

Hate Speech Analysis

Trend tweets di Indonesia periode 2019 - 2020

- Almero Dharaka
- Jose Alfred Benaya
- Diva Annisa Febecca
- Thomas Ken Ronaldi



Selama dekade terakhir, jumlah waktu yang dihabiskan pengguna dalam menggunakan media sosial tumbuh sebesar 62% di seluruh dunia. Di Indonesia rata - rata user menghabiskan 8 jam 36 menit sehari untuk bermain media sosial. Kami asumsikan ada kecenderungan pengguna media sosial di Indonesia aktif berkomentar dalam dunia maya.

Dalam aktivitasnya, menurut hasil studi yang dilakukan Digital Civility Index (DCI) tahun 2020 menyatakan bahwa Indonesia berada di peringkat pertama sebagai **warganet paling tidak sopan** se-Asia Tenggara. Angka ini meningkat dari tahun sebelumnya dan mengalami peningkatan kembali di tahun selanjutnya.

Penelitian ini bertujuan untuk membuktikan pernyataan tersebut. Kami melakukan analisis dengan menghitung data tweet. Disini kami melihat banyak tweet yang mengandung kalimat abusive dan non-abusive selama periode waktu 2019 - 2020.

Pemilihan data set pada periode ini bertepatan dengan pasca Pilpres 2019. Saat itu, Joko Widodo kembali terpilih sebagai Presiden Republik Indonesia untuk masa jabatannya yang kedua.

Latar Belakang dan Tujuan



Rumusan Masalah

- Bagaimana bentuk trend tweets Indonesia periode 2019 - 2020?
- Apakah bentuk trend tweets di Indonesia pada periode 2019 - 2020 termasuk *abusive*?



Metode Penelitian

Metode penelitian yang digunakan adalah **Deskriptif** atau biasa disebut dengan **Exploratory Data Analysis (EDA)** yang bertujuan untuk mendeskripsikan data text.

Sesuai rumusan masalah, kami mencoba menjawabnya dengan pengkajian tingkat *abusive* pada data tweets yang telah di cleansing.

Setelah melakukan pengkajian, metode penelitian ini akan menghasilkan kesimpulan berbentuk univariate dan bivariate yang akan kami jelaskan pada bab terakhir.



Tampilan API menggunakan FAST API merupakan hasil dari *challenge* pertama .

The screenshot displays the FastAPI documentation interface, version 0.1.0, generated from an OAS3 schema. The interface is organized into sections: **default**, **Cleansing API**, **Sentiment API**, and **Schemas**.

- default**: Contains a single endpoint: **GET / Index**.
- Cleansing API**: Contains one endpoint: **GET /cleansing Text Cleansing**.
- Sentiment API**: Contains two endpoints:
 - GET /sentiment Sentiment Analytics**
 - POST /sentiment-upload Upload File**
- Schemas**: Shows a schema named **Body_upload_file_sentiment_upload_post**.



Sumber Data

Data yang digunakan adalah Indonesian Abusive and Hate Speech Twitter Text yang bersumber dari kaggle. Data tersebut memiliki beberapa kolom yang kami jabarkan dalam tabel dibawah ini:

Columns	Description
Tweet	Text
HS	Hate speech label
Abusive	Abusive language label
HS_Individual	Hate speech targeted to an individual
HS_Group	Hate speech targeted to a group
HS_Religion	Hate speech related to religion/creed
HS_Race	Hate speech related to race/ethnicity
HS_Physical	Hate speech related to physical/disability
HS_Gender	Hate speech related to gender/sexual orientation
HS_Weak	Weak hate speech
HS_Moderate	Moderate hate speech
HS_Strong	Strong hate speech

Dataset and Attribute

01

Data

	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\nnKU 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

Data Indonesian Abusive and Hate Speech Twitter Text memiliki tweets sebanyak 13.169 yang mana data tersebut dibagi menjadi 2 kategori. Kategori (0) menunjukan bahwa tweet tersebut tidak masuk dalam atribut sedangkan kategori (1) menunjukan bahwa tweet tersebut masuk dalam atribut.

Dataset and Attribute

02

Data Abusive

ABUSIVE	
0	alay
1	ampas
2	buta
3	keparat
4	anjing

Data Abusive memiliki 125 baris dan 1 kolom. Data ini berguna untuk menghapus kata yang mengandung abusive.



Dataset and Attribute

03

Data Stopword

stopword	
0	ada
1	adalah
2	adanya
3	adapun
4	agak

Data Stopword memiliki 758 baris dan 1 kolom. Data ini berguna untuk mengurangi kata yang tidak terpakai.



Dataset and Attribute

04

Kamus Alay

	original	replacement
0	anakjakartaasikasik	anak jakarta asyik asyik
1	pakcikdahtua	pak cik sudah tua
2	pakcikmudalagi	pak cik muda lagi
3	t3tapjokowi	tetap jokowi
4	3x	tiga kali

Kamus Alay memiliki 15.167 baris dan 2 kolom. Data ini mengandung kata alay (tidak baku) dan kata standard (baku).



Data Preparation

- 1 Tidak ada missing value
- 2 Ada data yang terduplikat (146 tweet)
- 3 Ada *outliers* data, persentasenya sedikit (0.65%)



7 Tahap Data Cleansing



1. Lowercase Letter
2. Remove unnecessary character
3. Remove nonalphanumeric character
4. Remove stopwords
5. Remove emoticon byte
6. Normalize "alay" words
7. Stemming words

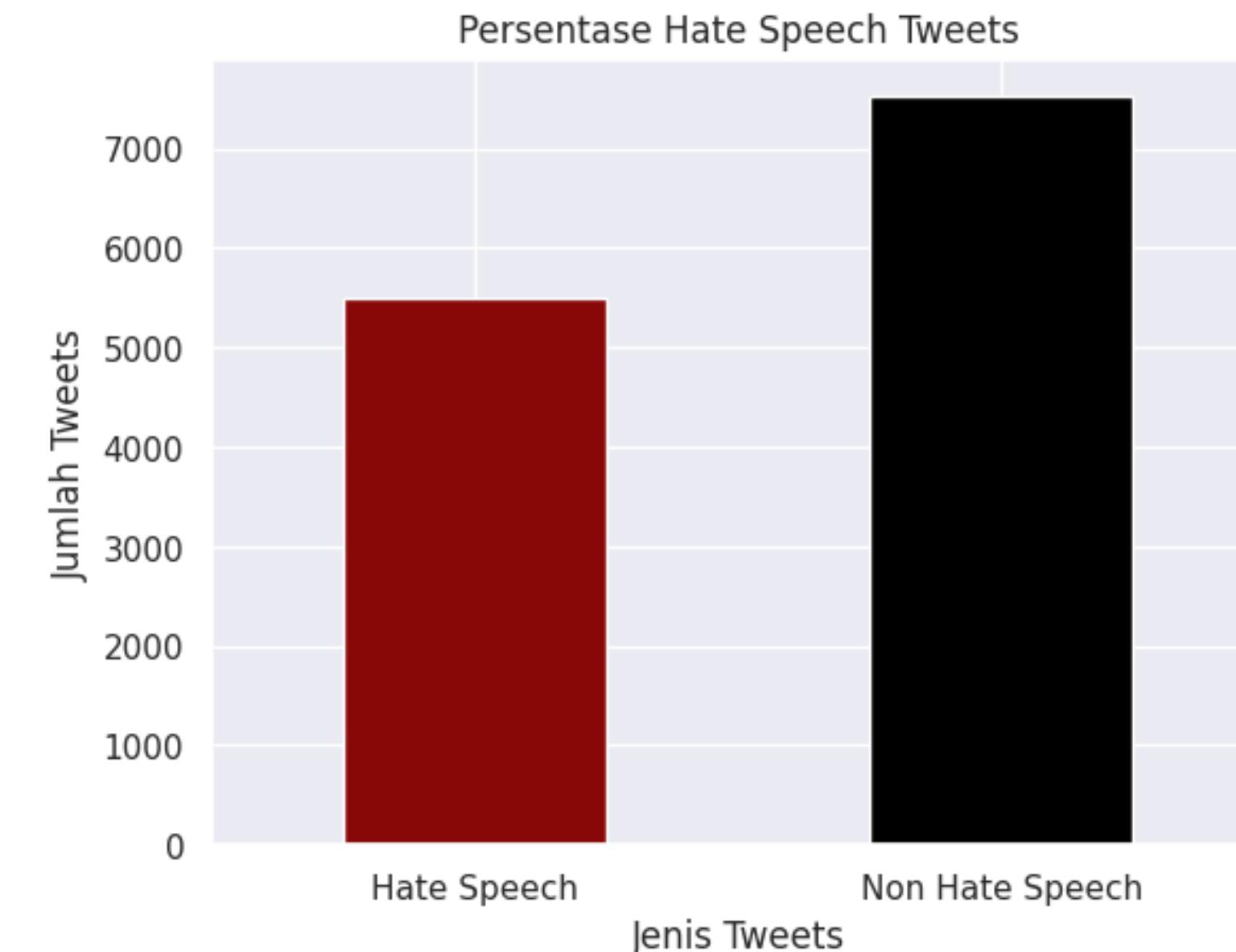
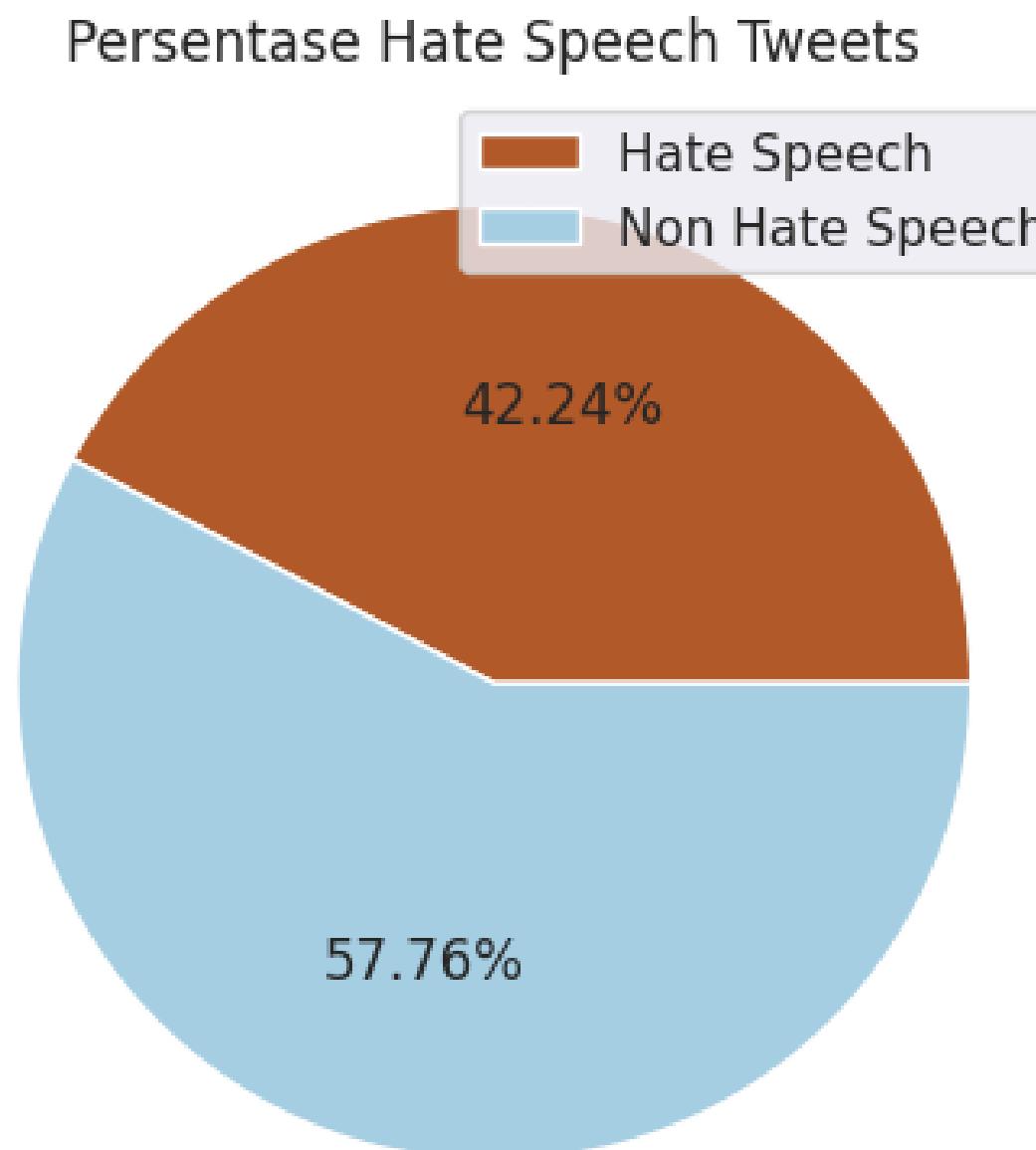


Visualisasi Data

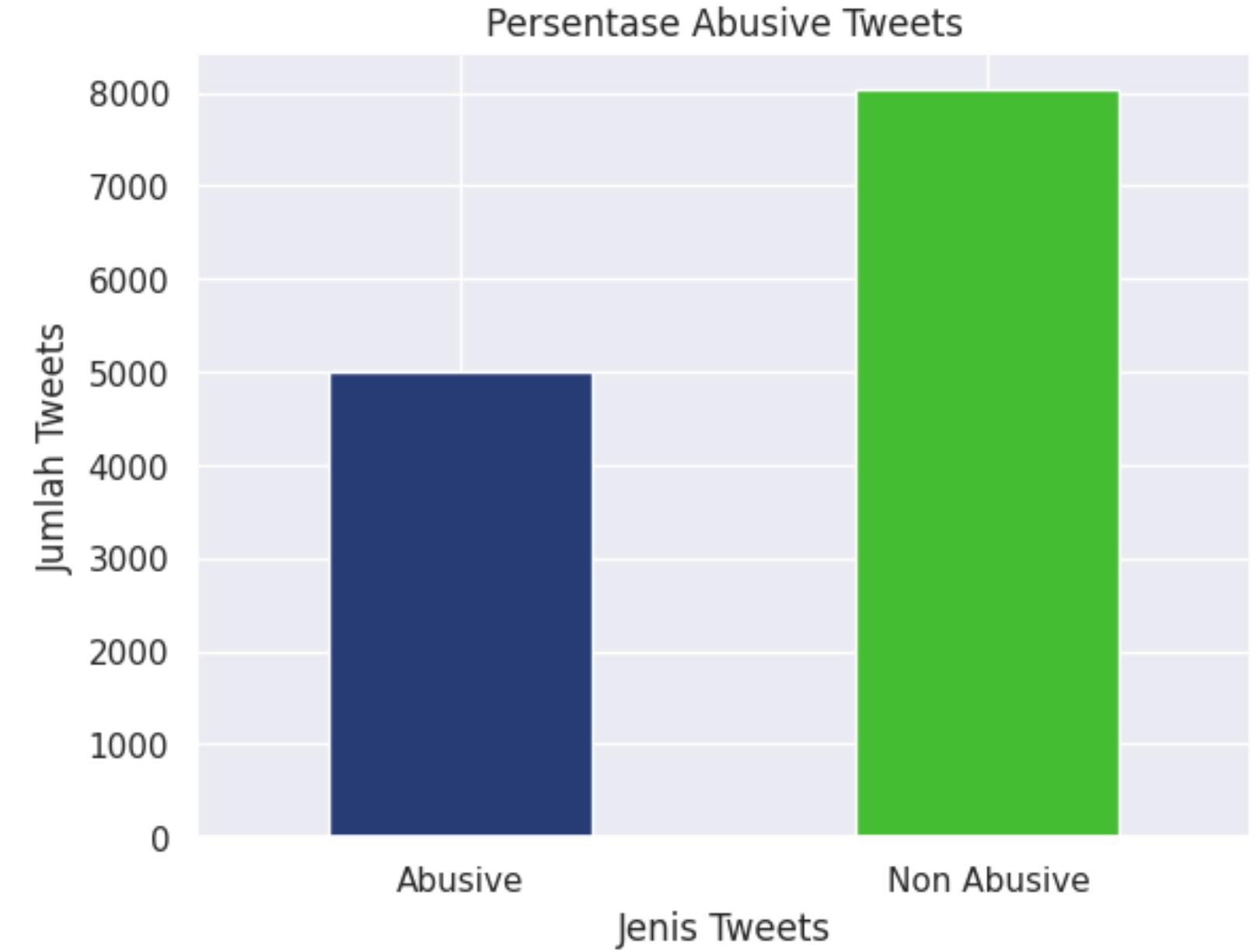
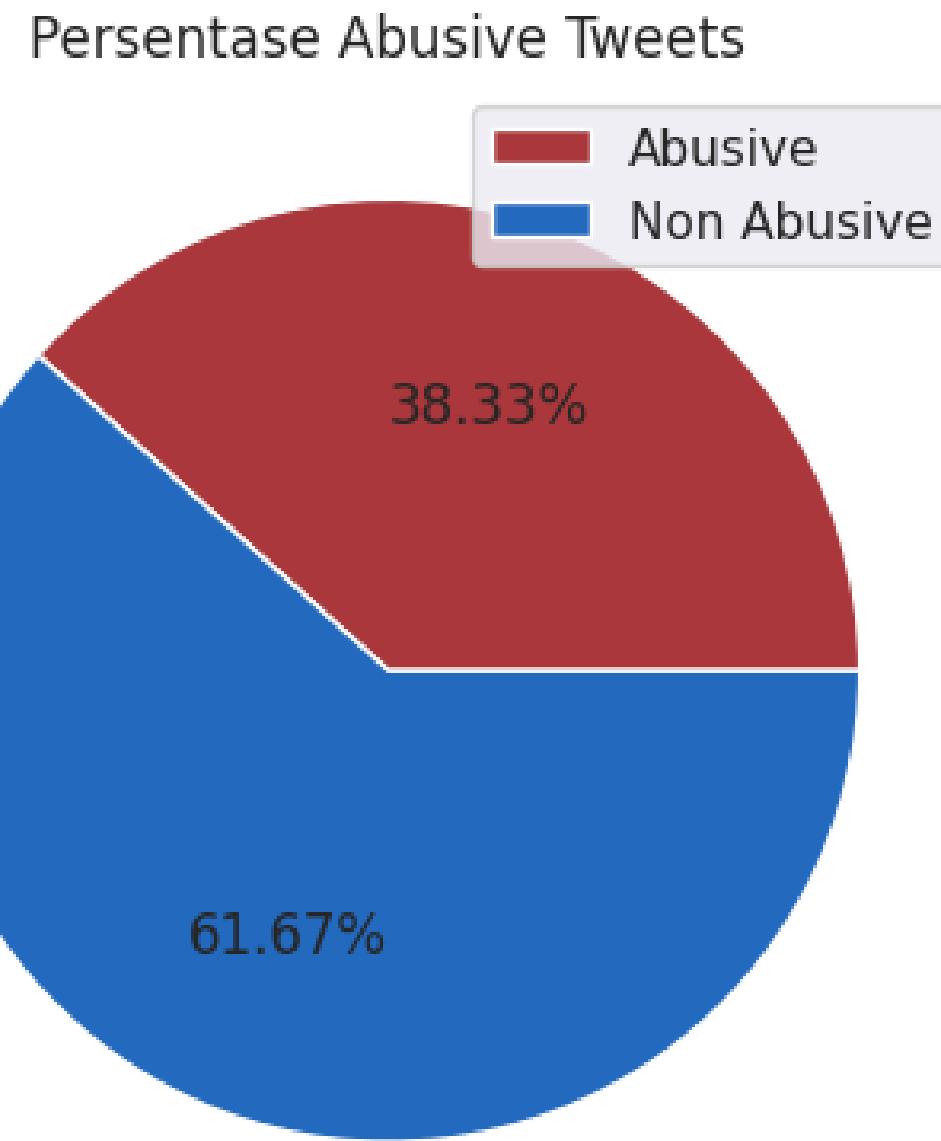
Menggunakan **4 model visualisasi** yang dapat merepresentasikan analisis dari data, yakni :

- **Pie Chart**
- **Histogram**
- **Mosaic**
- **Wordcloud**

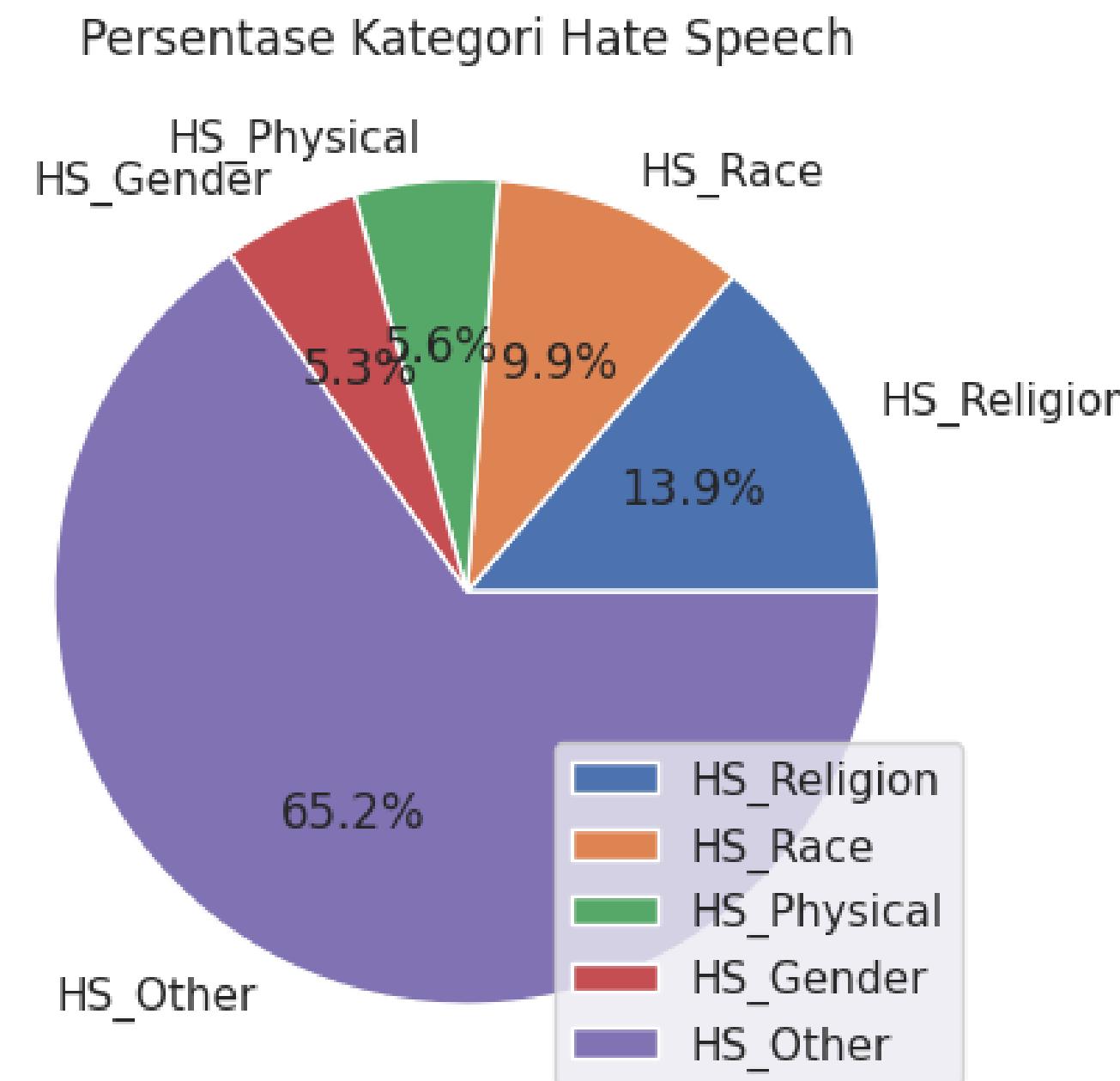
Data didominasi oleh *non-hate speech*, persentasenya sebesar 57,76% atau sebanyak 7522 tweets.



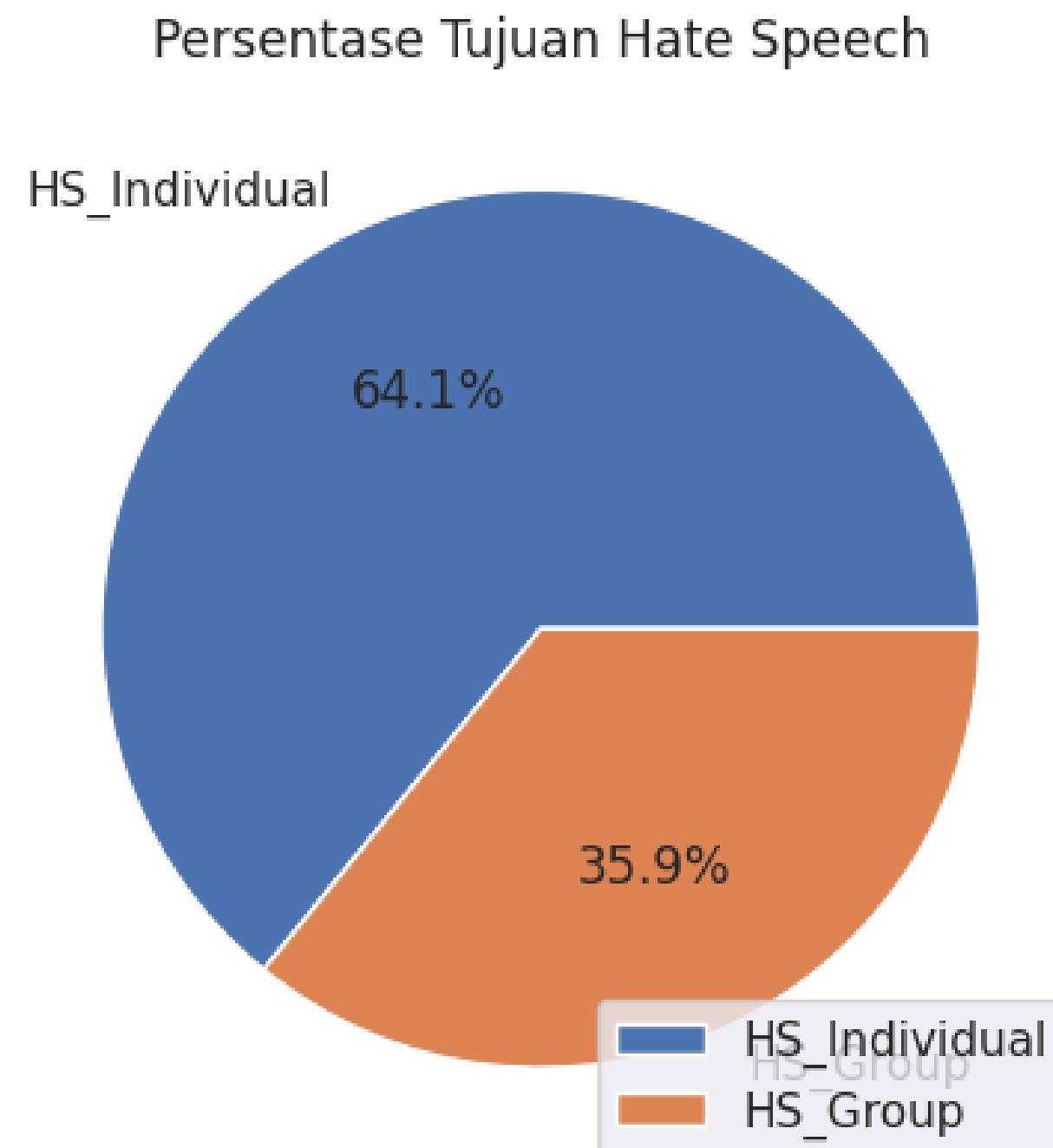
Label data didominasi oleh *non-abusive*, persentasenya sebesar 61.67% atau sebanyak 8031 tweets.



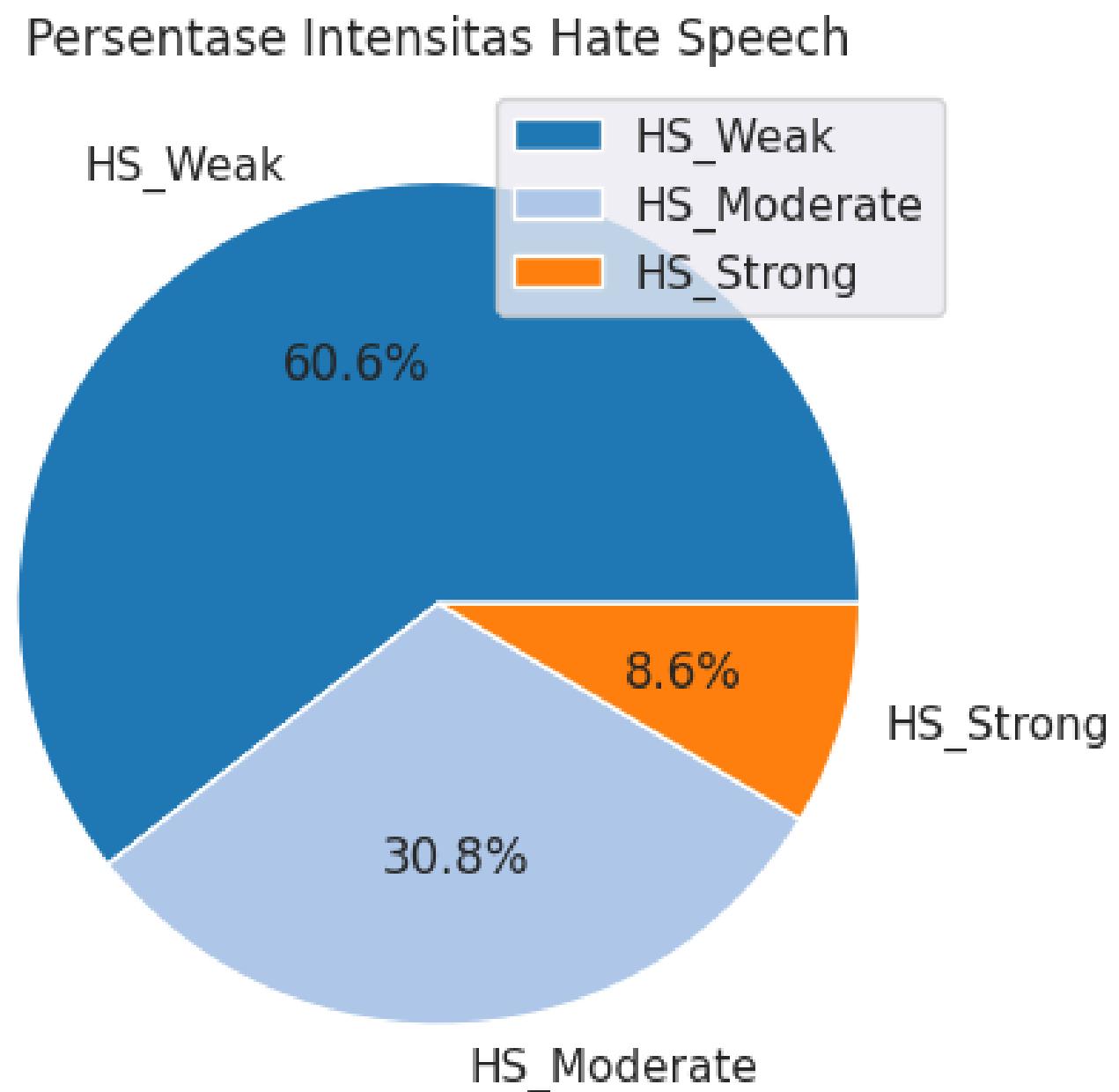
Kategori *Hate Speech Religion* menjadi sasaran utama dalam data ini.



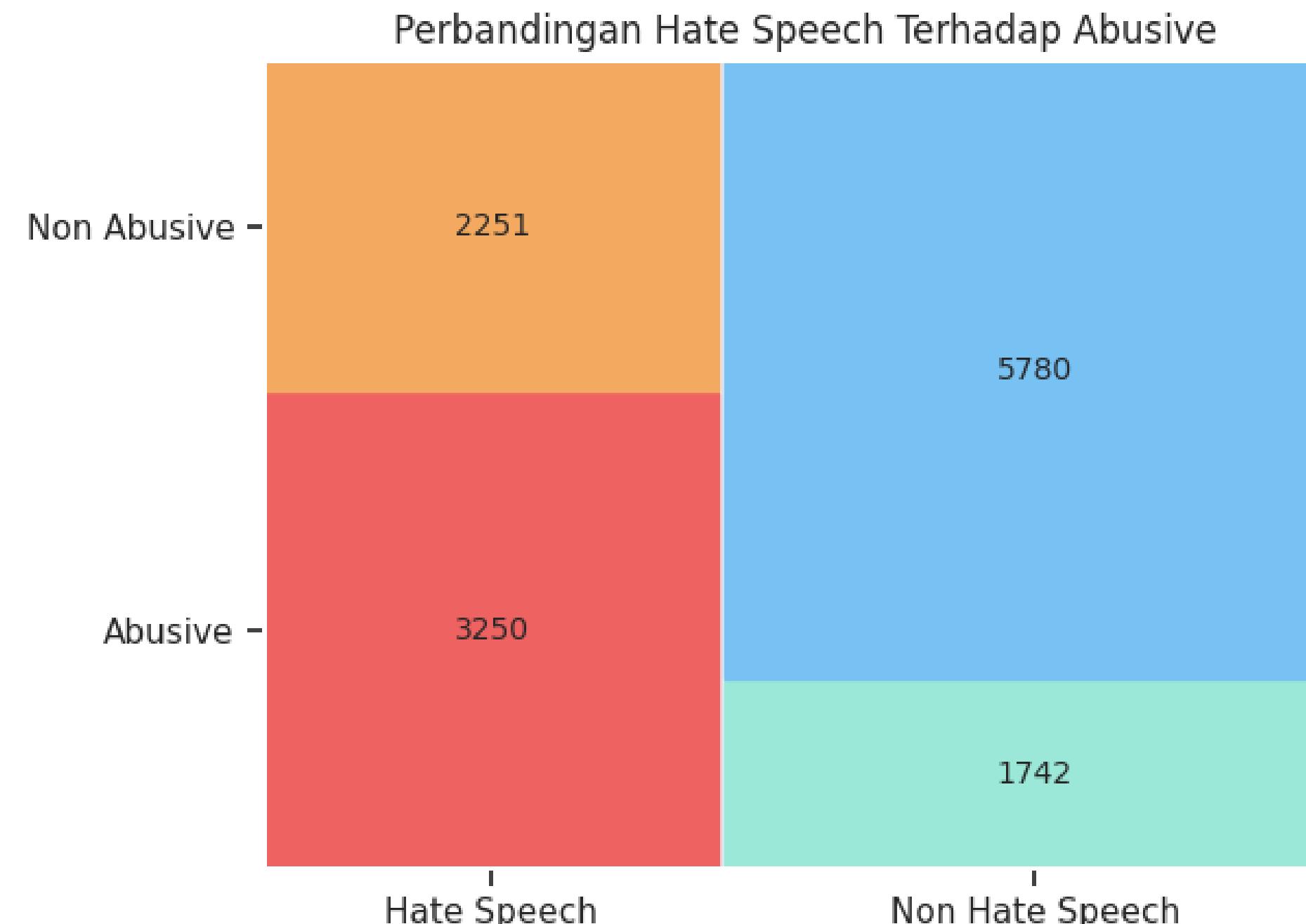
***Hate Speech* banyak ditujukan kepada *Individual* dengan persentase sebesar 64.1%.**



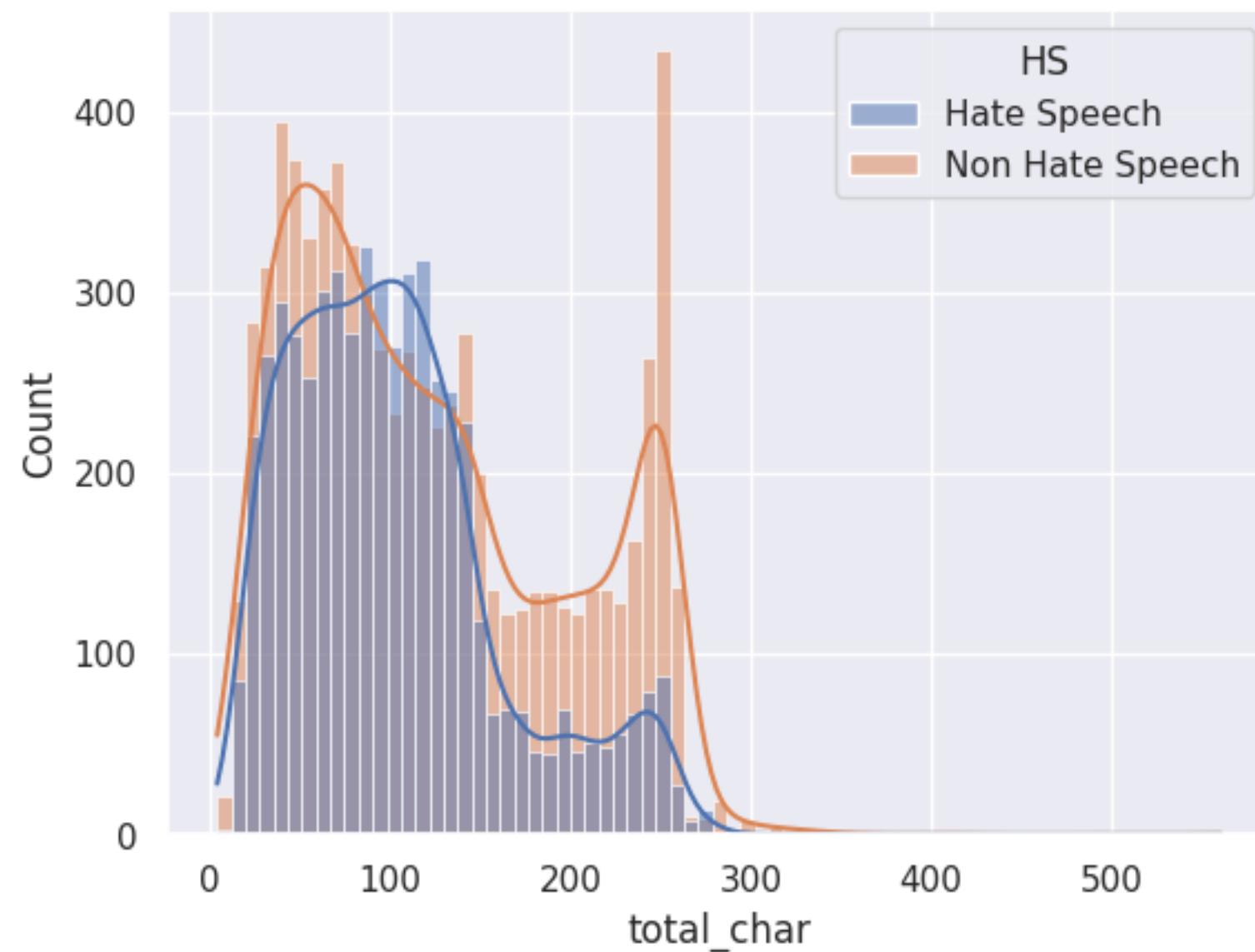
Intensitas *Hate Speech* didominasi oleh *Other* dengan persentase sebesar 42.4%.



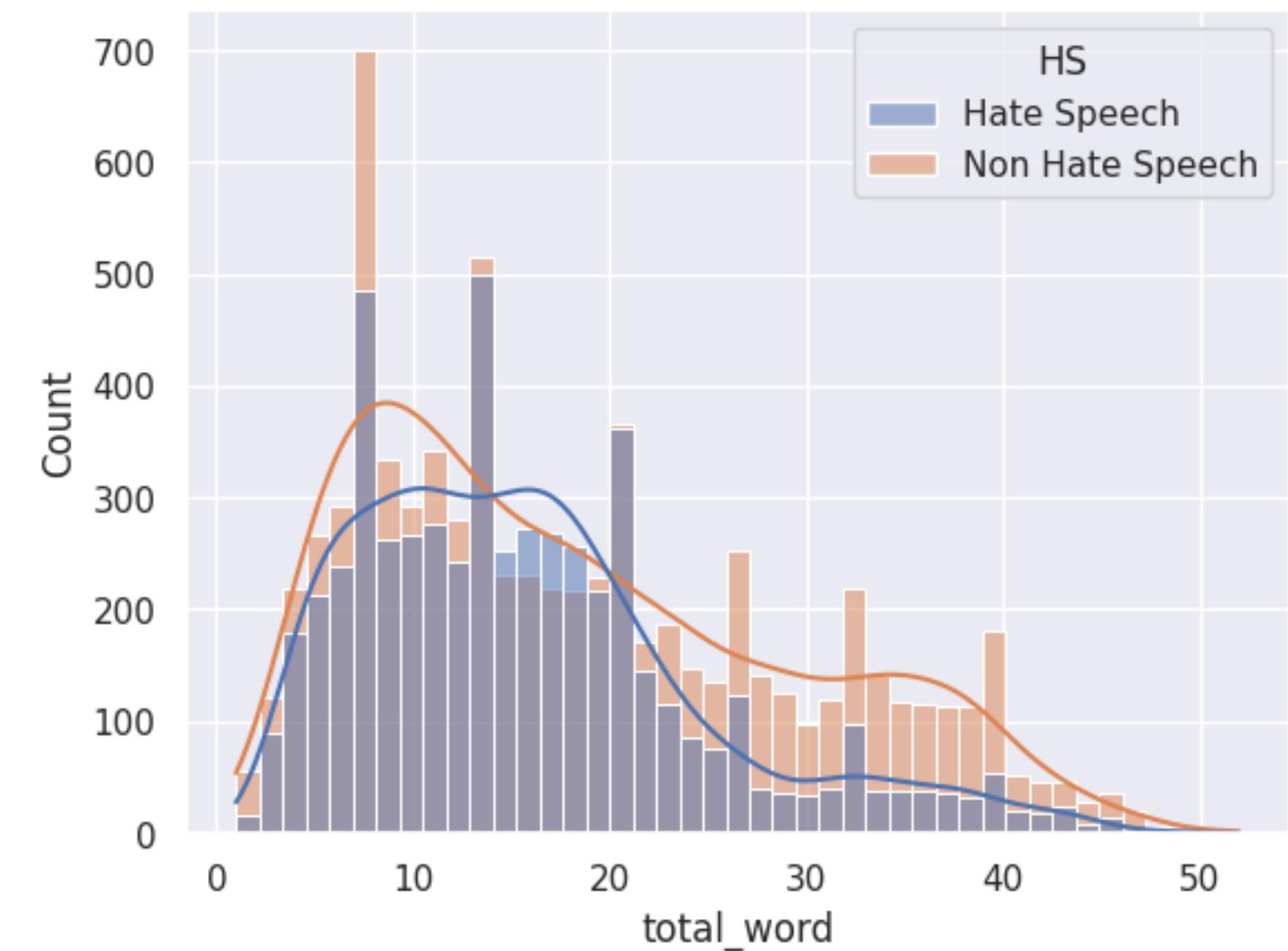
Visualisasi Mosaic memperlihatkan perbandingan Hate Speech Terhadap kategori Abusive didominasi oleh Hate Speech berkategori Abusive.



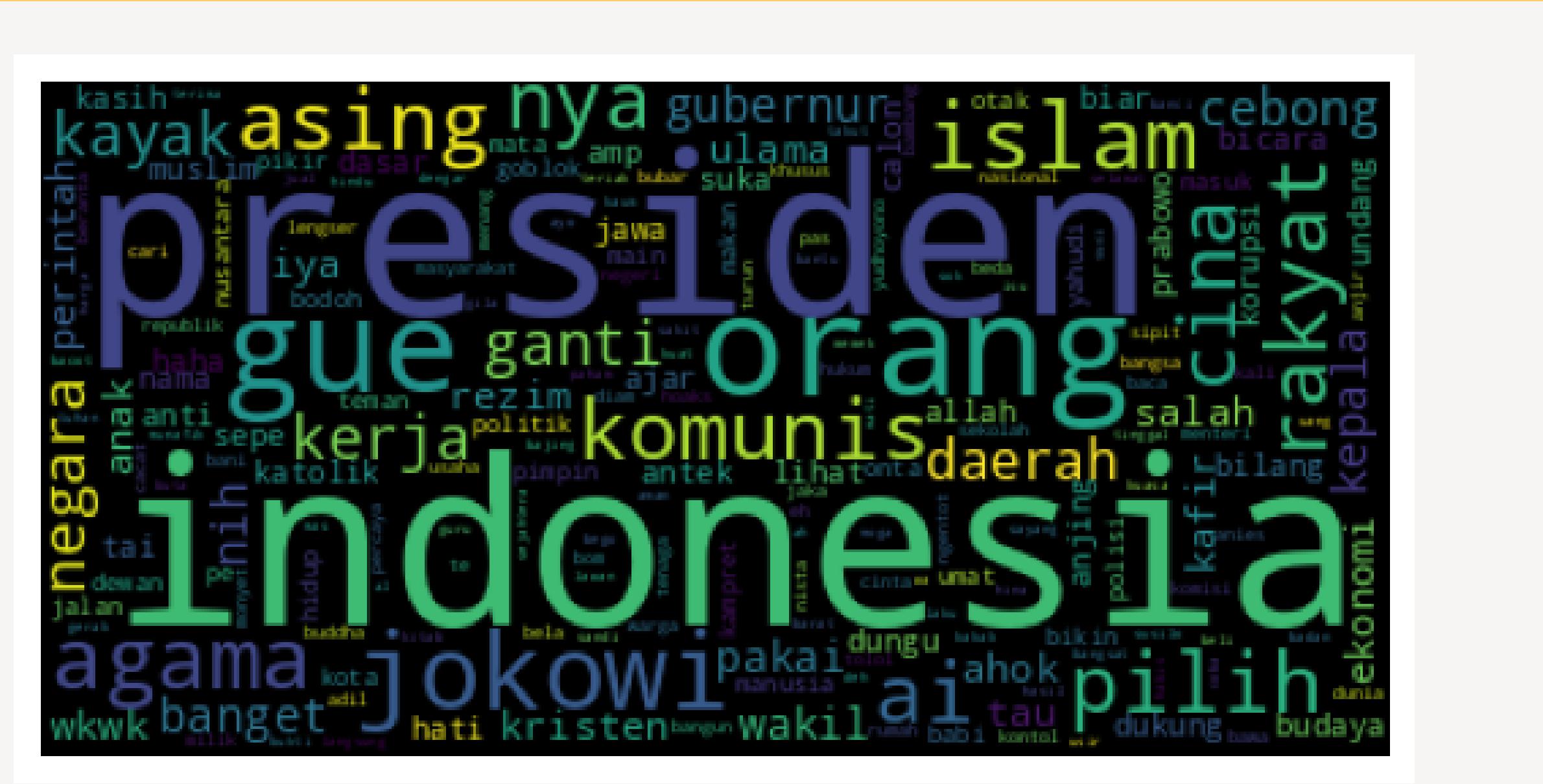
Tweet dengan Hate Speech berada di kisaran 20-150 karakter dan 8-20 kata



Rata-rata total karakter hate speech: 102



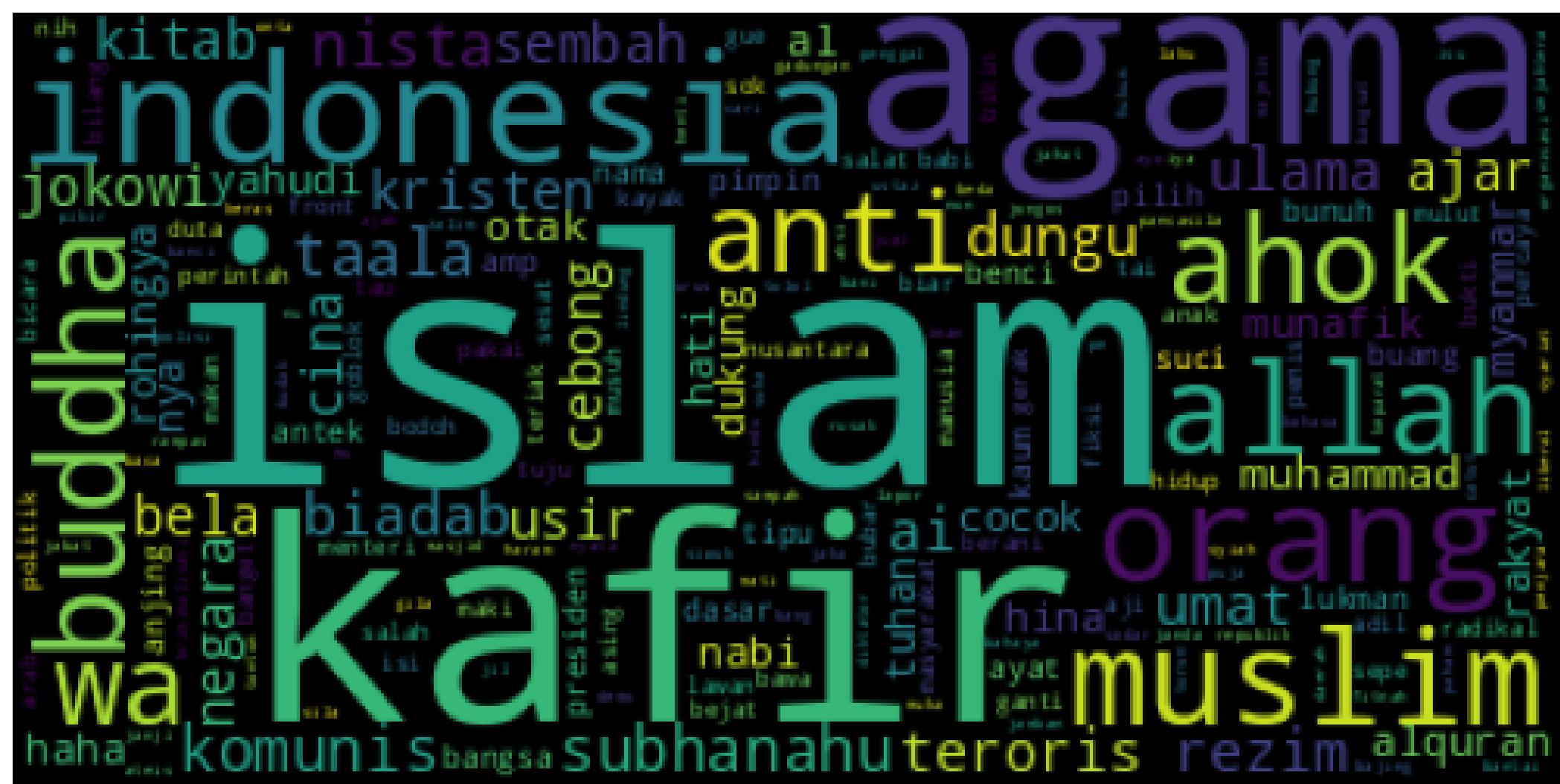
Rata-rata total kata hate speech: 16



Kata yang sering muncul dari **Kategori Individual** diantaranya; **Jokowi, Prabowo, Cebong, Ahok, Presiden.**



Kata yang sering muncul dari **Kategori Group** diantaranya; **Cina, Cebong, Indonesia, Partai komunis, Ganti presiden.**



Kata yang sering muncul dari **Kategori Religion** diantaranya; **Islam, Kafir, Agama, Muslim, Allah, Buddha**



Kata yang sering muncul dari Kategori *Race* diantaranya; **Cina, Komunis, Ai, Indonesia, Ahok**

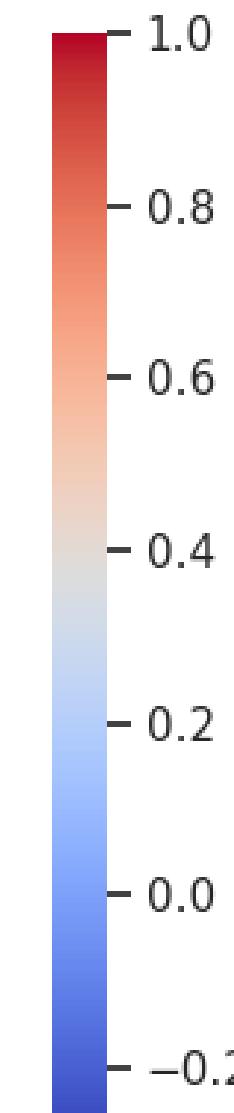
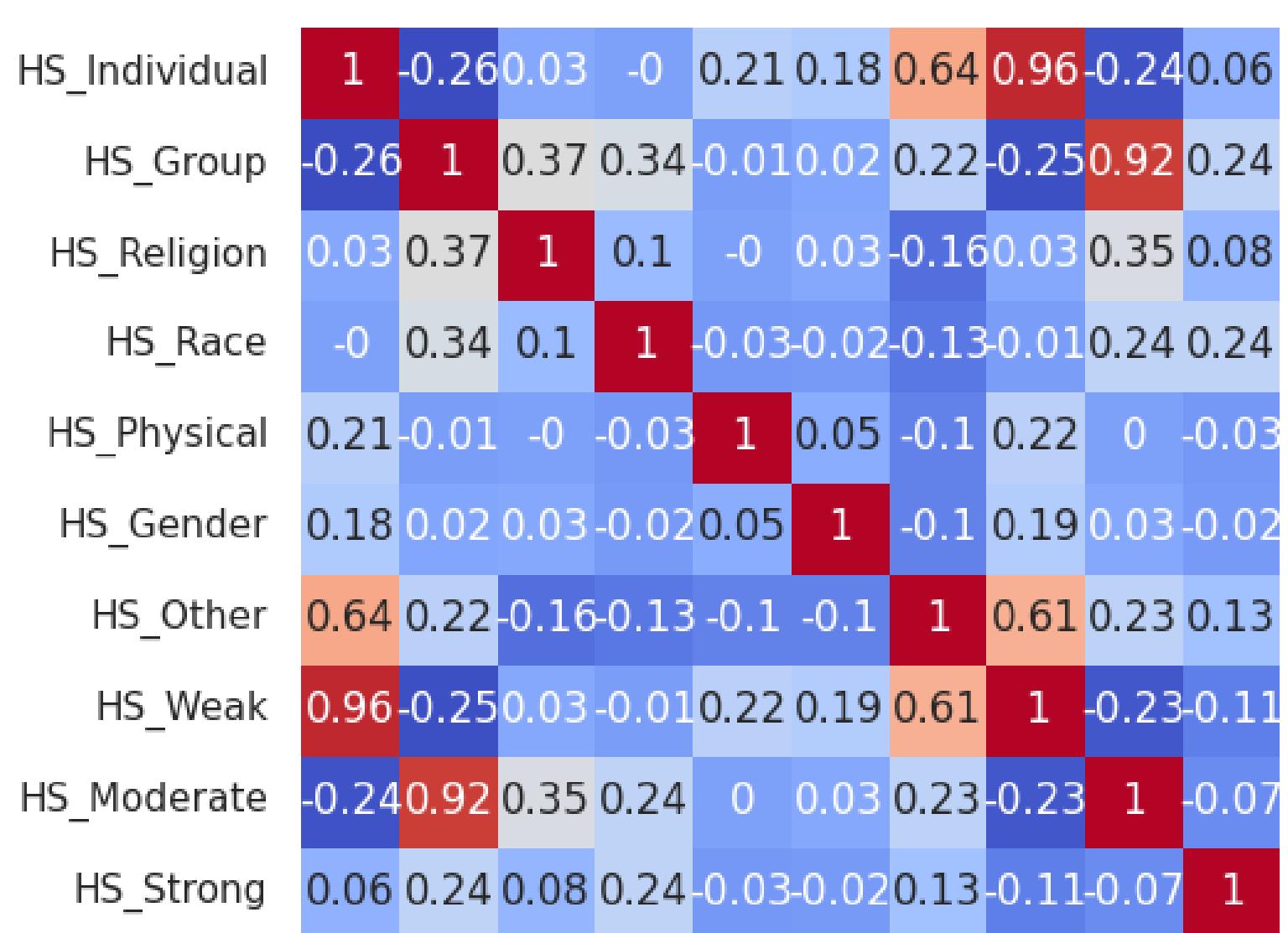


Kata yang sering muncul dari **Kategori Physical** diantaranya; **Idiot, Budek, Bolot, Cacat**



Kata yang sering muncul dari **Kategori Gender** diantaranya; **Banci, Bencong, Homo**

Analisis *Bivariate* menunjukan beberapa karakteristik Kategori yang berhubungan dengan Intensitas Hate Speech



1. Intensitas Rendah

- a. Individual
- b. Other
- c. Gender
- d. Race

2. Intensitas Moderate

- a. Group
- b. Religion
- c. Race

3. Intensitas Tinggi

Seperti yang sudah kelompok kami awali dibab "Metode Penelitian" bahwa terdapat 2 hasil kesimpulan dari metode penelitian deskriptif, yakni:

- **Univariate Analysis** (melibatkan 1 variabel)
- **Bivariate Analysis** (melibatkan 2 variabel)



Univariate Analysis

- Kebanyakan tweet pada periode 2019 - 2020 cenderung **tidak mengandung ujaran kebencian** (non-hate speech) meskipun selisih dengan tweet yang memiliki ujaran kebencian (hate speech) cukup rendah, yaitu berturut-turut 57.76% dan 42.24%.
- Tweet pada periode 2019 - 2020 cenderung **mengandung kata kasar** (abusive), yaitu sebesar 61,67%. Hal ini mengindikasikan perbandingan yang cukup kontras dengan tweet yang tidak mengandung kata kasar (non-abusive), yaitu 38.33%.
- Terdapat outlier yang **tidak signifikan** pada jumlah karakter dan jumlah kata pada tweet, yaitu berturut-turut 0.18% dan 0.65%.
- Lima kata yang paling sering muncul pada hate speech diantaranya adalah **jokowi, cebong, indonesia, presiden** dan **ahok**.
- Rata-rata jumlah karakter yang mengandung hate speech **sebanyak 102 karakter**. Sementara itu, rata-rata jumlah kata yang mengandung hate speech sebanyak **16 kata**.



Bivariate Analysis

- Hate speech dengan intensitas lemah memikili hubungan yang kuat bila ditujukan kepada pribadi tertentu, yaitu sebesar 0.96 diikuti dengan hate speech dengan intensitas menengah yang ditujukan kepada kelompok atau kaum tertentu, yaitu 0.92.
- Hate speech yang **ditujukan kepada individu** memiliki persentase lebih besar, yaitu 64.1% daripada yang ditujukan kepada kelompok (group), yaitu 35.9%.
- Kategori tweet hate speech terbanyak tidak terkait kategori tertentu (other) dengan persentase 65.2%. diikuti dengan **kategori agama, yaitu 13.9%**. Selain itu, hate speech dengan jumlah paling sedikit berkaitan dengan gender sebesar 5.3%.
- Hate speech **cenderung mengandung kata kasar** (abusive) dengan jumlah sebanyak 3250 tweet jika dibandingkan dengan hate speech yang tidak mengandung kata kasar (non-abusive), yaitu sebanyak 2251 tweet.
- Jumlah hate speech yang memiliki **intensitas kuat (strong)** lebih banyak, yaitu 60.6% daripada hate speech dengan intensitas menengah, yaitu 30,8%, dan intensitas lemah, yaitu 8.6%.



Berdasarkan hasil dari hasil analisis *univariate* dan *bivariate*, ada beberapa kesimpulan.

1. Data Tweet didominasi oleh ***non-hate speech***
2. Data *hate speech* ditujukan kepada ***Individu*** menjadi mayoritas data *hate speech*.
3. Intensitas *hate speech* didominasi dengan kategori **rendah**.
4. Kategori *hate speech* memiliki hubungan dengan **intensitas** dan **targetnya**.





Thank You

By:
DSC-Gold Group 2

Special Thanks:
Mentor kami
Wisnu Anggara