INDONESIAN ABUSIVE AND HATE SPEECH TWITTER

by
Joanes Ferdienand

Pendahuluan

Hate speech atau ujaran kebencian merupakan perilaku yang dilakukan oleh individu maupun kelompok dengan cara memprovokasi atau menghasut individu atau kelompok tertentu yang biasanya menyangkut aspek gender, orientasi seksual, ras, keagamaan, dan lain sebagainya.

Ujaran kebencian umumnya banyak ditemukan di media sosial. Hal ini dapat dilihat dari hasil program Virtual Police yang dibentuk dengan tujuan menegur akun yang dinilai melakukan pelanggaran UU ITE yang berisi ujaran kebencian dan SARA. Sejak pembentukan Virtual Police tersebut dalam rentang 100 hari kerja (23 Februari 2021 – 31 Mei 2021), Twitter menjadi media sosial yang paling banyak mendapat teguran sebanyak 215 akun.

Berdasarkan hal tesebut, dapat diambil rumusan masalah serta tujuan sebagai berikut:

Rumusan Masalah:

1. Seberapa banyak tweet yang mengandung hate speech beserta abusive?
2. Topik apa yang sering menjadi bahan ujaran kebencian?
3. Siapa yang sering ditargetkan dalam ujaran kebencian?

Tujuan:

1. Untuk mengetahui topik-topik sensitif apa saja yang sering menjadi bahan perdebatan di Twitter Indonesia.
2. Sebagai bahan pembelajaran guna membentuk sistem filtrasi teks yang tidak pantas di sosial media.

Metode Penelitian
Deskripsi Data

- Data yang digunakan dalam metode penelitian ini adalah data sekunder yang bersumber dari akun Kaggle Ilham Firdausi.
- Dari file tersebut, diperoleh data sebanyak 13.044 baris dengan 13 kolom yang mana telah dihilangkan data yang terduplikasi sebelumnya.

```
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Tweet           13169 non-null  object
 1   HS              13169 non-null  int64
 2   Abusive         13169 non-null  int64
 3   HS_Individual   13169 non-null  int64
 4   HS_Group        13169 non-null  int64
 5   HS_Religion     13169 non-null  int64
 6   HS_Race         13169 non-null  int64
 7   HS_Physical     13169 non-null  int64
 8   HS_Gender       13169 non-null  int64
 9   HS_Other        13169 non-null  int64
 10  HS_Weak         13169 non-null  int64
 11  HS_Moderate     13169 non-null  int64
 12  HS_Strong       13169 non-null  int64
dtypes: int64(12), object(1)
memory usage: 1.3+ MB
Dataset ini terdiri dari 13169 baris dan 13 kolom
```

Pengecekan adanya duplikasi di dalam dataset

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

```
125
```

Menghilangkan data duplikasi dan pengecekannya kembali

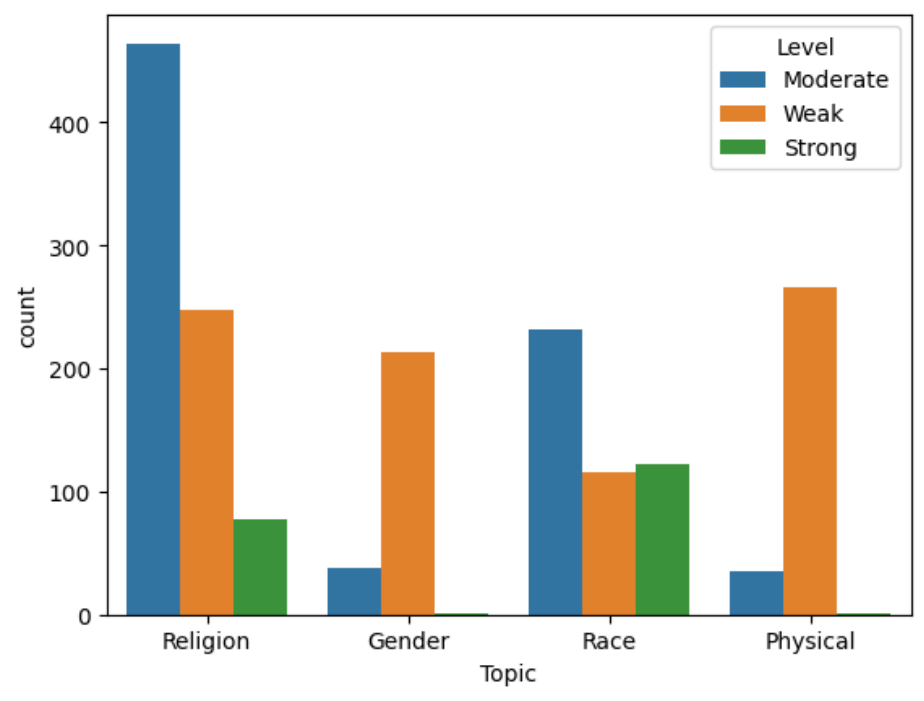
```
df = df.drop_duplicates()
df.duplicated().sum()
```

```
0
```

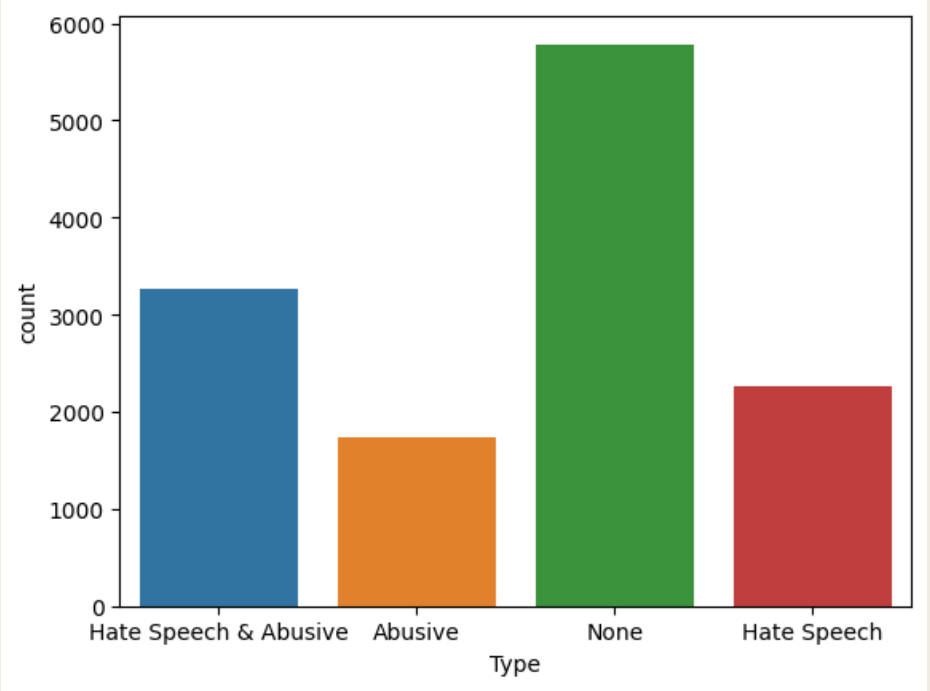
```
print(f"Dataset ini sekarang terdiri dari {df.shape[0]} baris")
```

```
Dataset ini sekarang terdiri dari 13044 baris
```

Metode Penelitian
Visualisasi



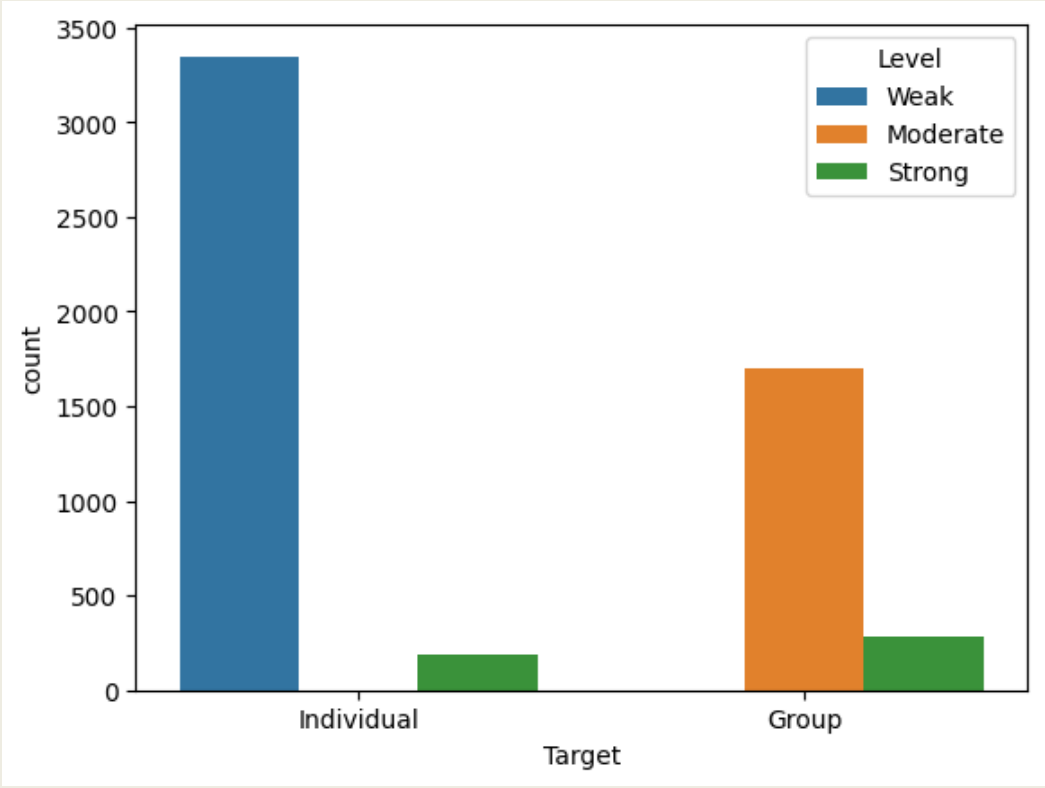
Mengesampingkan topik non-kategorial, topik agama menjadi salah satu topik yang paling sensitif dan rawan menjadi bahan hate speech meskipun masih dalam hate speech level rendah.



Berdasarkan hasil visualisasi di atas, tipe tweet bernon-kategorial memiliki jumlah paling banyak (5.783). Disusul oleh tweet mengandung hate speech dan abusive (3.262). Kemudian hate speech saja (2.256), dan terakhir abusive saja (1.743).

Metode Penelitian Visualisasi

Berdasarkan hasil visualisasi di samping, dapat diambil kesimpulan bahwa hate speech yang ditujukan kepada perseorangan lebih tinggi daripada kelompok meski dengan tingkat hate speech level di kategori rendah.



Metode Penelitian
Text Processing

Di bawah ini, merupakan salah satu proses cleansing data dengan menghilangkan tanda baca yang tidak dibutuhkan serta membuat seluruh kalimatnya dalam keadaan lower case (huruf kecil).
Setelah selesai diproses, kata atau kalimat tersebut akan disimpan di database sebagai histori.

```python
"""
Function untuk membersihkan data text
"""
import re
import pandas as pd

def text_cleansing(text):
    clean_text = ' '.join(text.split())
    clean_text = re.sub(r'[^a-zA-Z0-9\s]', '', text)
    clean_text = clean_text.lower()
    # Bersihkan dengan kamus alay
    # Bersihkan dengan kamus abusive
    return clean_text
```

Request URL

```
http://127.0.0.1:5000/cleansing_form
```

Server response

Code	Details
200	Response body

```
{
  "raw_text": "Reni!!!!!--",
  "clean_text": "reni"
}
```

Response headers

```
connection: close
content-length: 47
content-type: application/json
date: Mon, 19 Jun 2023 18:57:16 GMT
server: Werkzeug/2.2.2 Python/3.10.7
```

```
http://127.0.0.1:5000/show_cleansing_result
```

Server response

Code	Details
200	Response body

```
    "clean_text": "coba 1"
  },
  "1": {
    "raw_text": "coba 1...",
    "clean_text": "coba 1"
  },
  "2": {
    "raw_text": "ahog",
    "clean_text": "ahog"
  },
  "3": {
    "raw_text": "saya ahog",
    "clean_text": "saya ahog"
  },
  "4": {
    "raw_text": "reni123-/'",
    "clean_text": "reni123"
  },
  "5": {
    "raw_text": "reni!!!!!--",
    "clean_text": "reni"
  },
  "6": {
    "raw_text": "Reni!!!!!--",
    "clean_text": "reni"
  }
}
```

Response headers

```
connection: close
content-length: 357
content-type: application/json
date: Mon, 19 Jun 2023 18:58:52 GMT
server: Werkzeug/2.2.2 Python/3.10.7
```

Hasil dan Kesimpulan

Hasil:

1. Dari 13.044 data yang telah disaring, sebanyak 7.261 (55.7%) tweet mengandung hate speech dan atau abusive.
2. Di antara agama, gender, ras, dan fisik, ujaran kebencian dengan topik agama menjadi pembahasan paling sensitif di Twitter Indonesia.
3. Target ujaran kebencian di Indonesia lebih banyak ditujukan kepada perseorangan dibandingkan kelompok.

Kesimpulan:

Berdasarkan hasil analisa, dapat diambil kesimpulan bahwa masyarakat Indonesia cenderung sensitif terkait pembahasan mengenai topik agama dikarenakan tiap individu memiliki pemahaman keagamaannya masing-masing. Hal ini dapat dilihat dari hasil visualisasi sebelumnya. Oleh sebab itu, sistem filtrasi dalam teks tweet di sosial media dapat menjadi salah satu cara, guna mengurangi hal hal ujaran-ujaran kebencian yang dibagikan ke umum.