

# Exploratory Data Analysis On Multi Label Speech And Abusive Language


By Khadhi Musaid Syah



## Latar Belakang Masalah

Hate speech (HS) dan abusive language di media sosial, khususnya Twitter, telah menjadi perhatian serius karena dampaknya yang merugikan terhadap individu dan kelompok. Penyebaran ujaran kebencian dan bahasa kasar dapat menciptakan lingkungan yang tidak aman dan memicu konflik sosial. Di Indonesia, fenomena ini semakin meningkat seiring dengan bertambahnya pengguna media sosial. Mengidentifikasi dan memahami pola penyebaran hate speech dan abusive language sangat penting untuk mengambil langkah-langkah pencegahan yang efektif.

## Rumusan Masalah

1. Kategori HS apa yang paling banyak di Twitter?
  2. Berapa frekuensi dari penggunaan *Abusive Words* dari kategori HS yang paling banyak di Twitter?
  3. Apa Topik yang berkaitan dengan Abusive Words yang digunakan?
- 



## Tujuan Penelitian

1. Mencari kategori HS apa yang paling banyak di Twitter
2. Mencari frekuensi dari penggunaan *Abusive Words* dari kategori HS yang paling banyak di Twitter
3. Mencari Topik yang berkaitan dengan Abusive Words yang digunakan

## Batasan Masalah

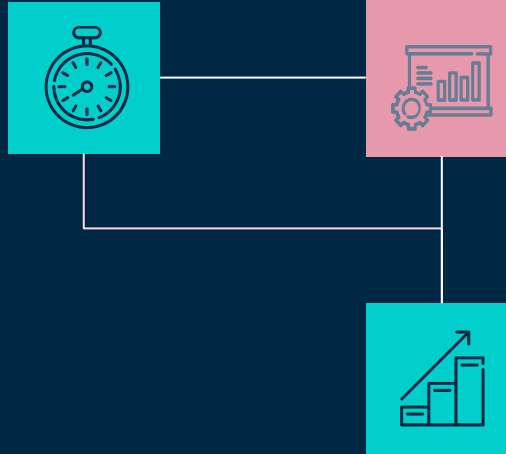
1. Dataset yang digunakan merupakan Bahasa Indonesia, sehingga tidak bisa diterapkan ke bahasa lain.
2. Definisi dari kategori hate speech (HS) dan abusive language mungkin tidak sepenuhnya mencakup semua bentuk ujaran kebencian atau bahasa kasar yang ada di Twitter
3. Keterbatasan dalam normalisasi teks karena kamus ini mungkin tidak lengkap dan tidak mencakup semua varian typo dan slang yang ada dalam bahasa Indonesia di Twitter.



# Metode Penelitian

## Data Cleansing

1. Menggunakan RegEx untuk menghapus simbol, angka, serta kata-kata yang tidak diperlukan.
2. Menghapus row yang duplikat atau mengandung nilai Null
3. Menggunakan Stemming untuk menghapus imbuhan dan mengubah seluruh kata menjadi kata dasar.



## Descriptive Analysis

Mencari nilai modus dari dataset yang sudah tersedia atau hasil modifikasi dataset yang dibuat.

## Data Visualization

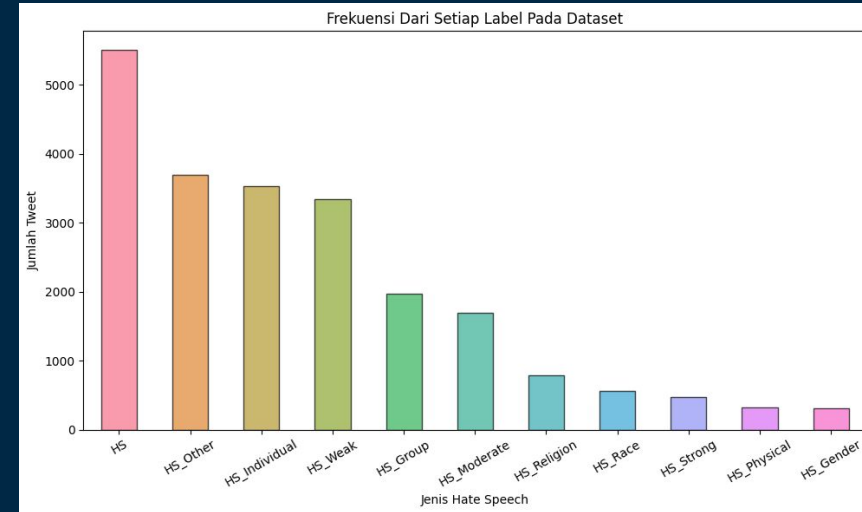
Untuk Data Visualization saya menggunakan Bar Chart untuk melihat perbandingan jumlah antar variabel atau kolom.

# Hasil Penelitian

Dari bar chart di samping, dapat disimpulkan bahwa :

1. **Kategori HS** memiliki frekuensi tertinggi dengan lebih dari 5000 tweet.
2. **Kategori HS\_Other** , **HS\_Individual** , dan **HS\_Weak** memiliki frekuensi yang hampir sama, berkisar antara 3000 hingga 4000 tweet.
3. **Kategori HS\_Moderate** memiliki sekitar 2000 tweet, yang masih cukup signifikan dibandingkan dengan kategori lainnya.
4. **Kategori HS\_Group** dan **HS\_Religion** memiliki frekuensi yang lebih rendah, dengan masing-masing sekitar 1500 dan 1000 tweet.
5. **Kategori HS\_Race** , **HS\_Strong** , **HS\_Physical** , dan **HS\_Gender** memiliki frekuensi yang paling rendah, dengan kurang dari 1000 tweet masing-masing.

Secara keseluruhan, kategori **HS** mendominasi jumlah tweet dalam dataset ini, diikuti oleh kategori **HS\_Other** , **HS\_Individual** , dan **HS\_Weak** . Kategori dengan frekuensi terendah adalah **HS\_Gender** .



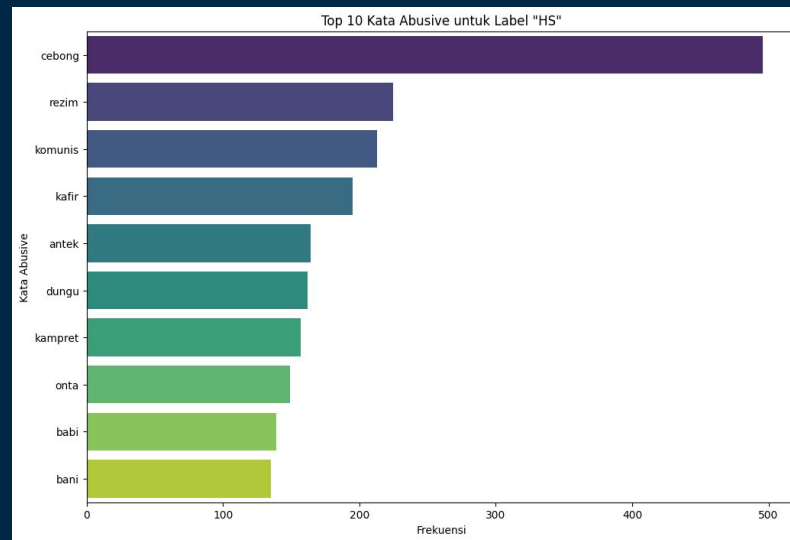
Gambar 1 Frekuensi Dari Setiap Label Hate Speech pada Dataset

# Hasil Penelitian

Dari bar chart di samping, dapat disimpulkan bahwa :

1. Kata "**cebong**" adalah yang paling sering muncul dengan frekuensi tertinggi, mencapai sekitar 500 kali.
2. Kata "**rezim**" dan "**komunis**" juga muncul cukup sering, masing-masing dengan frekuensi sekitar 300 hingga 400 kali.
3. Kata "**kafir**" muncul lebih dari 200 kali.
4. Kata-kata seperti "**antek**", "**dungu**", "**kampret**", "**onta**", "**babi**", dan "**bani**" muncul dengan frekuensi yang sedikit lebih rendah, berkisar antara 100 hingga 200 kali.

Kesimpulannya, kata "**cebong**" adalah yang paling dominan dalam kategori "**HS**", diikuti oleh kata-kata lain yang memiliki konotasi negatif atau menghina. Kata-kata ini mencerminkan jenis ujaran kebencian yang sering digunakan dalam tweet yang termasuk dalam label "**HS**".



Gambar 2 Perbandingan jumlah abusive words di Twitter pada kategori HS secara umum.

1



Tweet

kaum cebong kapir udah keliatan dongknya . awal tambah dongk hahahah  
mah sdh nentak nenek sy heran sama cebong biasay bohong terus  
prokaski mayat politisasi agama penyebab kekalahhan pilkada di beginilah cara cebong mendeskripsikan kekalahhan junjungannya fyi ahog blm pernah ikut pemilihan apapun kac jd wakil bukanlah etnis mayoritza  
cebond bang merasa paling benar  
gobgob bani cebong tukang tpu penjilat penguasa ketahuhan gerakan masa dibayar pakai nasi bungkus propaganda nasi bungkus memang selatol gagal  
wuh cebong sewot n x f x f x x d f x x x d f x f x x  
si tolod udah tau pemerintah rezim komunis bobrok koruptor pimpinan sidang dpr paya cebong urli  
pemberitahuan saracen diframka katak soakan tim prabowo anies sandi sebenarnya media manupahpahkan sakti hati katak cebong ehe  
kamu ane bukan cebong udah diblangin maren ngeyel ntilu maksaa ane jadi cebong  
tahu rakyat makan aspal beton solusi jitu mengurangi kemiskinan keahlian cebong tukang tpu cebong mah gitu goblok ketulungan plonga plongo mengakibatkan optimisme berlebihan gantipresiden  
hahahahha blunder lu cebong gak ngasih tuh pem pusat alasan hah kalau rakyat aceh mau kali cobak kau datangin pemerintah bilang suruh buat hukum islam kafah aceh kalau dpr a  
cebond kacung kit temen gu dungunya sampe tulang sum jd jgn buang energi debat sm x f f x x  
nikita mirzani udah pake hijab ngomong cebond cebond udah pernah fantasin dki balom uh ah cokokers  
kotor ga masalah penting bagiana warnanya cebong mana ngerti  
brahahahaha cebong nya ngahok x f x f x a x a x f x f x a x a cebong blm cebok x f x f x x  
belah duren yaaaaa ngak dungu gak cebong x f x f x x d f x f x x nmaaf  
ahhh kan perasaanmu udah menghibur diri mana cebong tdk bagiana presiden cebong punya tol cebong punya hie elo punya cuma germo x f x x x d f x x x f x f x x  
semua terdapat kelakuan bani cebong  
cebond makin lepanasan jadi makin ngaurur x f x f x x d f x f x x  
gantir rezim kacung lah cebong mah gitu kalau ditanya  
cebond kampret ontak ngak bersinggungan laut  
kamu malah cari musuh dasar peranakan cebong kampret  
mungkin mrk sli tulus sli sebut aja nu nmkt nya mungkin cebong muda kumpangen x f x f x x d f x f x x d f x f x x  
presiden paling buruk sejarah bangsa cebong  
cebond kask fakta bilang dusta cebond dikasih dusta bilang realita cebonger s emang bikin games x f x f x x  
rakyat kaum bani cebong  
biasa bro cili cebong adu domba so lugu namanya cebong buzzer rechehan begini x f x f x x d f x f x x d f x f x x  
cebond pernah senang lihat kejayaan pk anies Irti sipri parahnya lu kumpangen duplii habis habis habis meskipun melakukan buruk  
nggak barisan cebong sakti hati mau pake isu saracen uti nyaran anies sandi gagal dilantik

Gambar 3 Konteks penggunaan kata "cebong" di Tweet user

# Kesimpulan

1. Kategori Tweet: Kategori HS mendominasi dengan lebih dari 5000 tweet, diikuti oleh HS\_Other, HS\_Individual, dan HS\_Weak dengan frekuensi sekitar 3000-4000 tweet. Kategori HS\_Moderate memiliki sekitar 2000 tweet, sementara HS\_Group dan HS\_Religion lebih rendah, sekitar 1500 dan 1000 tweet. Kategori HS\_Race, HS\_Strong, HS\_Physical, dan HS\_Gender memiliki frekuensi terendah, masing-masing kurang dari 1000 tweet.
2. Frekuensi Kata Kunci: Kata "cebong" muncul paling sering dengan sekitar 500 kali, diikuti oleh "rezim" dan "komunis" dengan 300-400 kali. Kata "kafir" muncul lebih dari 200 kali, sementara kata-kata lain seperti "antek" dan "kampret" muncul antara 100-200 kali.
3. Konteks Penggunaan: Kata "cebong" sering digunakan dalam konteks politik untuk merendahkan pendukung Joko Widodo pada Pemilu 2019.

Secara keseluruhan, data menunjukkan dominasi ujaran kebencian politik dengan penggunaan kata-kata negatif yang merendahkan kelompok tertentu.