

GOLD CHALLENGE

Analisis Deskriptif Sentimen Tweet
Berbahasa Indonesia Berdasarkan
Kategorisasi Abusif & Alay

Citra Dwikasari (DSC_Wave 7)



Table of Contents

- 01 Latar Belakang
- 02 Metode Penelitian
- 03 Hasil
- 04 Kesimpulan
- 05 Demo Aplikasi

1. Latar Belakang

Berikut ini merupakan latar belakang dilakukannya analisis deskriptif mengenai sentimen pengguna Twitter berbahasa Indonesia berdasarkan kategorisasi Abusif dan Alay.

Indonesia sebagai Peringkat 4 di Dunia

Menurut data StatCounter, pada bulan Februari 2022, Indonesia menempati peringkat ke-4 sebagai pengguna Twitter terbesar di dunia dengan pangsa pasar sebesar 5,58% dari total pengguna Twitter global.

Sentimen Abusif & Alay Berdampak Buruk

Dampak buruk akibat Tweet bersentimen Abusif dan Alay, yaitu 1) Meningkatkan risiko cyberbullying dan online harassment, 2) Merusak citra diri sendiri dan menurunkan kredibilitas, dan 3) Merusak hubungan sosial dalam komunitas online.

Urgensi Analisis Tweet Indonesia yang Bersentimen Abusif & Alay

Untuk melihat sejauh mana sentimen Abusif dan Alay yang dilakukan oleh Pengguna Twitter di Indonesia, maka dilakukan analisis deskriptif terhadap Tweet berbahasa Indonesia sehingga Stakeholder dapat melakukan tindakan preventif.

2. Metode Penelitian

Sumber data penelitian yang akan dianalisis secara deskriptif diperoleh dari website Kaggle yakni
<https://www.kaggle.com/datasets/ilhamfp31-indonesian-abusive-and-hate-speech-twitter-text>



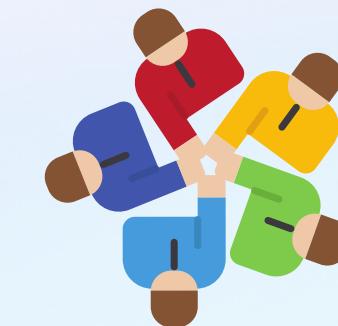
Data.csv

berisi Data Twitter berbahasa Indonesia (13169 baris × 13 kolom), selanjutnya akan dianalisis.



abusive.csv

berisi daftar kata abusive (126 kata), selanjutnya digunakan sebagai referensi analisis.



new_kamusalay.csv

berisi daftar kata alay (15166 kata), selanjutnya digunakan sebagai referensi analisis.

Metode Analisis

Pada tahap pertama, Data.csv kemudian akan dilakukan proses cleansing menggunakan code disamping.

```
# Inisialisasi fungsi
# rules 1: membuat semua huruf kecil
def lowercase(text):
    return text.lower()

# rules 2: menghilangkan karakter-karakter yang tidak diperlukan
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
    return text

# rules 3 : mengganti kata-kata yang ada di kamus pertama
def remove_abusive(text):
    """review: disini pakai huruf kapital : dict1.ABUSIVE
    karena di dalam data abusvie.csv, kolom nya pakai huruf kapital
    """
    text = ' '.join(['' if word in dict1.ABUSIVE.values else word for word in text.split(' ')])
    text = re.sub(' +', ' ', text) # Remove extra spaces
    text = text.strip()
    return text

# rules 4: mengganti kata-kata yang ada di kamus kedua
alay_dict_map = dict(zip(dict2['original'], dict2['replacement']))
def normalize_alay(text):
    return ' '.join([alay_dict_map[word] if word in alay_dict_map else word for word in text.split(' ')])

# rules 5: menjadikan seluruh fungsi dalam 1 fungsi
def cleansing(text):
    text = lowercase(text) # rules 1
    text = remove_unnecessary_char(text) # rules 2
    text = remove_abusive(text) # rules 3
    text = normalize_alay(text) # rules 4
    return text

# end inisialisasi fungsi
```

Metode Analisis

Pada tahap kedua, data akan diterapkan metode Statistika, lalu pada tahap terakhir akan dilakukan Visualisasi.



STATISTIKA

Beberapa metode statistika digunakan untuk mengidentifikasi:

1. Shape data mentah yang akan dianalisis.
2. Jumlah Sentimen Abusive yang ditemukan pada Tweet.
3. Jumlah Kata Alay yang digunakan pada Tweet.



VISUALISASI

Univariat Analisi digunakan untuk melihat visualisasi :

1. Pie Chart merepresentasikan persentase penggunaan kata Abusive dan Non Abusive pada Tweet.
2. Word Cloud merepresentasikan penggunaan kata terbanyak pada Tweet.

Shape Data

Data.csv yang akan dianalisis berbentuk 13.169 Baris dan 13 Kolom.

```
[10] DATASET.shape
```

```
(13169, 13)
```

Jumlah Sentimen Abusive

Terdapat 6280 Tweet yang bersentimen Abusive

```
DATASET['label'].value_counts()
```

```
Non Abusive    6889
```

```
Abusive       6280
```

```
Name: label, dtype: int64
```

3. Hasil

Jumlah Kata Alay

Terdapat 4600 kata alay pada Tweet

```
▶ Alay = set(DATASET['Clean_Tweet'].str.split().explode())
kamus_words = set(Dict2['Replacement'])

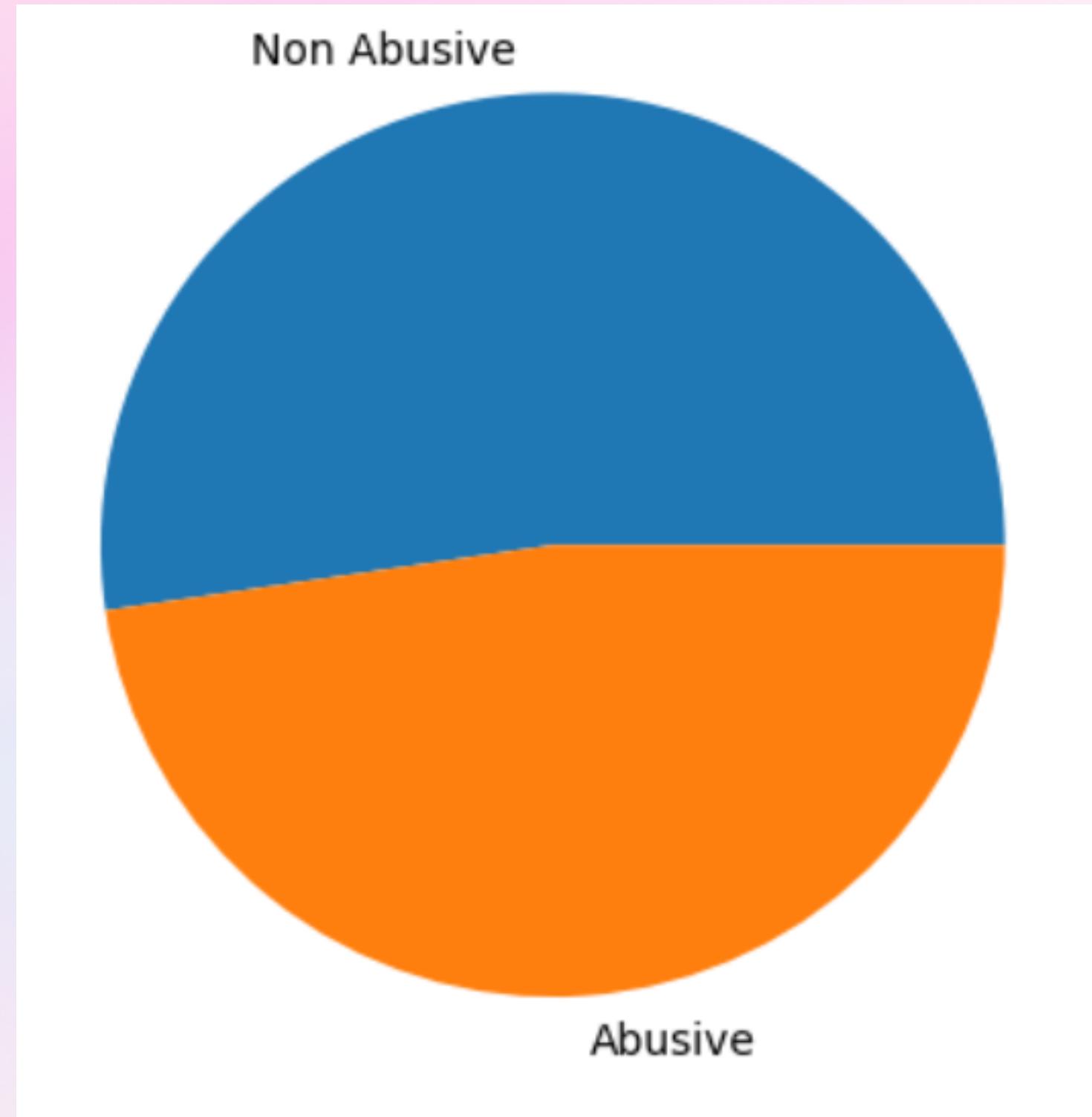
matching_words = Alay.intersection(kamus_words)

print('Number of Alay Words:', len(matching_words))

⇨ Number of Alay Words: 4600
```

Univariat

Berikut ini merupakan visualisasi data yang menggambarkan persentase penggunaan kata Abusive dan Non Abusive pada Tweet.



Words Mapping

Berikut ini adalah kata - kata yang paling sering digunakan oleh Pengguna Twitter berbahasa Indonesia



Hasil Cleansing dapat dilihat pada kolom Clean_Tweet

	Tweet	Clean_Tweet	label	HS	A
141	RT USER: Admin tolol di partai tolol\nPartai T...	admin di pak ai tololnpa ai untuk orang tolol	Abusive	1	
426	USER USER Matamu picek, cukk..lnOra delok pres...	matamu picek cukknora lihat prestasi pak jokow...	Abusive	1	
749	USER USER Tolol...apa hubungannya dgn rumahtan...	tololapa hubungannya dengan rumah tangga rus sa...	Non Abusive	1	
841	USER USER Apan jir? dahak?KWKWKW TOLOL'	apaan jirdahakkwkwl tolol	Non Abusive	1	
951	USER USER USER USER USER USER USER U...	daripada kamu monyet tolol	Non Abusive	1	
...
12009	USER In anak sama bapak sama" tukang KIBUL,dis...	in anak sama bapak sama tukang kibul disekolah...	Non Abusive	1	



4. Kesimpulan

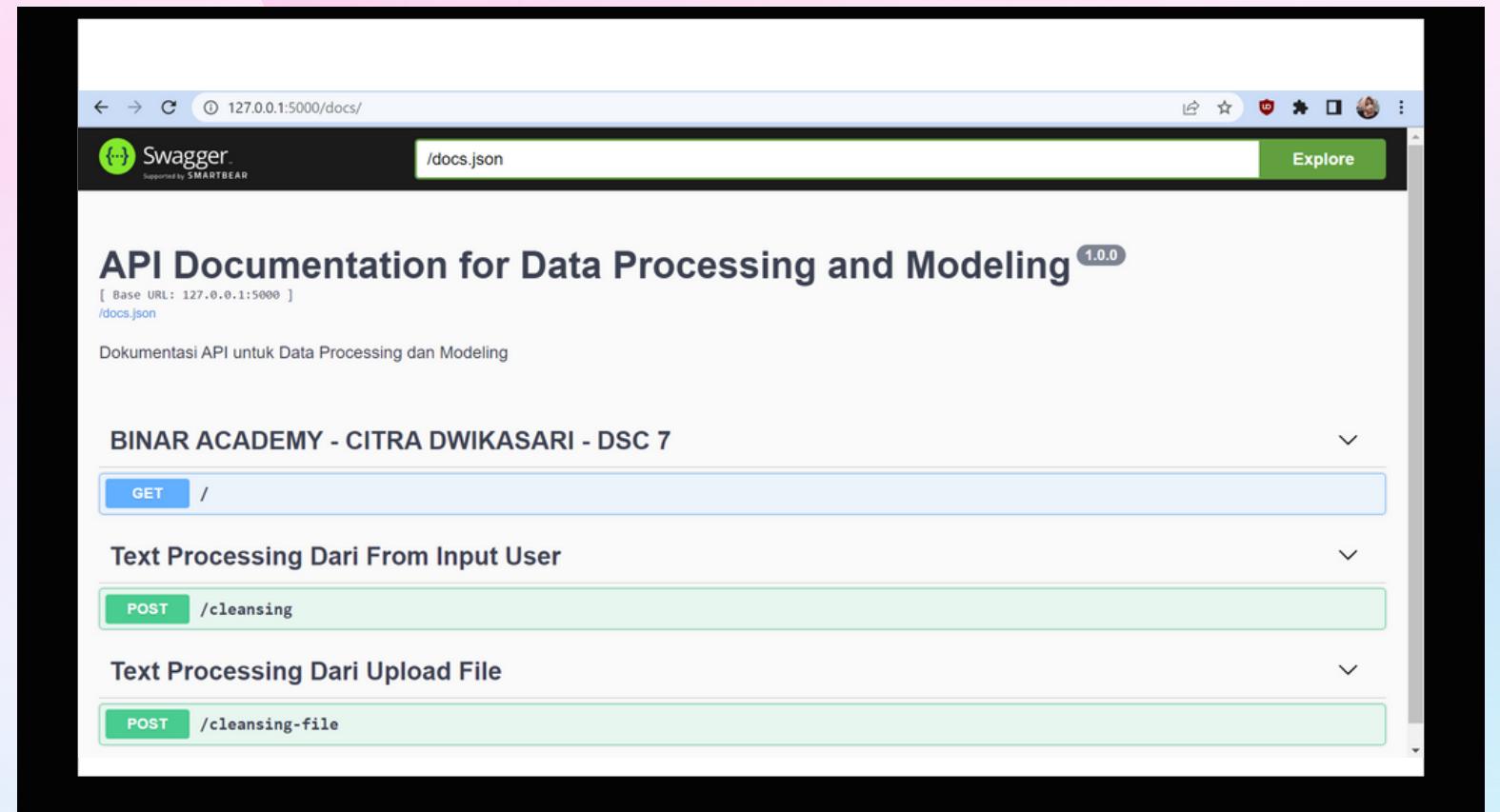
terdapat kalimat toxic di twitter sebanyak 7309 dan kalimat non-toxic sebanyak 5860.

```
print("Toxic shape: ", DATASET[(DATASET['HS'] == 1) | (DATASET['Abusive'] == 1)].shape)
print("Non-toxic shape: ", DATASET[(DATASET['HS'] == 0) & (DATASET['Abusive'] == 0)].shape)

Toxic shape: (7309, 13)
Non-toxic shape: (5860, 13)
```

5. Demo Aplikasi

Aplikasi Cleansing untuk membersihkan kata Alay dan Abusif dapat diakses melalui Github https://github.com/citradwikaSARI/BINAR_DSC_Gold-Challenge_Citra-Dwikasari .





**THANK
YOU**