Analisis Deskriptif & Visualisasi Data Hate Speech Tweet di Indonesia

Dessen

(Binar Academy)

Outline

1. Problem Definition
2. Data Preparation
3. Analyze Data
4. Summary

Problem Definition

BINAR

Latar Belakang

Ujaran kebencian atau hate speech banyak ditemukan di media sosial khususnya di Indonesia. Dalam 100 hari terakhir, Twitter menjadi media sosial yang paling banyak mendapat teguran dari polisi siber terkait hate speech yaitu sebanyak 215 akun.

BINAR

Tujuan Penelitian

Dari fakta di atas, peneliti ingin mengetahui lebih dalam target, kategori, dan level hate speech yang dilakukan melalui Twitter oleh masyarakat Indonesia.

Oleh karena itu, penelitian ini bertujuan untuk menganalisis jumlah karakter, jumlah kata, target, kategori & level hate speech pada tweet masyarakat Indonesia serta kata yang sering digunakan dalam setiap klasifikasi. Peneliti berharap analisis yang dilakukan dapat memberi pengetahuan baru bagi pembaca.

BINAR

Rumusan Masalah

1. Berapa rentang jumlah karakter dan jumlah kata tweet hate speech di Indonesia?
2. Siapa target hate speech yang paling banyak?
3. Kategori hate speech apa yang paling banyak?
4. Level hate speech apa yang paling banyak?
5. Apa kata yang paling sering muncul pada hate speech setiap klasifikasi dengan persentase terbesar?

BINAR

Sumber Data

Sumber data pada penelitian ini merupakan data sekunder dimana diambil peneliti dari situs open source Kaggle yang dapat dilihat melalui link berikut. Data yang dianalisis adalah data tweet dalam bahasa indonesia dari platform social media Twitter.

Metode Penelitian

Metode analisis yang dipakai peneliti adalah Descriptive Analytics. Descriptive Analytics bertujuan mengetahui dan mempelajari kondisi dan pola dari suatu data menggunakan berbagai operasi matematika dan statistika.

Jenis Exploratory Data Analysis (EDA) yang digunakan adalah 1 variabel (Univariate Analysis). Proses analisis yang digunakan adalah metode statistik deskriptif dan visualisasi data. Deskriptif statistik digunakan untuk mengetahui persebaran data. Visualisasi data dipakai untuk memudahkan pembaca memahami tren data tweet hate speech di Indonesia.

Data Preparation

BINAR

Load Data

Untuk load data dan membuat data frame menggunakan library pandas

```
1 import pandas as pd
2 import re
3 import seaborn as sns
4 import matplotlib.pyplot as plt
```

```
1 df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/data.csv', encoding='latin1')
2 df
```

BINAR

Cek Head & Tail

Cek data 5 teratas dan bawah untuk melakukan eksplorasi awal data tweet.

Dari data ini kita pada melihat bahwa kata-kata seperti RT, USER, \n, dan simbol yang tidak penting harus di hapus agar pada saat analisis data tidak bias

```
1 df.head()
```

	Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1	0	0	0	0	0	1	1	0	0
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0	0	0	0	0	0	0	0	0	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0	0	0	0	0	0	0	0	0	0
3	USER USER AKU ITU AKU\n\nKU TAU MATAMU SIPIT T...	0	0	0	0	0	0	0	0	0	0	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1	0	1	1	0	0	0	0	0	1	0

```
1 df.tail()
```

	Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
13164	USER jangan asal ngomong ndasmu. congor lu yg ...	1	1	1	0	0	0	1	0	0	1	0	0
13165	USER Kasur mana enak kunyuk'	0	1	0	0	0	0	0	0	0	0	0	0
13166	USER Hati hati bisu :(.g\n\nlagi bosan huft \...	0	0	0	0	0	0	0	0	0	0	0	0
13167	USER USER USER USER Bom yang real mudah terdet...	0	0	0	0	0	0	0	0	0	0	0	0
13168	USER Mana situ ngasih(": itu cuma foto ya kuti...	1	1	1	0	0	0	0	0	1	1	0	0

BINAR

Cek Shape

Cek jumlah kolom dan baris pada data dimana ada 13 kolom dan 13169 baris

```
1 df.shape
```

```
(13169, 13)
```

Cek Data Duplicates & Missing Value

```
1 df.duplicated().sum()
```

```
125
```

```
1 df = df.drop_duplicates()
2 df.duplicated().sum()
```

```
0
```

Duplicate Data

Terdapat 125 data duplikat dan sudah dihapus pada dataframe

```
1 df.isnull().sum()
```

```
Tweet           0
HS              0
Abusive         0
HS_Individual   0
HS_Group        0
HS_Religion     0
HS_Race         0
HS_Physical     0
HS_Gender       0
HS_Other        0
HS_Weak         0
HS_Moderate     0
HS_Strong       0
```

Missing Value

Tidak terdapat missing value pada dataframe

BINAR

Membuat Kolom Baru Total Karakter dan Total Kata

Kolom baru dibuat untuk menghitung jumlah karakter dan jumlah kata pada setiap tweet

```
1 df['total_char'] = df.Tweet.apply(len)
2 df['total_word'] = df.Tweet.apply(lambda sent: len(sent.split()))
```

Data Cleansing

Library regex digunakan untuk melakukan data cleansing pada tweet karena regex adalah library spesialis untuk melakukan substitusi / menghapus teks pada tweet yang tidak diperlukan dalam analisis

Input

```
def data_cleaning(text):
    text = re.sub('USER', '', text) #Remove USER
    text = re.sub('RT', '', text) #Remove RT
    text = re.sub('URL', '', text) #Remove URL
    text = re.sub(r'\\n+', '', text) #Remove \n
    text = re.sub(r'https\S+','', text) #Remove https
    text = re.sub(r'\\x[A-Za-z0-9./]+', '', text) #Remove \x96 etc
    text = re.sub('#[A-Za-z0-9./]+', '', text) #Remove hastag
    text = re.sub('  +', '', text) #Remove extra space
    return text.lower() #Lowercase text

def data_cleaning2(text):
    text = re.sub('[^0-9a-zA-Z]+', ' ', text) #Remove non alpha numeric
    return text

def preprocess(text):
    text = data_cleaning(text)
    text = data_cleaning2(text)
    return text
```

Output

	Tweet
0	disaat semua cowok berusaha melacak perhatian...
1	siapa yang telat ngasih tau elu edan sarap gu...
2	41 kadang aku berfikir kenapa aku tetap percay...
3	aku itu akuku tau matamu sipit tapi diliat dar...
4	kaum cebong kapir udah keliatan dongoknya dari...

	Tweet
13164	jangan asal ngomong ndasmu congor lu yg sekat...
13165	kasur mana enak kunyuk
13166	hati hati bisu glagi bosan huft
13167	bom yang real mudah terdeteksi bom yang terkub...
13168	mana situ ngasih itu cuma foto ya kutil onta

Analyze Data

```
1 df.mean()
total_char        97.601426
total_word        15.729454
```

```
1 df.median()
total_char        84.0
total_word        13.0
```

```
1 df['total_char'].mode()
0    36
dtype: int64
```

```
1 df['total_word'].mode()
0    7
dtype: int64
```

Mean

Rata-rata panjang karakter adalah 97.6 dan rata-rata panjang kata adalah 15.73.

Median

Nilai tengah panjang karakter adalah 84 dan nilai tengah panjang kata adalah 13.

Mode

Frekuensi karakter yang paling sering muncul yaitu 36 dan frekuensi kata yang paling sering muncul totalnya 7.

BINAR

```
[25]  1 range_total_char = df.total_char.max()-df.total_char.min()
      2 range_total_char

      280

[26]  1 range_total_word = df.total_word.max()-df.total_word.min()
      2 range_total_word

      52
```

Range

Range dari total word perbedaannya hanya sebesar 280 kata, sedangkan di total char punya perbedaan sebesar 52 karakter.

```
1 df.var()
total_char      4102.606365
total_word       106.554186
```

```
1 df.mean()
total_char      97.601426
total_word      15.729454
```

Variance

Variance dari total char sebesar 4102.6 menjauhi nilai mean. Karena nilai mean total char yaitu 97.6 karakter, maka nilai variance lebih dari nilai mean.

Sedangkan untuk variance dari total word sebesar 106.55 menjauhi nilai mean. Karena nilai mean total_word yaitu 15.73 kata, maka nilai variance lebih besar dari mean.

```
1 df.std()
total_char      64.051591
total_word      10.322509
```

```
1 df.mean()
total_char      97.601426
total_word      15.729454
```

Standard Deviation

Nilai standard deviation-nya total char yaitu sebesar 64.05. Artinya standard deviation total char lebih kecil dari nilai mean total char yaitu 97.6 karakter.

Begitupun untuk nilai standard deviation dari total word sebesar 10.32 lebih kecil dari nilai mean total word yaitu 15.73.

```python
p0 = df.total_word.min()
p100 = df.total_word.max()
q1=df.total_word.quantile(0.25)
q2=df.total_word.quantile(0.5)
q3=df.total_word.quantile(0.75)
iqr = q3-q1
lower_limit=q1-1.5*iqr
upper_limit=q3+1.5*iqr
print("Batas bawah 'total_word':", lower_limit)
print("Nilai minimum",p0)
if lower_limit <p0:
  print("Tidak ada outlier dari sisi batas bawah")
else:
  print("Ada outlier dari sisi batas bawah")
print("Batas atas 'total_word':",upper_limit)
print("Nilai maksimum",p100)

if upper_limit>p100:
  print("Tidak ada outlier dari sisi batas atas")
else:
  print("Ada outlier dari sisi batas atas")
```

```
Batas bawah 'total_word': -11.5
Nilai minimum 0
Tidak ada outlier dari sisi batas bawah
Batas atas 'total_word': 40.5
Nilai maksimum 52
Ada outlier dari sisi batas atas
```

```python
p0 = df.total_char.min()
p100 = df.total_char.max()
q1=df.total_char.quantile(0.25)
q2=df.total_char.quantile(0.5)
q3=df.total_char.quantile(0.75)
iqr = q3-q1
lower_limit=q1-1.5*iqr
upper_limit=q3+1.5*iqr
print("Batas bawah 'total_char':", lower_limit)
print("Nilai minimum",p0)
if lower_limit <p0:
  print("Tidak ada outlier dari sisi batas bawah")
else:
  print("Ada outlier dari sisi batas bawah")
print("Batas atas 'total_char':",upper_limit)
print("Nilai maksimum",p100)

if upper_limit>p100:
  print("Tidak ada outlier dari sisi batas atas")
else:
  print("Ada outlier dari sisi batas atas")
```

```
Batas bawah 'total_char': -80.5
Nilai minimum 0
Tidak ada outlier dari sisi batas bawah
Batas atas 'total_char': 259.5
Nilai maksimum 280
Ada outlier dari sisi batas atas
```

Outlier

Total word dan total char tidak ditemui adanya outlier dari sisi batas bawah.

Tetapi ada outlier dari sisi batas atas total word dan total char.

```
1 df.skew()
total_char      0.792862
total_word      0.861397
```

```
1 df.kurtosis()
total_char      -0.271545
total_word      -0.015995
```

Skew

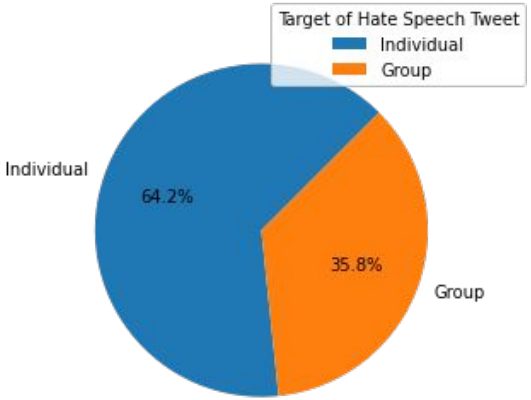
Nilai skewness di total char 0.79 dan total word 0.86 nilainya positif.

Hal ini karena nilai skewness kedua variabel/kolom > 0.

Kurtosis

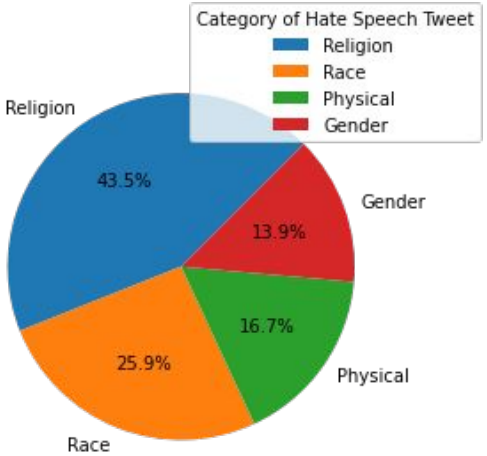
Nilai kurtosis total char dan total word bernilai kurang dari 3.

Artinya punya sifat playkurtik yaitu cenderung menghasilkan lebih sedikit nilai outlier.



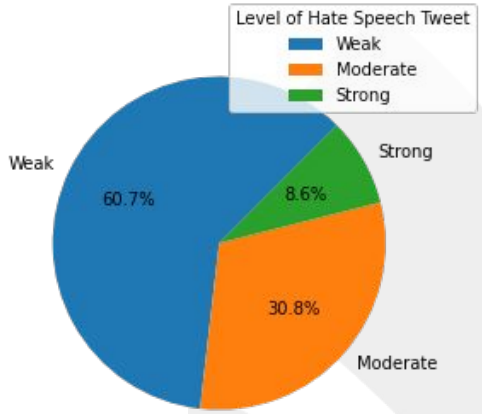
Target

64.2% tweet ditargetkan kepada individual dan 35.8% kepada grup.



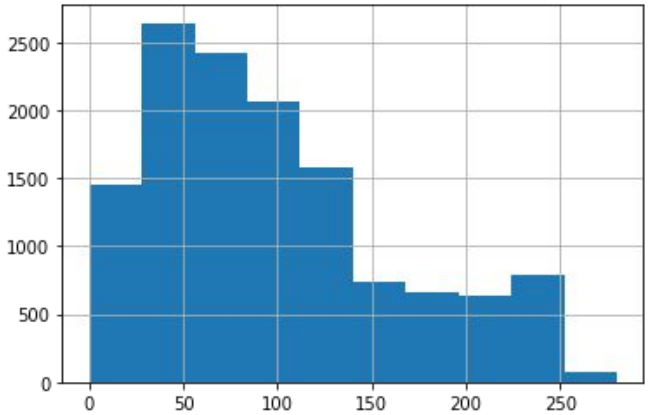
Category

43.5% tweet dikaitkan dengan agama, 25.9% dikaitkan dengan ras, 16.7% dikaitkan dengan fisik, dan 13.9% dikaitkan dengan jenis kelamin



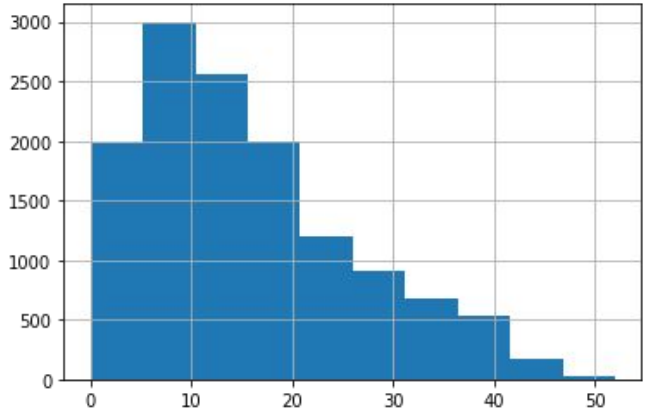
Level

60.7% tweet termasuk level lemah, 30.8% termasuk level sedang, & 8.6% termasuk level kuat



Total Char

Rata-rata panjang karakter dari data teks yang ada adalah sekitar 25-75 karakter.



Total Word

Rata-rata panjang kata dari data teks yang ada adalah sekitar 5-15 kata.

Target : Individual

Untuk individual hate speech kata yang paling sering muncul jika exclude kata hubung adalah "Jokowi", “Prabowo”, & “Ahok”

Category : Religion

Untuk hate speech terkait agama kata yang paling sering muncul jika exclude kata hubung adalah "Islam”, “Muslim”, & “Kafir”

Level : Low

Untuk hate speech yang bersifat weak kata yang paling sering muncul jika exclude kata hubung adalah "Jokowi” & “Cebong”

Summary

Kesimpulan

Berdasarkan analisis yang sudah kita lakukan dapat hasilnya dapat dijabarkan sebagai berikut::

- Dalam Descriptive Statistic menunjukkan data yang peneliti olah memiliki outlier namun tidak terlalu signifikan
- Berdasarkan Univariate Analysis:
    - Dalam visualisasi menunjukkan:
        - Total karakter dan total kata memiliki panjang 25-75 karakter dan 5-15 kata.
        - Target hate speech paling banyak untuk individual
        - Kategori hate speech paling banyak adalah terkait agama
        - Level hate speech paling banyak adalah weak
        - Untuk individual hate speech kata yang paling sering muncul jika exclude kata hubung adalah "Jokowi", “Prabowo”, & “Ahok”
        - Untuk hate speech terkait agama kata yang paling sering muncul jika exclude kata hubung adalah "Islam”, “Muslim”, & “Kafir”
        - Untuk hate speech yang bersifat weak kata yang paling sering muncul jika exclude kata hubung adalah "Jokowi” & “Cebong”
        - Dari kata yang paling sering muncul dapat disimpulkan bahwa hate speech di Indonesia berhubungan dengan Politik