

DATA CLEANSING API FLAGGER

Rachmawati Oktavia

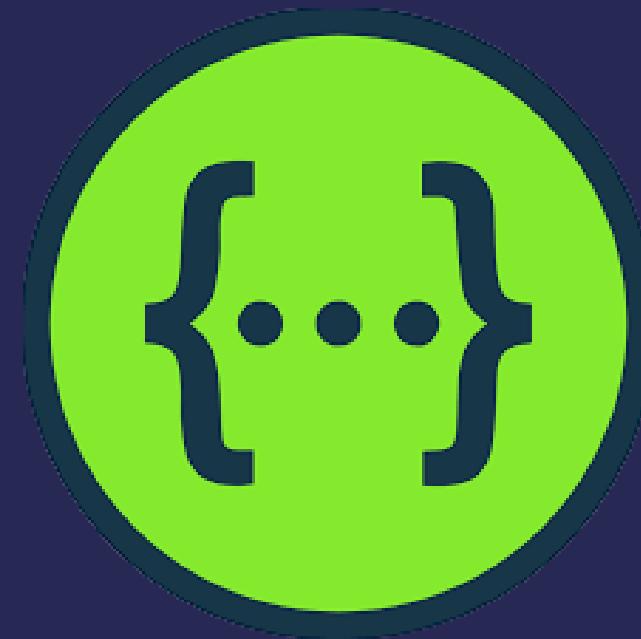


PENDAHULUAN

Merupakan syarat untuk menyelesaikan Level Gold pada Data Scientist Class of Binar Academy

API digunakan untuk melakukan cleansing pada kalimat yang mengandung kata-kata alay dengan penulisan yang tidak proper dan menggantikan tulisan bermuatan kata abusive dengan sensor

Tools dan bahasa yang digunakan:



METODOLOGI

1. Import ke dalam sqlite csv data tweet, kamus alay, dan list abusive kemudian ditranslate menjadi dataframe dan list.
2. Bersihkan tweet dari tulisan yang tidak diperlukan, rubah value menjadi string lower untuk mempermudah proses selanjutnya
3. Proses cleansing alay untuk membersihkan kata-kata yang menggunakan bahasa/tulisan alay
4. Proses cleansing tulisan yang bermuatan kata-kata abusive untuk dilakukan sensor
5. Keluaran berupa kata atau file csv yang telah dibersihkan dari alay dan abusive



HASIL & KESIMPULAN

API telah berjalan untuk cleansing data, namun masih terdapat hal yang belum dapat diproses.

Cleansing data dapat dibangun menggunakan Python dengan bahasa yang cukup sederhana dan berbagai tools yang tersedia gratis

Memampukan para newbie untuk mempelajari dan menjadi seorang data scientist, cleansing data dan memproses data lebih lanjut

ANALYSIS FROM DATA

56%

7.309 record from total tweet data of 13.169 record
is worth to be censored because:



17%

2.266 tweet contain hate speech
1.748 tweet contain abusive

25%

3.295 contain both hate speech & abusive



ANALYSIS FROM DATA

1ST

most hate speech towards group or individual both contain invective or slanders

2ND

2nd most hate speech towards group are hate speech targeted religion and hate speech towards individual target the physical or disability



GROUP

racist hate
speech

targeted
gender

targeted
physical

3RD

targeted
religion

4TH

targeted
gender

5TH

targeted
race



INDIVIDUAL

THANK YOU

