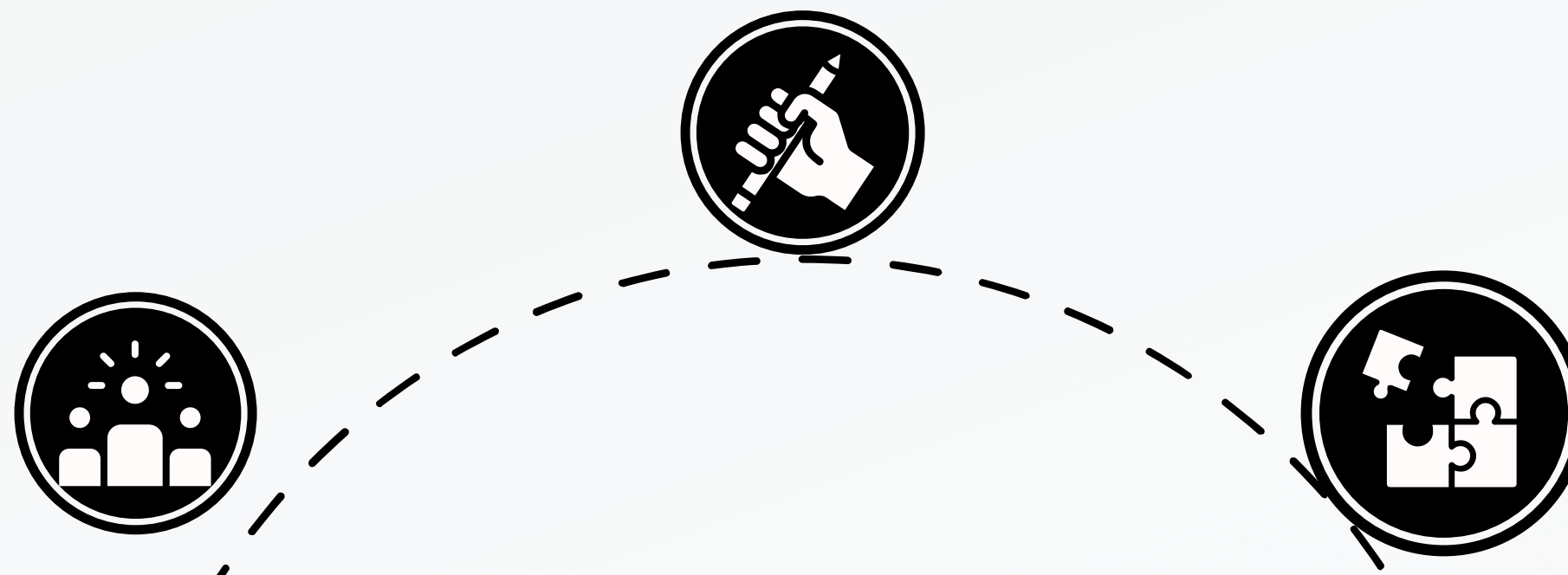
Penelitian ini akan menggambarkan bagaimana pembuatan API untuk cleansing data tweet "*Hate Speech*" yang ada di Twitter serta akan menganalisis data teks tersebut dengan metode-metode yang relevan.

RUMUSAN MASALAH

- Bagaimana gambaran secara statistik pola interaksi di media sosial tweeter yang mengandung unsur *Hate Speech* ?
- Bagaimana melakukan cleansing atas data tweet yang mengandung *Hate Speech* ?

TUJUAN PENELITIAN

- Untuk mengetahui gambaran pola interaksi para pengguna twitter yang me "Tweet" hal yang bersifat "*Hate Speech*" di media sosial Twitter
- Untuk membuat API yang digunakan untuk cleansing data tweet yang mengandung "*Hate Speech*"



METODE PENELITIAN



DESKRIPSI DATA

Data yang digunakan dalam penelitian ini adalah tweet dari user-user twitter yang mengandung unsur hate speech dan data ini termasuk kedalam jenis data sekunder karena bersumber web kaggle.



CLEANSING DATA

- Melakukan Cleansing data text tweet dengan menggunakan python library Regular Expression (RegEx): seperti: url, baris baru, double backslash (//), backslash n dan x, double space, user name, hashtag, bcc, dan semua symbol selain angka dan huruf.
- Melakukan Cleansing data text dengan mengganti kata-kata alay dan kata-kata kasar dengan kata-kata baku yang telah dilampirkan dalam database
- Membuat API dibuat dengan Flask dan Swagger UI
- Penyimpanan data menggunakan SQLite (SQLite 3)



METODE ANALISIS DATA & VISUALISASI

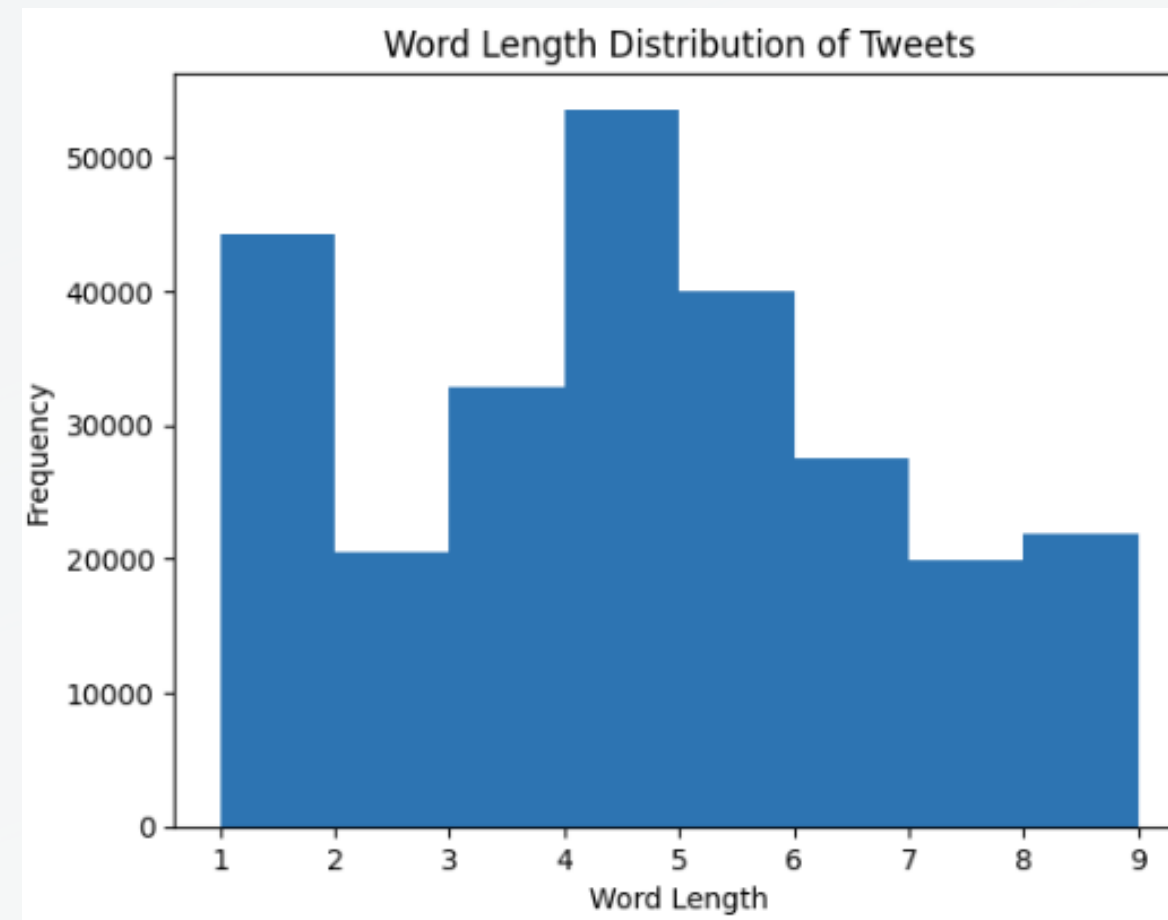
- Menggunakan metode statistik "*descriptive analytics*" dan melakukan melakukan Exploratory Data Analytics (EDA) untuk menggambarkan kondisi, pola dan tren dari data.
- Menggunakan fungsi-fungsi python library untuk pengolahan data seperti : Pandas untuk dataframe, numpy untuk operasi numerik, wordcloud untuk menampilkan kata-kata yang sering muncul, dll
- Menggunakan fungsi dari python library seperti seaborn dan matplotlib untuk Memvisualisasi data statistik dalam bentuk grafik

HASIL PENELITIAN

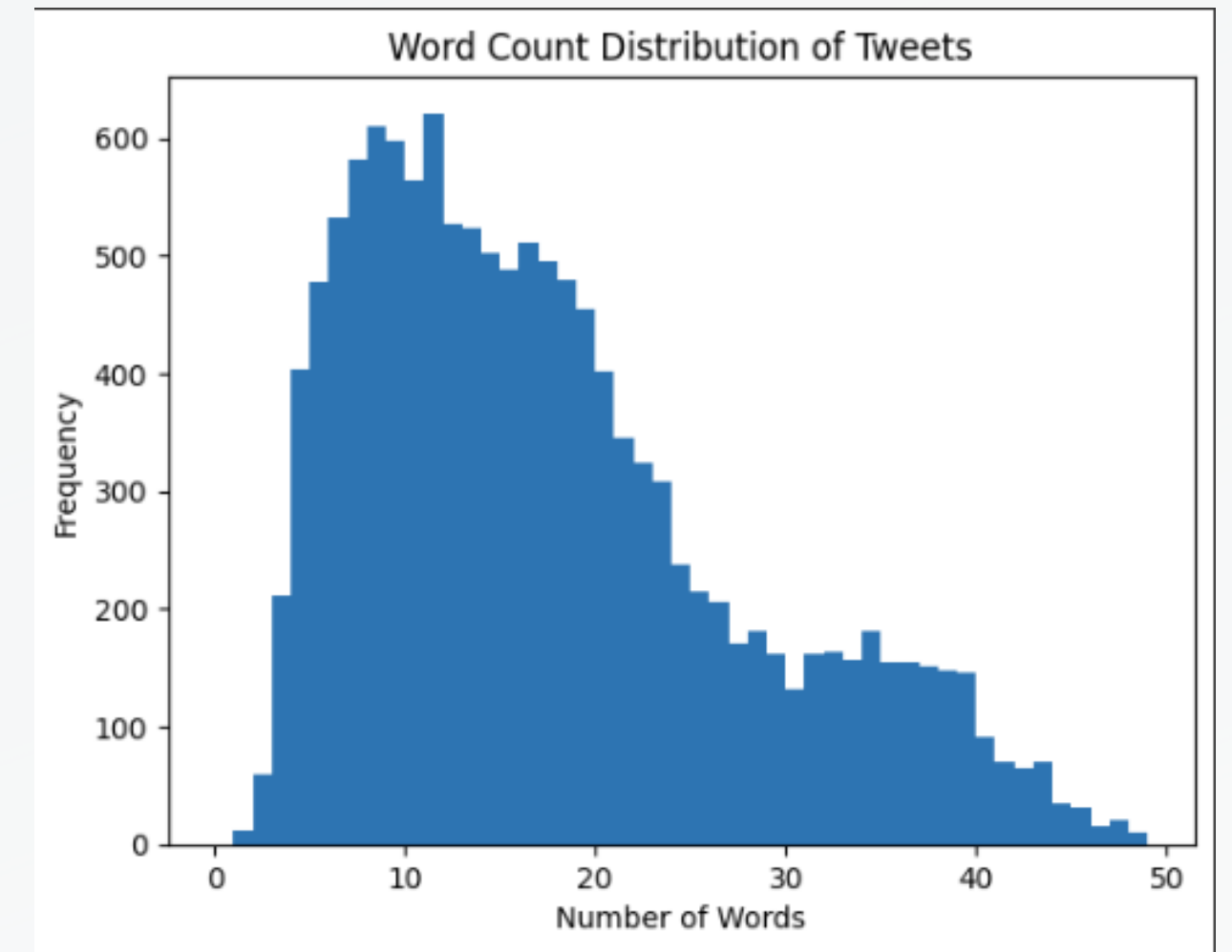
Kata-kata yang paling banyak muncul pada tweet



Distribusi Panjang Kata pada Data Tweet




Distribusi Jumlah Kata pada Data Tweet



HASIL PENELITIAN

Tampilan Text Cleansing dengan Output Text

 Swagger
Supported by SMARTBEAR

/docs.json

Explore

API Text Output 1.0.0 // BETA

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

Text Cleansing Output Clean text

Please Input the Word

POST /input_data

[Powered by [Flasgger](#) 0.9.5]

input_data * required
string
(formData)

- disaat semua cowok berusaha melacak per

Execute

Clear

Responses

Response content type [application/json](#)

Curl

```
curl -X POST "http://127.0.0.1:5000/input_data" -H "accept: application/json" -H "Content-Type: application/x-www-form-urlencoded" -d "input_data=-%20disaat%20semua%20cowok%20berusaha%20melacak%20perhatian%20gue.%20lo%20lantas%20remehkan%20perhatian%20yg%20gue%20kasih%20khusus%20ke%20lo.%20basic%20elo%20cowok%20bego%20!%20!%20!"
```

Request URL

http://127.0.0.1:5000/input_data

Server response

Code	Details
200	<div>Response body</div> <div><pre>{ "input": "- disaat semua cowok berusaha melacak perhatian gue. loe lantas remehkan perhatian yg gue kasih khusus ke elo. basic elo cowok bego ! ! !", "output": "di saat semua cowok berusaha melacak perhatian saya kamu lantas remehkan perhatian yang saya kasih khusus ke kamu basic kamu cowok"}</pre></div> <div>Download</div>


Response headers

```
connection: close
content-length: 293
content-type: application/json
date: Mon, 22 May 2023 10:02:09 GMT
server: Werkzeug/2.3.4 Python/3.11.3
```

Responses

Code	Description
------	-------------

Tampilan Text Cleansing dengan Output File

 Swagger
Supported by SMARTBEAR

/docs.json

Explore

API File Output 1.0.0 // BETA

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

File Cleansing Output Clean File

Text Processing File

POST /input-file

[Powered by [Flasgger](#) 0.9.5]

Name	Description
file <small>* required</small> file (formData)	File to upload <div>Choose File data.csv</div>

Execute

Clear

Responses

Response content type [application/json](#)

Curl

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

Request URL

http://127.0.0.1:5000/input-file

Server response

Code	Details
200	<div>Response body</div> <div><pre>{ "data": ["twit hs abusive hs individual hs grup hs religion hs race hs physical hs gender hs other hs weak hs moderate hs strong", "di saat semua cowok berusaha melacak perhatian saya kamu lantas remehkan perhatian yang saya kasih khusus ke kamu basic kamu cowok", "rt pengguna pengguna siapa yang telat memberi tau kamu saya bergaul dengan cigax jifla calis sama siapa itu licew juga", "di kadang aku berpikir kenapa aku tetap percaya pada tuhan padahal aku selalu jatuh berkali kali kadang aku merasa tuhan itu meninggalkan aku sendirian ketika orang tuaku berencana berpisah ketika kakaku lebih memilih jadi kristen ketika aku anak ter", "pengguna pengguna aku itu aku dan ku tau matamu tapi dilihat dari mana itu aku", "pengguna pengguna kaum sudah kelihatan dongoknya dari awal tambah lagi haha", "pengguna ya dan kawan kawan", "deklarasi pilihan kepala daerah 2018 aman dan anti hoaks warga dukuh sari jabon", "saya baru saja selesai re watch aldoah zero paling memang akhirnya karakter utama cowoknya kena friendzone bro xd uniform resource locator", "nah admin belanja satu lagi port terbaik nak makan ais kepal milo ais kepal horlicks atau cendol toping kau kau doket mana itu gerald rozak mertuaku talpan depan kembar baby ump romantika bank islam senang", "pengguna enak lagi kalau sambil", "setidaknya saya punya jari tengah buat kamu sebelum saya ukur nyali sama kamu", "pengguna pengguna pengguna pengguna kaleng malu tidak bisa jawab pertanyaan kami dari hari lalu nyungsep koe pengguna uniform resource locator", "kalau belajar ekonomi mestinya jago memprivatisasi hati orang aduh ironi pengguna", "aktor huru hara 98 prabowo si ingin lengserkan pemerintahan jokowi nyata", "pengguna bu guru enakan jadi atau guru sekolah dasar sih kayaknya menikmati jadi ini guru",]}</pre></div>

KESIMPULAN



Dengan dilakukannya penelitian ini, maka pembaca dapat mendapatkan gambaran terkait "*Hate Speech*" di media sosial Twitter.

Berdasarkan penelitian ini, terjadinya "*Hate Speech*" di Twitter relatif sering terjadi. Pembaca dapat melihat kata-kata yang paling sering digunakan, panjang karakter dan kata para user Twitter saat terjadinya "*Hate Speech*".

Selain itu dengan adanya API untuk cleansing data, kita dapat melakukan cleansing pada data Teks dengan tujuan agar informasi yang kita baca terhindar dari adanya kata-kata yang kurang sesuai, tanda baca yang kurang tepat dan penggunaan simbol-simbol yang tidak perlu.

SARAN



Melalui hasil penelitian ini, diharapkan kepada pembaca untuk menggunakan media sosial seperti Twitter secara bijaksana dan kehati-hatiannya pada saat akan me "Tweet" sesuatu sehingga tidak mengandung "*Hate Speech*". Hal ini dikarenakan isu ini cukup sensitif, untuk itu perlu diharapkan kehati-hatiannya ketika membuat tweet, seperti: menggunakan ejaan yang sesuai, tidak menggunakan simbol-simbol yang tidak perlu, dll.

Penelitian ini dapat dikembangkan lagi menjadi sebuah aplikasi yang lebih baik untuk mendeteksi adanya "*Hate Speech*" yang bersumber dari Twitter.