BINAR DSC#25 | Gold Challenge

Analyzing Indonesian Abusive and Hate Speech on Twitter: Impact on Mental Health

By Isumi Karina

Table of contents

01 Introduction

Background & Supporting Data

02 Methodology

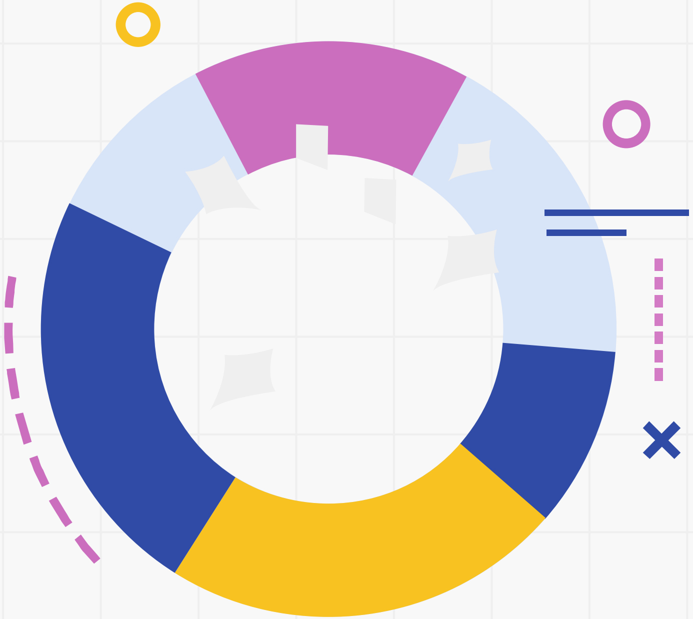
Cleansing Process, Statistical & EDA Methods

03 Visualization

Visualization Process

04 Results & Conclusion

Results, Conclusion, and Recommendation



01

Introduction

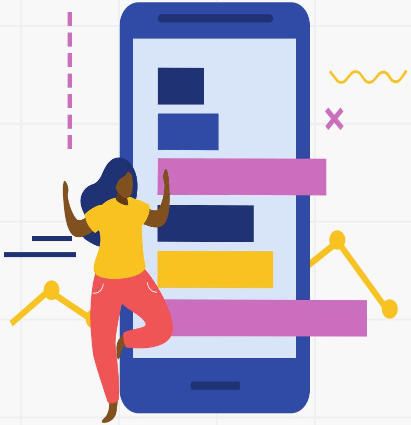
Background & Supporting Data

Background

In recent years, the spread of **hate speech** and **abusive speech** on social media platforms, particularly Twitter, has surged in Indonesia. This phenomenon poses serious risks to mental health, especially among younger users who are heavily engaged in online discourse.

Hate speech, often targeting individuals based on race, religion, gender, and other identity markers, not only fosters a hostile environment but also exacerbates **mental health issues** such as **stress, anxiety, and depression**.

According to a 2024 study by Monash University Indonesia, the presence of hate speech in social media increased dramatically, particularly during politically sensitive times (Monash University). UNESCO Indonesia, through the "Social Media 4 Peace" initiative, highlighted that hate speech and toxic online behavior are major contributors to deteriorating mental health (UNESCO). Data collected over the past three years shows that online abuse significantly affects individuals' well-being, with 30% of users exposed to hate speech experiencing symptoms of stress and depression (Oxford Academic) (UNESCO).



Notes:
The MDDRH team's findings showed that hate speech appeared most frequently on Twitter, at 51.2%. Meanwhile, hate speech was 45.15% on Facebook and 3.34% on Instagram.

Justification

Given this alarming trend and the increasing reports of mental health problems tied to social media interactions, this project was undertaken using **Kaggle's dataset on Indonesian abusive and hate speech on Twitter (here)**. The data includes **over 13,000 tweets labeled according to various categories**, such as **hate speech**, **abusive speech**, and specific targeted forms of hate (e.g., race, religion, gender). For this analysis, I used three specific files from the Kaggle dataset:

1. **dataset.csv**: Contains tweet text and 12 columns categorizing hate speech and abusive behavior. Each column contains a binary indicator (1 = present, 0 = not present) for different forms of hate speech, such as targeting individuals or groups based on religion, race, gender, etc.
2. **abusive.csv**: A collection of abusive words used in Indonesian tweets, stored in a single column.
3. **new_kamusalay.csv**: A two-column file containing slang or "alay" words and their normalized equivalents, aiding in the data cleansing process to better understand the context of the speech.

These files provide the basis for text analysis, allowing for the categorization, cleansing, and in-depth analysis of the data. This project aims to contribute to ongoing efforts to support mental health, especially during **World Mental Health Day** in October, by shedding light on the extent and impact of harmful online behavior in Indonesia.

02

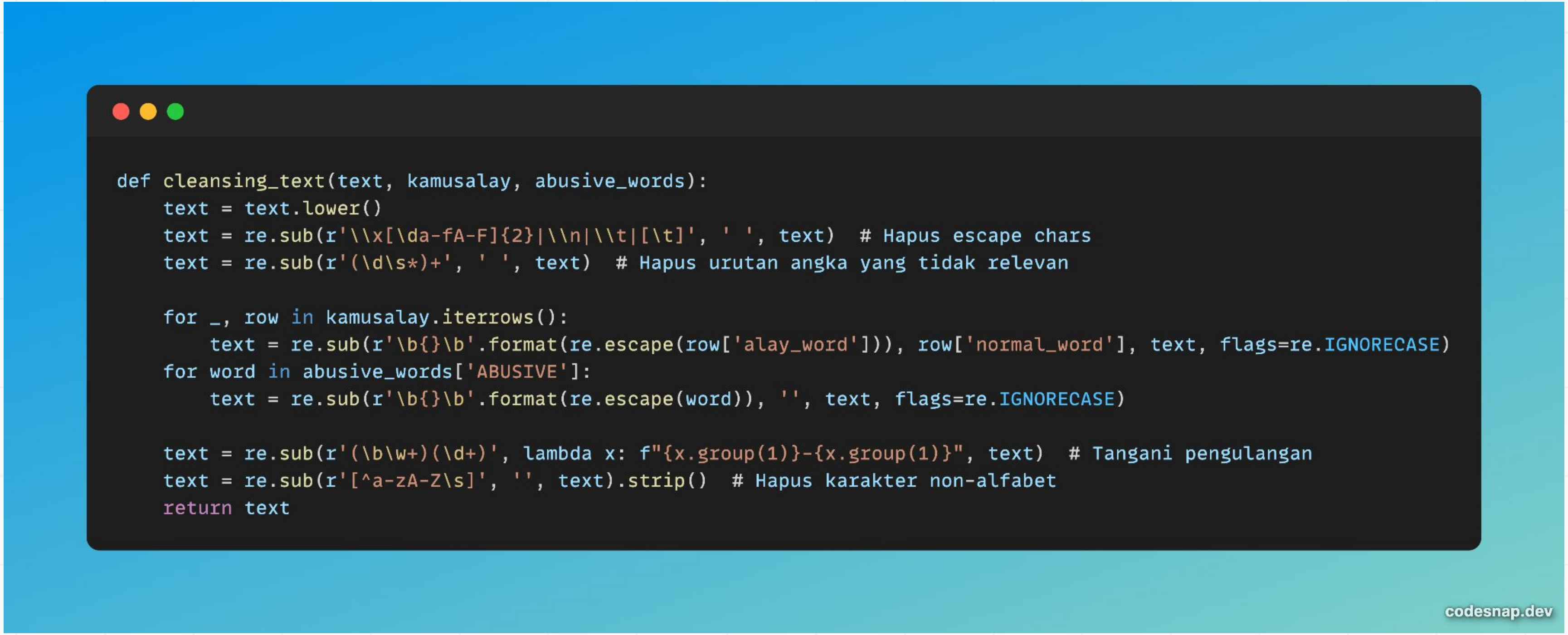
Methodology

Cleansing Process, Statistical & EDA Methods

Cleansing Process

Given the need to analyze the content of tweets, **data cleansing** plays a vital role in ensuring accurate analysis. The cleansing process also takes advantage of the use of **Regex** and other functions, including:

- **Lowercasing**: Converting the entire tweet text to lowercase ensures that all text is treated uniformly, which simplifies the cleansing process.
- **Removing Escape Characters and Unnecessary Whitespaces**: This step strips the text of any unwanted escape characters such as `\n`, `\t`, or hexadecimal sequences like `\xF0`. This ensures that the text is more readable and analyzable.
- **Removing Numeric Sequences**: Irrelevant numeric sequences are often present in tweets, especially in noisy data like social media text. This regex removes them while preserving meaningful text.
- **Replacing Slang Words (Alay)**: Using the `new_kamusalay.csv` dataset, slang words are normalized to standard Indonesian. This makes the text easier to interpret during analysis.
- **Removing Abusive Words**: The `abusive.csv` file contains a list of abusive words, and this step removes any such words found in the tweet.
- **Handling Repeated Word Patterns**: Using a **lambda function**, repeated words with numbers (e.g., `cantik2`) are converted into the proper format (e.g., `cantik-cantik`). The lambda captures the word and constructs the repeated form based on the number.

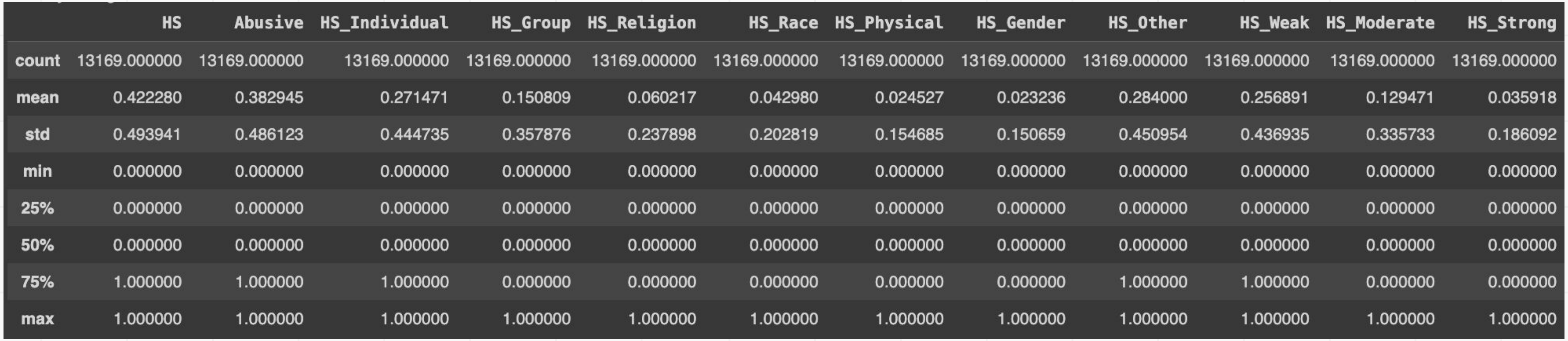


```python
def cleansing_text(text, kamusalay, abusive_words):
    text = text.lower()
    text = re.sub(r'\\x[\da-fA-F]{2}|\\n|\\t|[\t]', ' ', text)  # Hapus escape chars
    text = re.sub(r'(\d\s*)+', ' ', text)  # Hapus urutan angka yang tidak relevan

    for _, row in kamusalay.iterrows():
        text = re.sub(r'\b{}\b'.format(re.escape(row['alay_word'])), row['normal_word'], text, flags=re.IGNORECASE)
    for word in abusive_words['ABUSIVE']:
        text = re.sub(r'\b{}\b'.format(re.escape(word)), '', text, flags=re.IGNORECASE)

    text = re.sub(r'(\b\w+)(\d+)', lambda x: f"{x.group(1)}-{x.group(1)}", text)  # Tangani pengulangan
    text = re.sub(r'[^a-zA-Z\s]', '', text).strip()  # Hapus karakter non-alfabet
    return text
```

Statistical and EDA Methods

To uncover insights related to the influence of hate speech on mental health, the following statistical and exploratory data analysis (EDA) methods are employed:

- **Countplot**: This method visualizes the distribution of tweets containing hate speech and abusive speech, giving a clear understanding of the prevalence of these behaviors.
- **Correlation Heatmap**: Used to identify relationships between different types of hate speech (e.g., race, religion, gender) to understand which categories co-occur more frequently.

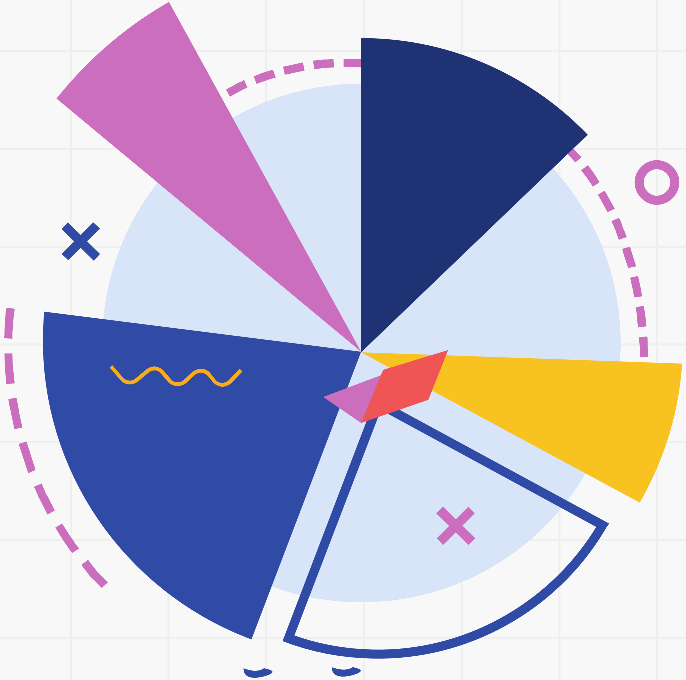


	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
count	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000
mean	0.422280	0.382945	0.271471	0.150809	0.060217	0.042980	0.024527	0.023236	0.284000	0.256891	0.129471	0.035918
std	0.493941	0.486123	0.444735	0.357876	0.237898	0.202819	0.154685	0.150659	0.450954	0.436935	0.335733	0.186092
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

03

Visualization

Visualization Process



Visualization Process

Visualizations (using Matplotlib & Seaborn) are created to help better understand the distribution and impact of hate speech and abusive speech:

1. **Distribution of Hate Speech and Abusive Speech**:
   Bar charts visualize the frequency of tweets containing hate speech and abusive speech.

2. **Correlation of Hate Speech Types**:
   A heatmap identifies correlations between various hate speech categories, revealing patterns and overlaps.

3. **Comparison of Hate Speech and Abusive Speech**:
   A chart comparing tweets containing both hate speech and abusive speech, those with only one, and those containing neither, providing insights into the overlap and severity.

Additional Note:

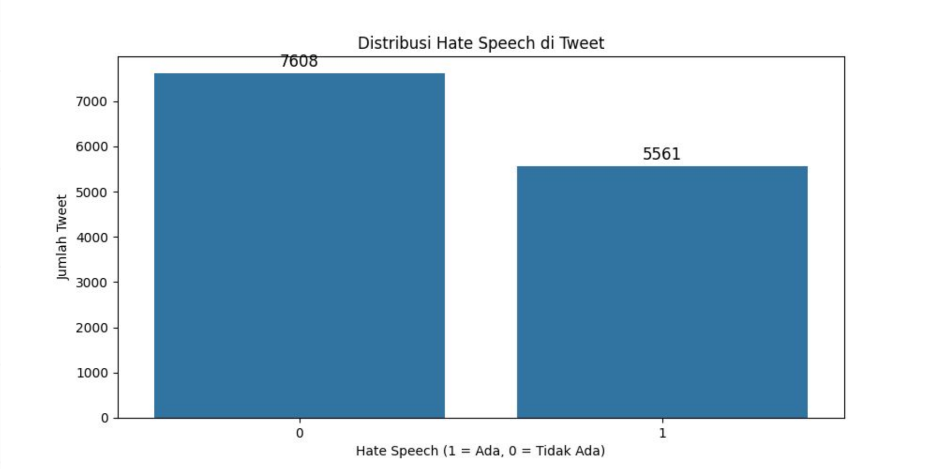
The visualizations from this analysis are also available in **Google Colab** at [here]. However, due to the large size of the dataset (over 13,000 tweets), if you plan to run the code in Google Colab, you can use the smaller `data_mini.csv` file (containing only 20 rows of data) for faster execution and testing purposes.

Example Visualization (Countplot)

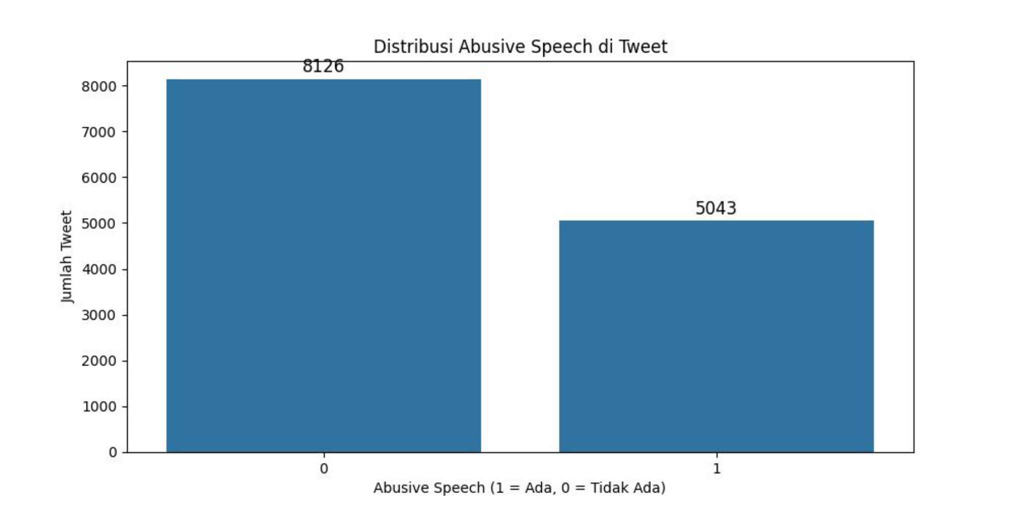
Distribution of Hate Speech

Hate Speech (HS): Out of a total of 13,169 tweets, approximately **42.2%** contain hate speech (HS = 1), while **57.8%** do not contain hate speech (HS = 0).

The number of tweets without hate speech is greater than those with hate speech (7,608 tweets without hate speech vs. 5,561 tweets with hate speech).



Example Visualization (Countplot)



Distribution of Abusive Speech

Abusive Speech: **38.3%** of the tweets contain abusive speech (Abusive = 1), while **61.7%** do not contain abusive speech (Abusive = 0).

The number of tweets containing abusive speech is fewer than those without abusive speech (8,126 tweets without abusive speech vs. 5,043 tweets with abusive speech).

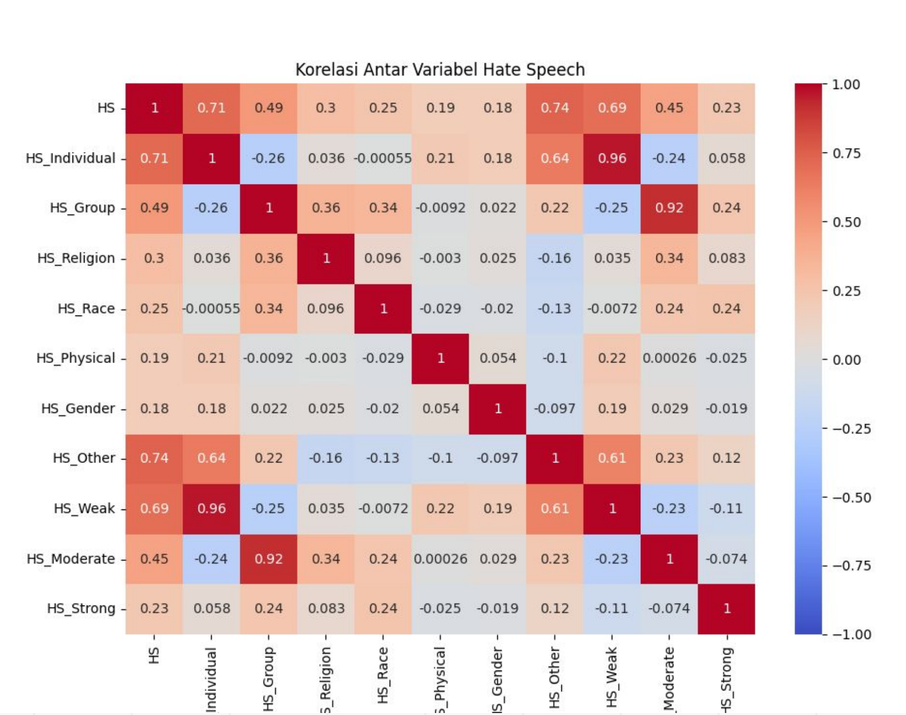
Example Visualization (Heatmap)

Correlation Between Hate Speech Variables

High Correlation: The variables **HS_Weak** and **HS_Individual** show a very high correlation (0.96), indicating that tweets categorized as weak hate speech (HS_Weak = 1) are often also categorized as hate speech targeted at individuals (HS_Individual = 1).

Moderate Correlation: The variables **HS** and **HS_Other** show a moderate correlation (0.74), suggesting that if a tweet contains general hate speech (HS = 1), it is often categorized as hate speech in other unspecified categories (HS_Other = 1).

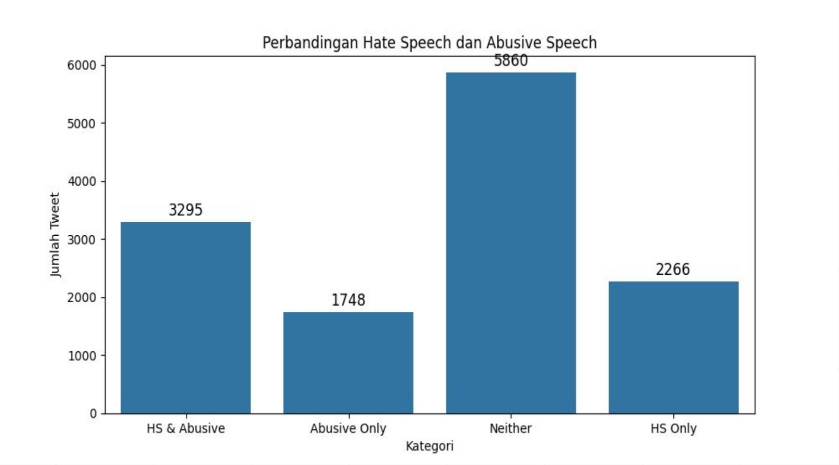
Negative Correlation: A negative correlation is found between **HS_Weak** and **HS_Group** (-0.25), suggesting that hate speech directed at groups (HS_Group = 1) is less likely to be categorized as weak hate speech.



Korelasi Antar Variabel Hate Speech

	HS	Individual	HS_Group	S_Religion	HS_Race	S_Physical	IS_Gender	HS_Other	HS_Weak	Moderate	HS_Strong
HS	1	0.71	0.49	0.3	0.25	0.19	0.18	0.74	0.69	0.45	0.23
HS_Individual	0.71	1	-0.26	0.036	-0.00055	0.21	0.18	0.64	0.96	-0.24	0.058
HS_Group	0.49	-0.26	1	0.36	0.34	-0.0092	0.022	0.22	-0.25	0.92	0.24
HS_Religion	0.3	0.036	0.36	1	0.096	-0.003	0.025	-0.16	0.035	0.34	0.083
HS_Race	0.25	-0.00055	0.34	0.096	1	-0.029	-0.02	-0.13	-0.0072	0.24	0.24
HS_Physical	0.19	0.21	-0.0092	-0.003	-0.029	1	0.054	-0.1	0.22	0.00026	-0.025
HS_Gender	0.18	0.18	0.022	0.025	-0.02	0.054	1	-0.097	0.19	0.029	-0.019
HS_Other	0.74	0.64	0.22	-0.16	-0.13	-0.1	-0.097	1	0.61	0.23	0.12
HS_Weak	0.69	0.96	-0.25	0.035	-0.0072	0.22	0.19	0.61	1	-0.23	-0.11
HS_Moderate	0.45	-0.24	0.92	0.34	0.24	0.00026	0.029	0.23	-0.23	1	-0.074
HS_Strong	0.23	0.058	0.24	0.083	0.24	-0.025	-0.019	0.12	-0.11	-0.074	1

Example Visualization (Countplot)



Comparison of Hate Speech and Abusive Speech:

- ○ **3,295** tweets contain both hate speech and abusive speech.
- ○ **1,748** tweets contain only abusive speech without hate speech.
- ○ **2,266** tweets contain only hate speech without abusive speech.
- ○ **5,860** tweets do not contain hate speech or abusive speech.

This data reveals that most tweets containing abusive speech also contain hate speech. However, there are also tweets that exclusively contain either hate speech or abusive speech.

LOOK

04

Results and Conclusion

Results and Conclusion

Results:

1. **Prevalence of Hate Speech**: Approximately **42%** of analyzed tweets contained hate speech, while **38%** contained abusive speech. A large overlap exists between the two, with many tweets exhibiting both characteristics.
2. **Correlation Findings**: The correlation analysis reveals that weak hate speech is strongly associated with hate speech targeting individuals, general hate speech often overlaps with other unspecified categories, and hate speech directed at groups tends to be less frequently categorized as weak hate speech.

Conclusion:

The findings suggest a serious issue with the prevalence of **hate speech** and **abusive speech** on Indonesian social media platforms. These toxic behaviors exacerbate **mental health issues** among users, particularly in sensitive or marginalized communities.

Recommendations:

1. **Platform Moderation**: Social media platforms should enhance content moderation efforts using AI-powered tools to detect and mitigate hate speech and abusive speech.
2. **Mental Health Awareness Campaigns**: Broaden awareness programs during **World Mental Health Day** to emphasize the detrimental impact of online toxicity on mental health.
3. **Policy Changes**: Governments and institutions should collaborate with social media companies to establish stricter policies on hate speech and provide mental health resources for affected individuals.

Thanks!

Do you have any questions?

isumi.karina72@gmail.com

Github: **@isumizumi**

Google Colab: **here**

CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik