

Analisis Karakter dan Identifikasi Kata-Kata Abusive dalam Data Tweet menggunakan Descriptive Analytics

Date: November 12th, 2023

Prepared by: Syaeful Bahri





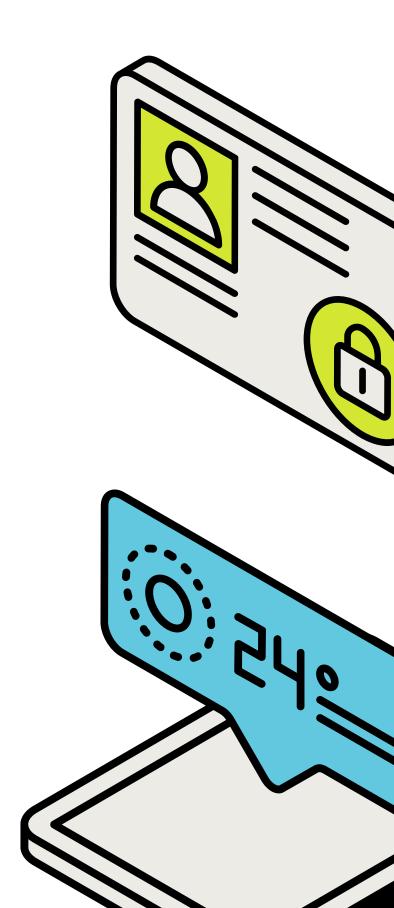
Pendahuluan



Twitter yang sekarang menjadi X menjadi salah satu media sosial yang digunakan baik skala indvidu atau kelompok untuk berbagi ide, pikiran, pandangan bahkan promosi tertentu yang terus berkembang. Hal ini berimplikasi langsung terhadap penggunaan bahasa dan perilaku dari platform tersebut.

Sehingga hal ini menjadi perhatian penting untuk kemudian dilakukan kajian yang mendalam untuk menganalisis karakter dan identifikasi kata-kata abusive dalam data tweet dalam memahamai interaksi digital yang lebih baik

Penelitian ini fokus terhadap identifikasi kata-kata abusive dengan menggunkan Descriptive Analytics. Sehingga harapannya dapat kemudian memberikan wawasan lebih lanjut terkait perilaku komunikatif di Twitter atau X.





Tujuan Penelitian



(1)

Analisis Karakter
Tweet

(2)

Identifikasi Kata-Kata
Abusive

3

Descriptive Analytics

Memahami karakteristik unik dari tweet, termasuk panjang teks bahasa untuk mengidentifikasi pola-pola yang mungkin terkait dengan sentimen

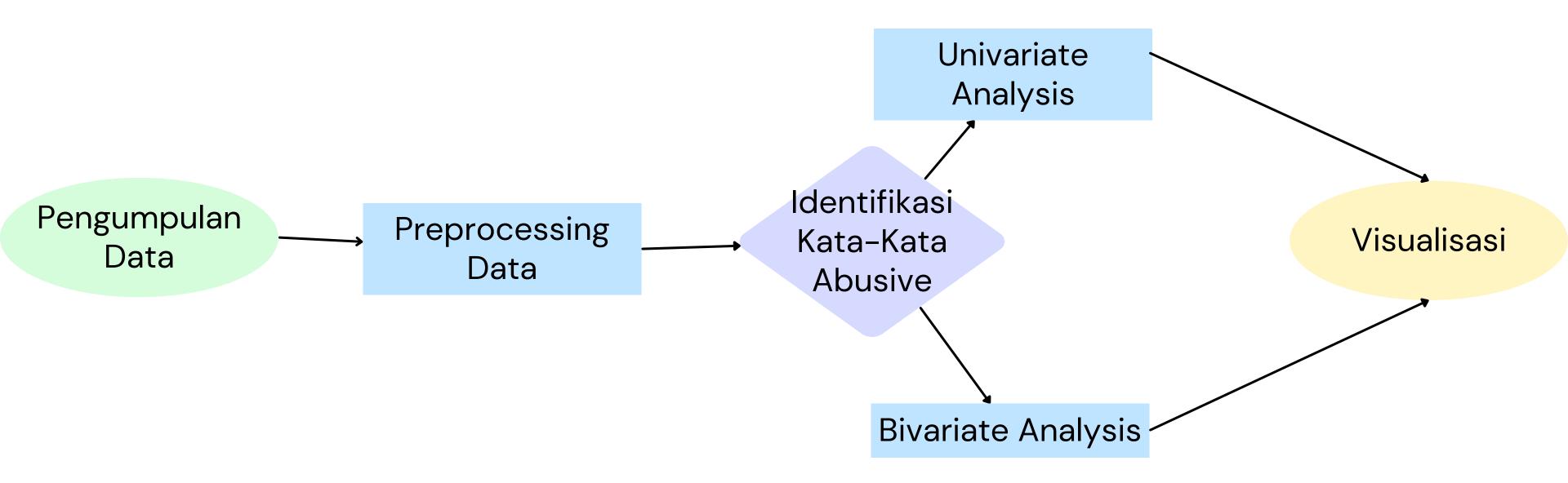
Menelusuri dataset tweet untuk mengidentifikasi kata-kata yang bersifat abusive

Menerapkan pendekatan deskriptif untuk memberikan gambaran umum tentang distribusi karakter dalam tweet dan frekuensi katakata abusive dalam dataset



Metode Penelitian





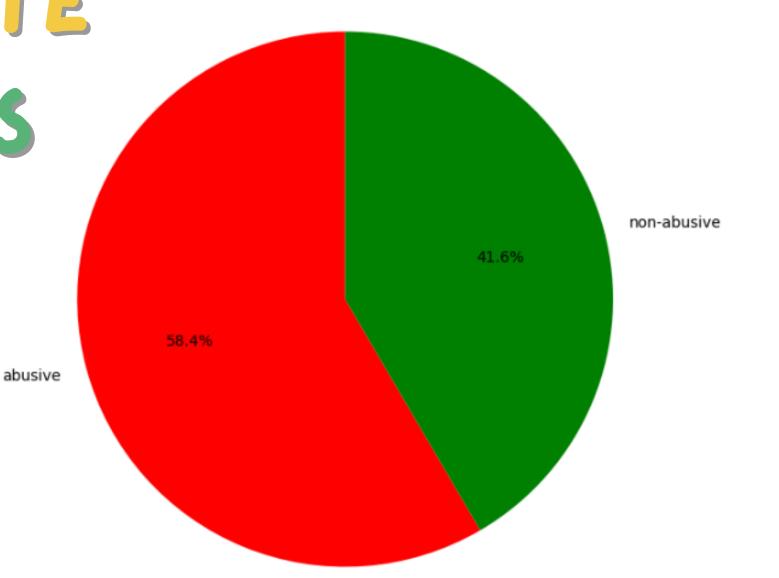


Hasildan Kesimpulan

Distribution of Labels

UNIVARIATE ANALYSIS





label abusive

38490

non-abusive 27355

Name: count, dtype: int64

	label	total_word	total_char		
		mean	mean		
0	abusive	19.119122	120.091322		
1	non-abusive	20.777006	137.470115		
mean					

	label	total_word	total_char
		median	median
0	abusive	17.0	105.0
1	non-abusive	18.0	123.0

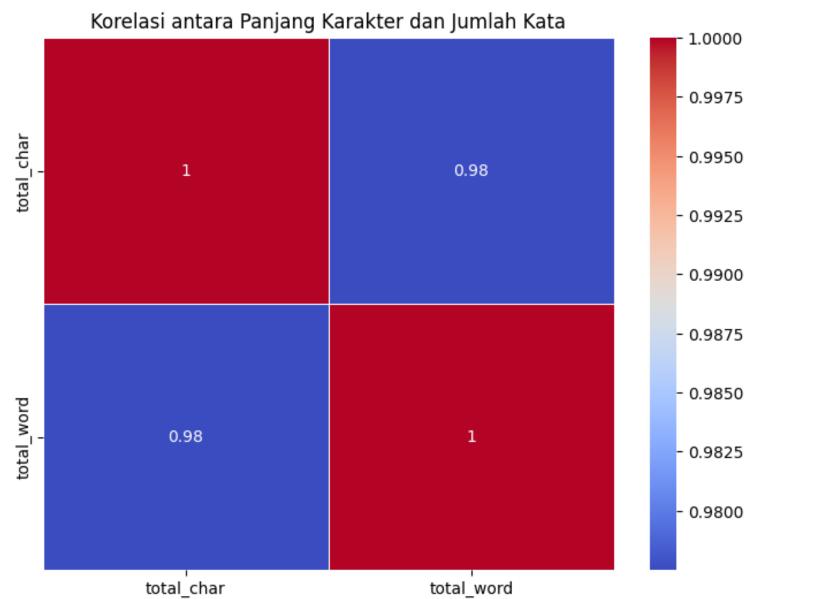
median

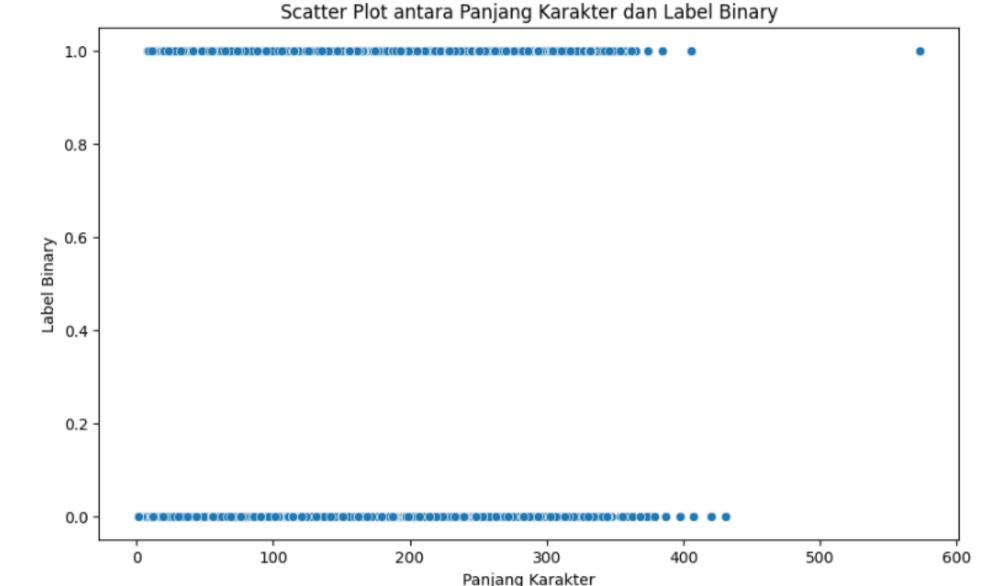
	label	total_word	total_char
0	abusive	7	61
1	non-abusive	[13, 16]	79

modus

BIVARIATE ANALYSIS







Dalam Descriptive Statistic menunjukkan variabel total karakter dan total kata memiliki korelasi positif.

Namun, saat hendak menentukan korelasi antara panjang karakter dengan label abusive, tidak memiliki korelasi yang positif

```
correlation = tweets['total_char'].corr(tweets['binary_label'])
print(f'Korelasi antara Panjang Karakter dan Label Abusive: {correlation}')
```

Korelasi antara Panjang Karakter dan Label Abusive: -0.11359752107065528

TOP 10 WORDS



10 kata terbanyak:

indonesia: 8360

presiden: 7550

orang: 7125

jokowi: 5805

islam: 4430

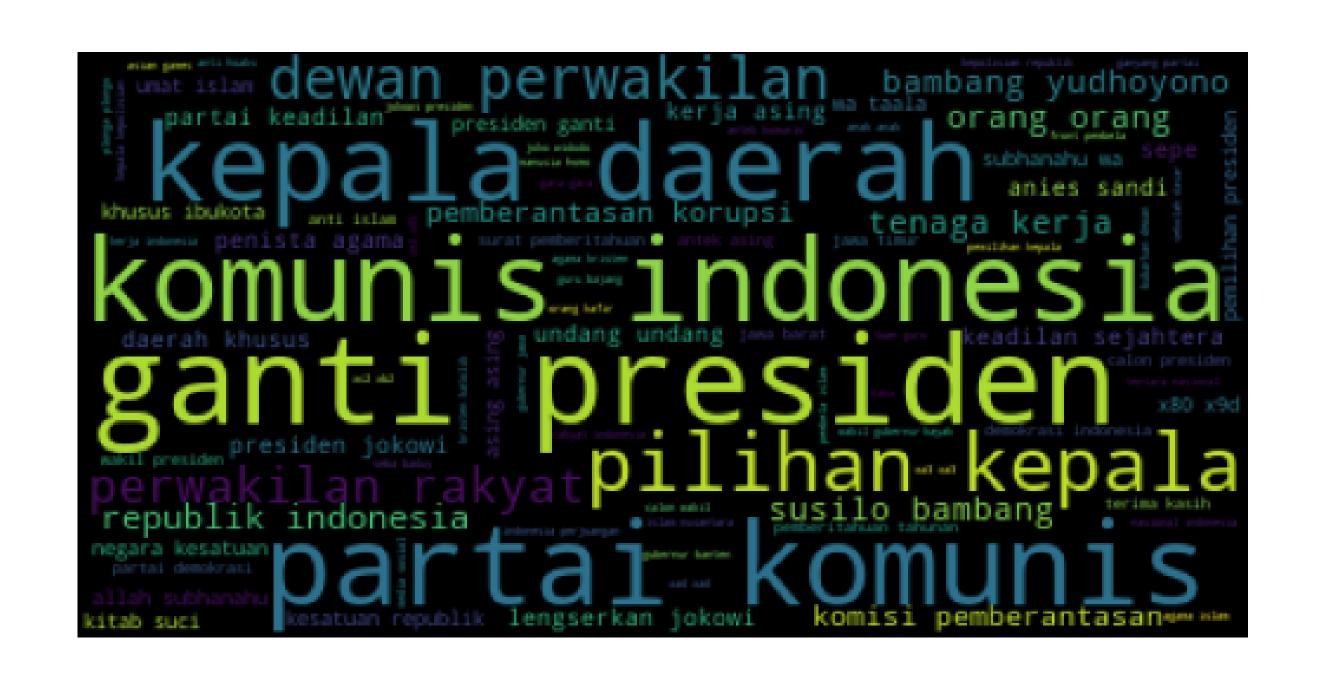
agama: 3835

asing: 3690

komunis: 3645

rakyat: 3445

cina: 3115



TOP 10 WORDS ABUSIVE LABEL

orang: 4760

indonesia: 4250

asing: 3690

komunis: 3645

cebong: 2750

partai: 2345

kayak: 2070

kafir: 1865

islam: 1860

banget: 1850



TOP 10 WORDS NON-ABUSIVE LABEL

presiden: 5835

indonesia: 4110

jokowi: 4015

agama: 2800

islam: 2570

daerah: 2410

orang: 2365

rakyat: 2255

gubernur: 2210

kepala: 1880











Data yang diperoleh panjang karakter ataupun jumlah kata dengan label abusive lebih kecil dibandingkan nonabusive, namun jumlah tweet yang ber-label abusive lebih banyak dibandingkan dengan non-abusive



Hasil identifikasi kata-kata abusive dapat menginikasikan suatu topik yang sedang di bahas baik dalam koteks perdebatan politik atau diskusi yang sedang hangat atapupun kontroversi



Hal ini dapat kemudian ditindak lanjuti untuk penyusunan kebijakan pada paltform twitter atau X seperti deteksi yang otomatis terhadap kata-kata Abusive

Thank you.

