BINAR GOLD CHALLAGE

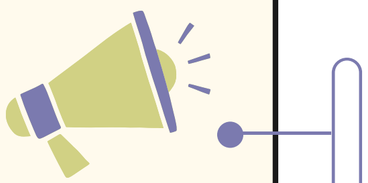
Tweet

Trends

PROGRAM API UNTUK CLEANSING DAN ANALISIS DATA

TWEET UJARAN KEBENCIAN BAHASA INDONESIA

Presented by Yaya Hidayana



Hatespeech

Feedback

LATAR BELAKANG

1 PERTUMBUHAN INTERNET

Pertumbuhan internet di Indonesia sangat pesat dalam beberapa tahun terakhir. Menurut data survei yang diterbitkan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), mencatat penetrasi 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.

2 TWITTER

Salah satu platform media sosial yang sangat populer di Indonesia adalah Twitter. Menurut data StatCounter, pada bulan Maret 2023, Twitter merupakan platform media sosial terpopuler kedua di Indonesia setelah Facebook, dengan pangsa pasar sebesar 12,5%.

3 DAMPAK NEGATIF

Penggunaan sosial media khususnya Twitter juga membawa dampak negatif, seperti meningkatnya hate speech atau ujaran kebencian yang dapat memicu konflik sosial dan memperburuk situasi keamanan di masyarakat.

LATAR BELAKANG

4 RUMUSAN MASALAH

- Bagaimana gambaran secara statistik para pengguna twitter yang membahas isu terkait ujaran kebencian?
- Bagaimana melakukan cleansing atas data tweet terkait ujaran kebencian?

5 TUJUAN

- Untuk mengetahui gambaran secara statistik para pengguna twitter yang membahas isu terkait ujaran kebencian
- Untuk membuat API yang digunakan untuk cleansing data tweet terkait ujaran kebencian

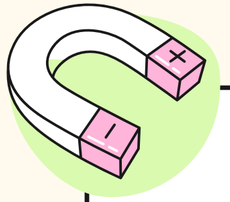
METODE PENELITIAN

DATA

Data tweet **Ujaran Kebencian Bahasa Indonesia** bersifat sekunder karena bersumber web kegel. Data terdiri dari 3 file, yaitu:

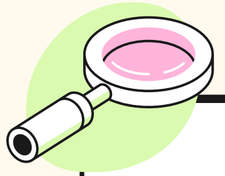
- abusive.csv > *kumpulan kata-kata kasar*
- data.csv > *file data utama*
- new_kamusalay.csv > *kumpulan kata tidak baku dan baku*

File utama **(data.csv)** terdiri dari **13.169 baris** dan **15 kolom**. Dipilih kolom "**Tweet**" untuk dibersihkan menggunakan program cleansing ini, karena sisa kolomnya hanya berisi parameter untuk melakukan klasifikasi dari jenis tweet itu sendiri, seperti: apakah termasuk topik ras, agama, fisik, jenis kelamin, ujaran kebencian, dll.



ANALISA DATA

- Memanipulasi data dalam bentuk DataFrame mengunakan pandas.
- Memvisualisasi data statistik dalam bentuk grafik dengan seaborn.
- Memvisualisasi data statistik dalam bentuk grafik dengan matplotlib.
- menghitung frekuensi elemen dalam sebuah list atau tuple dengan collections.
- Menghitung operasi numerik pada data dengan numpy.
- Menggambarkan kata-kata yang sering muncul dalam teks dengan wordcloud.

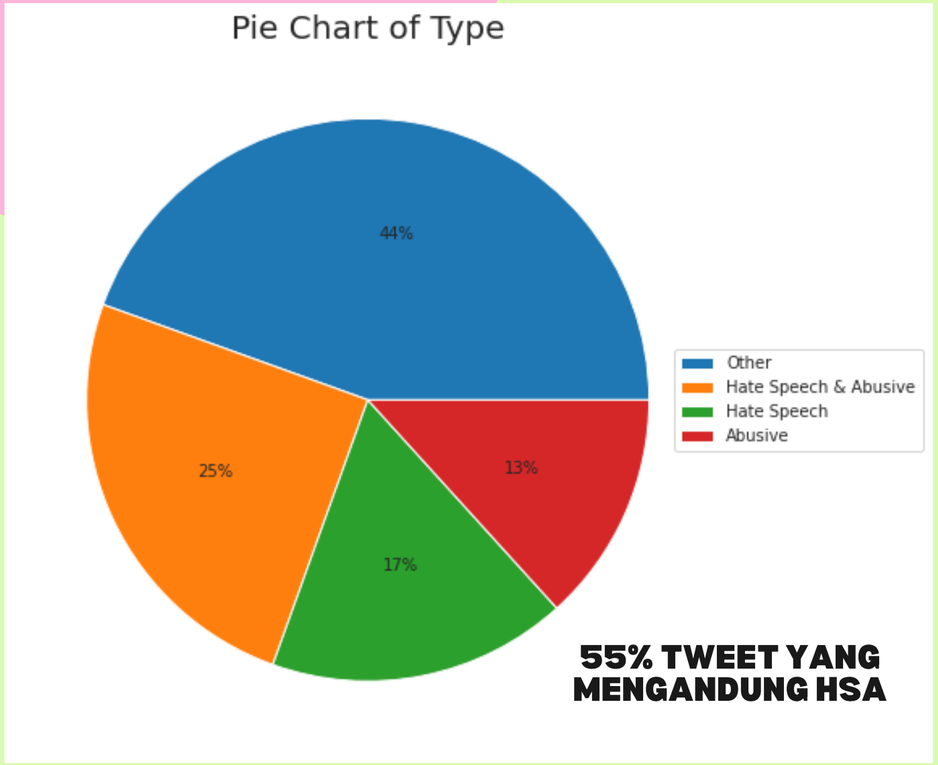


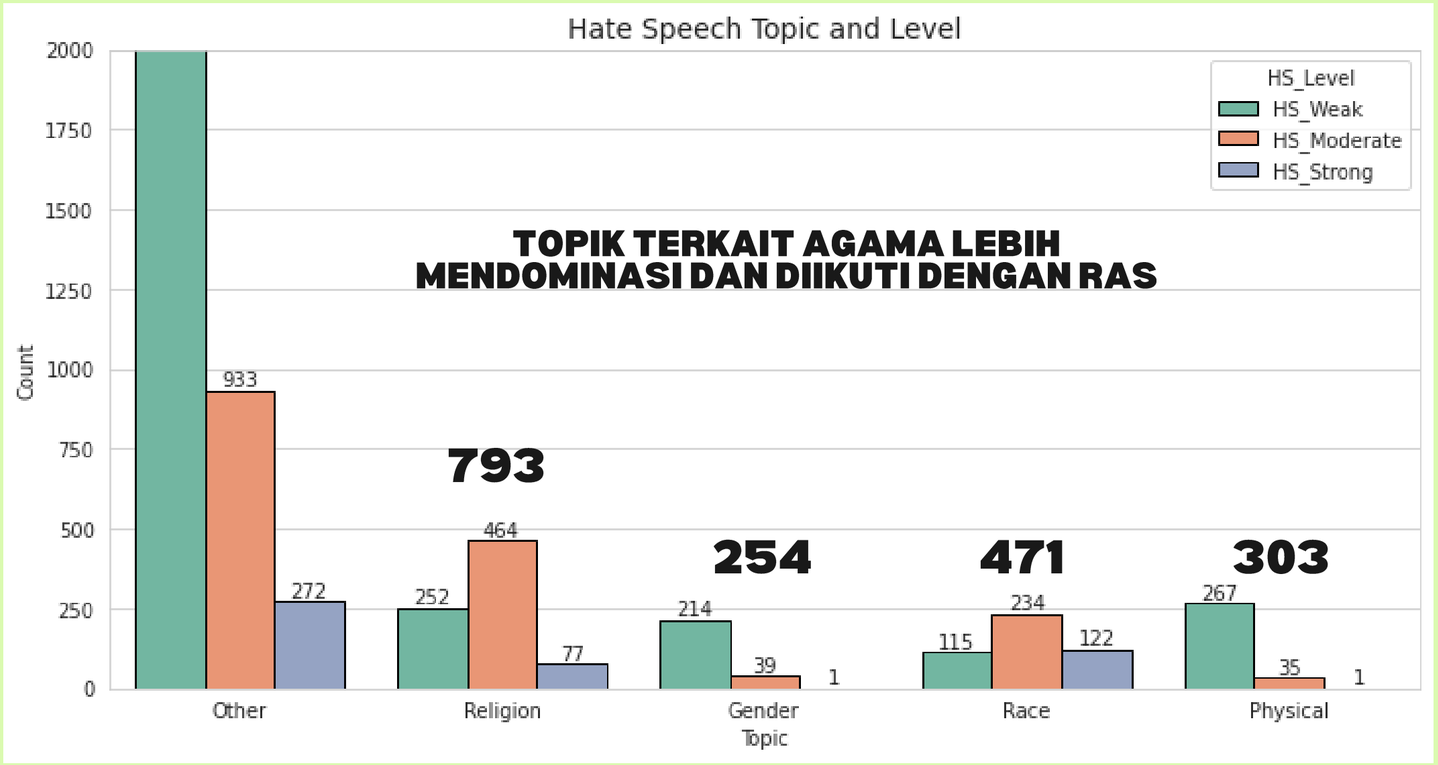
VISUALISASI

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

VISUALISASI

- Exploratory Data Analysis (EDA)





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

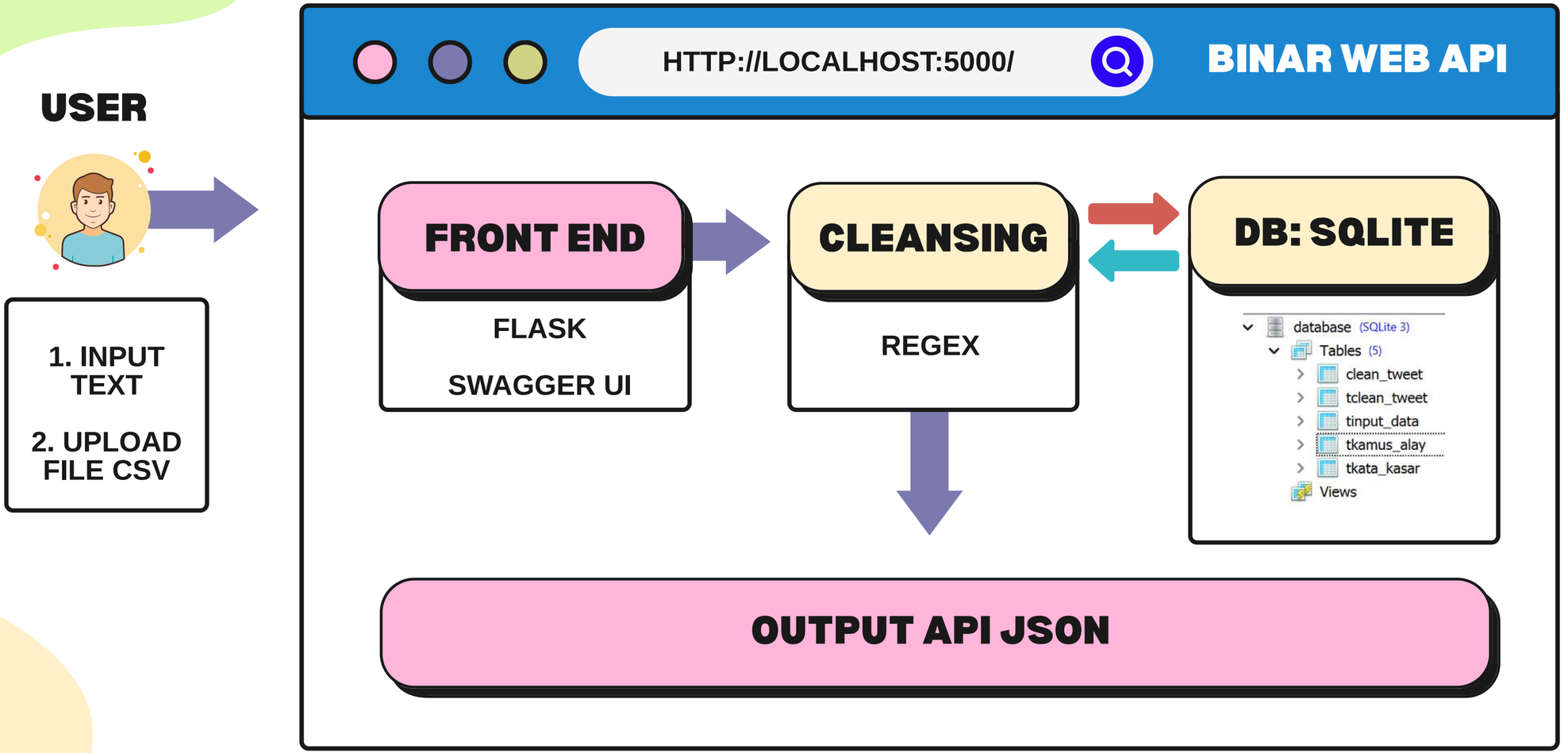
RELIGION

GENDER

RACE

PHYSICAL

VISUALISASI - Alur Proses Pembersihan Dan Penyimpanan Data

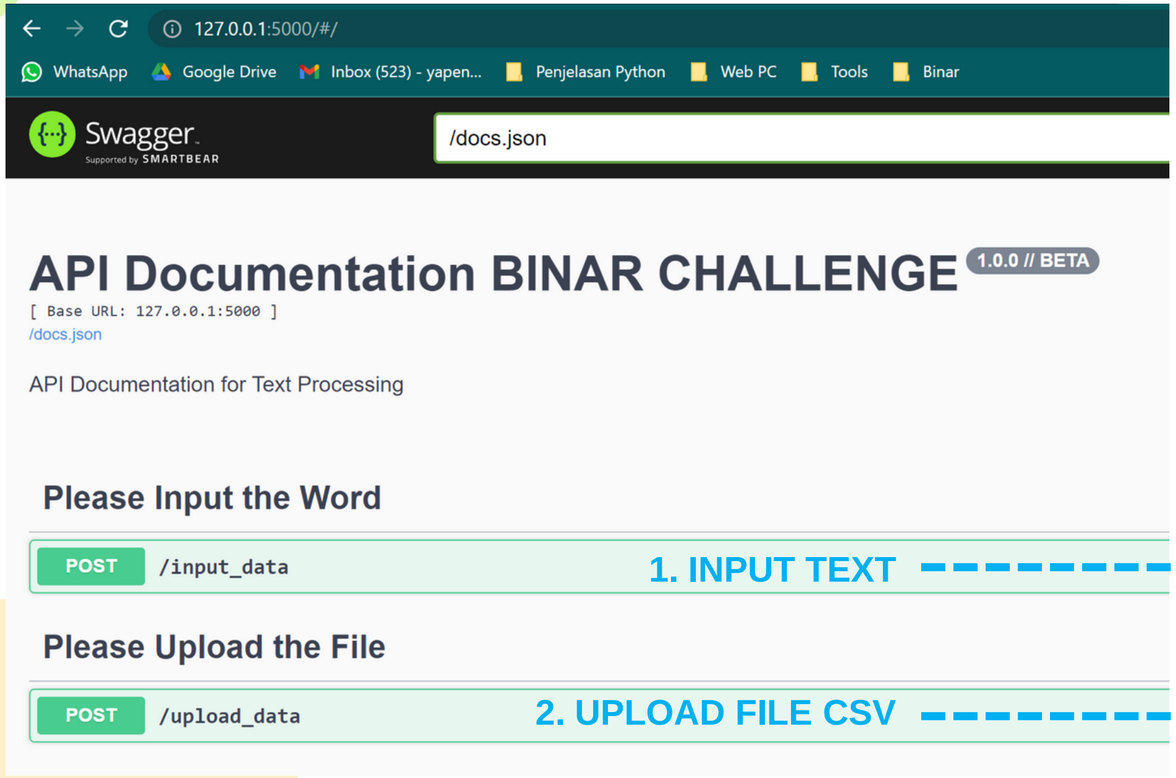


Ket:

Pada proses cleansing, input yang dimasukan pada fontend akan dibersihkan dari simbol, gambar, link, dll menggunakan modul regex, kemudian secara otomatis menghilangkan kata kasar dan kata yang tidak sesuai dengan ejaan baku (alay) sebagaimana data yang tersimpan pada tabel **tkamus_alay** dan **tkata_kasar**, kemudian hasilnya disimpan ke dalam database.

VISUALISASI - Tampilan Hasil Program

Tampilan Frontend Web API



OutPut

Code | Details

200

Response body

```
{
  "input": "lo anjing",
  "output": "kamu"
}
```

Response headers

```
connection: close
content-length: 38
content-type: application/json
date: Sun, 26 Mar 2023 14:43:03 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

OutPut

Code | Details

200

Response body


```
{
  "output": [
    "di saat semua cowok berusaha melacak perhatian saya kamu lantas remehkan perhatian yang saya k
    "siapa yang telat memberi tau kamu saya bergaul dengan cigax jifla calis sama siapa itu licew j
    "41 kadang aku berpikir kenapa aku tetap percaya pada tuhan padahal aku selalu jatuh berkali ka
cana berpisah ketika kakakku lebih memilih jadi kristen ketika aku anak ter",
    "aku itu aku dan ku tau matamu tapi dilihat dari mana itu aku",
    "kaum sudah kelihatan dongoknya dari awal tambah lagi haha",
    "ya dan kawan kawan",
    "deklarasi pilihan kepala daerah 2018 aman dan anti hoaks warga dukuh sari jabon",
    "saya baru saja selesai re watch aldnoah zero paling memang akhirnya 2 karakter utama cowoknya
    "nah admin belanja satu lagi po terbaik nak makan ais kepal milo ais kepal horlicks atau cendol
mantika bank islam senawang",
    "enak lagi kalau sambil",
    "setidaknya gue punya jari tengah buat kamu sebelum gue ukur nyali sama kamu",
    "kaleng malu tidak bisa jawab pe anyaan kami dari 2 hari lalu nyungsep koe uniform resource loc
    "kalau belajar ekonomi mestinya jago memprivatisasi hati orang aduh ironi",
    "aktor huru hara 98 prabowo si ingin lengserkan pemerintahan jokowi nyata",
    "bu guru enakan jadi atau guru sekolah dasar sih kayaknya menikmati jadi ini guru",
    "lawan bicara gue tidak intelek kayak kamu yang otak tidak punya tentang kencing gue mengakui h
    "belakangan ini kok pikiran banget ya",
    "ari sama beki adalah rapi",
    "jadi cowok itu harus gantle kalau tidak gantle itu namanya",
    "alga mnr bom",
    "ya tapi gue jarang mengambek takut wkwk gue kan budak cinta",
    "kalau kamu pasti peluang disakiti nya lebih gede sih",
    "joko widodo dinilai sebagai presiden terlemah dalam sejarah indonesia hal ini terjadi bukan sa
karena ketidakmampuan pemerintahannya menghadapi situasi ekonomi tidak",
```


Response headers

```
connection: close
content-length: 1442827
content-type: application/json
date: Sun, 26 Mar 2023 14:44:21 GMT
server: Werkzeug/2.2.3 Python/3.11.2
```

Ket: Pada contoh kasus sederhana, user memasukan input text secara langsung berisi kata: “Lo Anjing” maka output akan menjadi “Kamu”, karena kata “Lo” akan diubah menjadi bentuk baku yaitu “Kamu”, sedangkan kata “Anjing” akan dihapus karena termasuk kata kasar pada kalimat itu.

KESIMPULAN

Dari hasil analisa data, kita menadapat gambaran bahwa topik agama yang paling dominan ditulis dan direspon oleh pengguna twitter, karena bagi sebagian orang, agama merupakan nilai dan identitas yang sangat penting.

Dengan dibuatnya aplikasi ini, diharapkan dapat membantu dalam proses pembersihan data yang bersumber dari Twitter, sehingga dapat mempercepat dan menyederhanakan proses tersebut.

SARAN

Penelitian sederhana ini dapat diperluas dengan menambahkan analisis sentimen (positif dan negatif) dari setiap tweet menggunakan framework Natural Language Processing (NLP), sehingga dapat memperkaya analisis yang dilakukan.

Dapat dikembangkan menjadi program otomatis yang dapat mendeteksi penyebaran tweet yang dapat memecah persatuan dan kesatuan bangsa yang terhubung dengan K/L terkait.