BINAR ACADEMY

THE PHENOMENON OF BADMOUTHING AMONG INDONESIAN TWITTER USER

Regina Aprilia Roberto

COMPONENTS

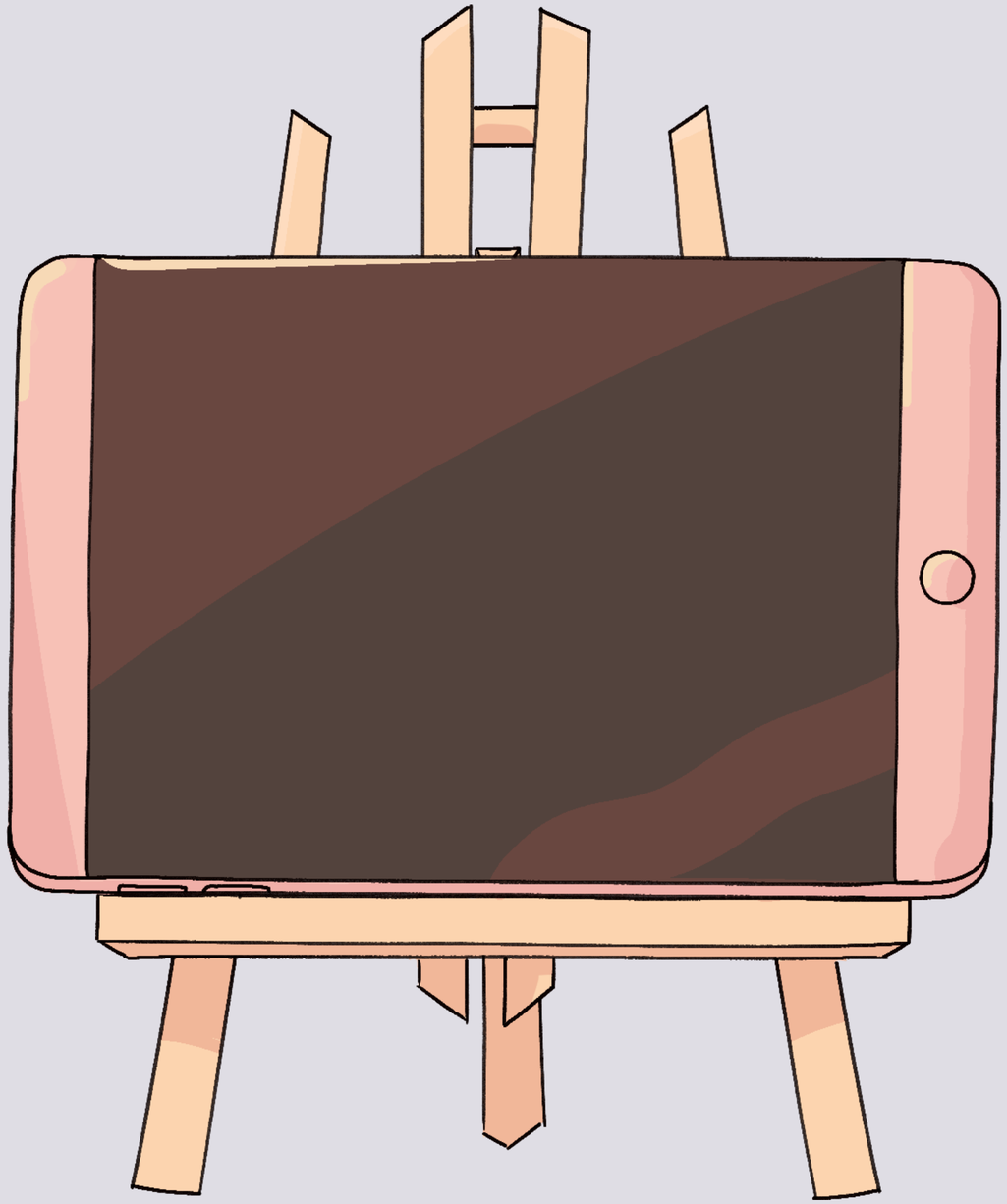
Introduction:
Background, Problem Statements, and Aims

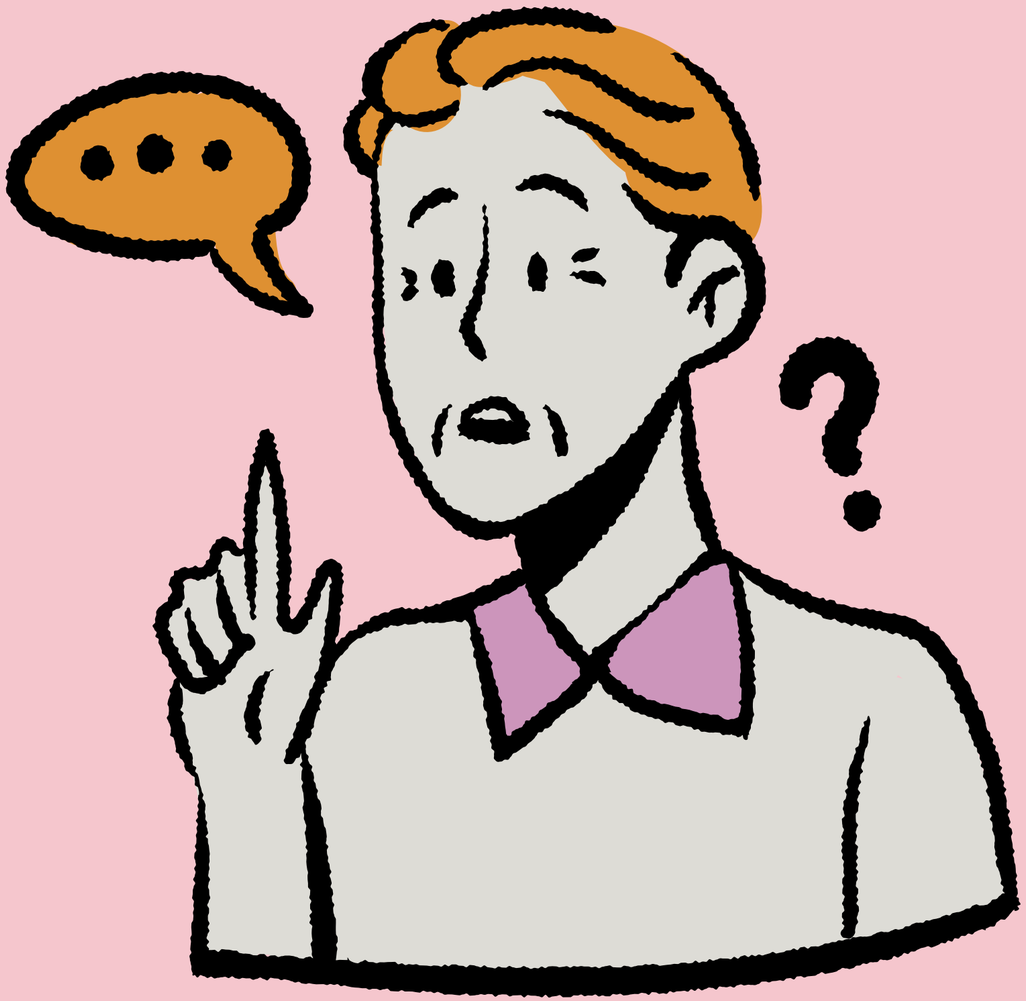
Methodology:
Data description, Preprocessing, Relevant findings

Results & Summary:
Visualizations, Key Points

BACKGROUND

- **Twitter** was the **second biggest platform** for **online abuse**.
- The most **significant type of abuse** was from **online insults** (O'driscoll, 2022).
- **Hate speech** becomes **Indonesia's most frequently reported online crime** since 2016 and has been rapidly increase.
- **Indonesia's digital civility index** is the **worst** in **Southeast Asia**. This assessment is based on Indonesia citizen online behavior such as spreading hoaxes, hate speech, trolling and cyberbullying.





PROBLEM STATEMENTS

- How **frequent** tweets contain hate speech appear?
- How many **category** does tweets contain hate speech have?
- What is the **most common word** used in tweet contain hate speech?
- How frequent tweets contain hate speech can **lead social conflict**?

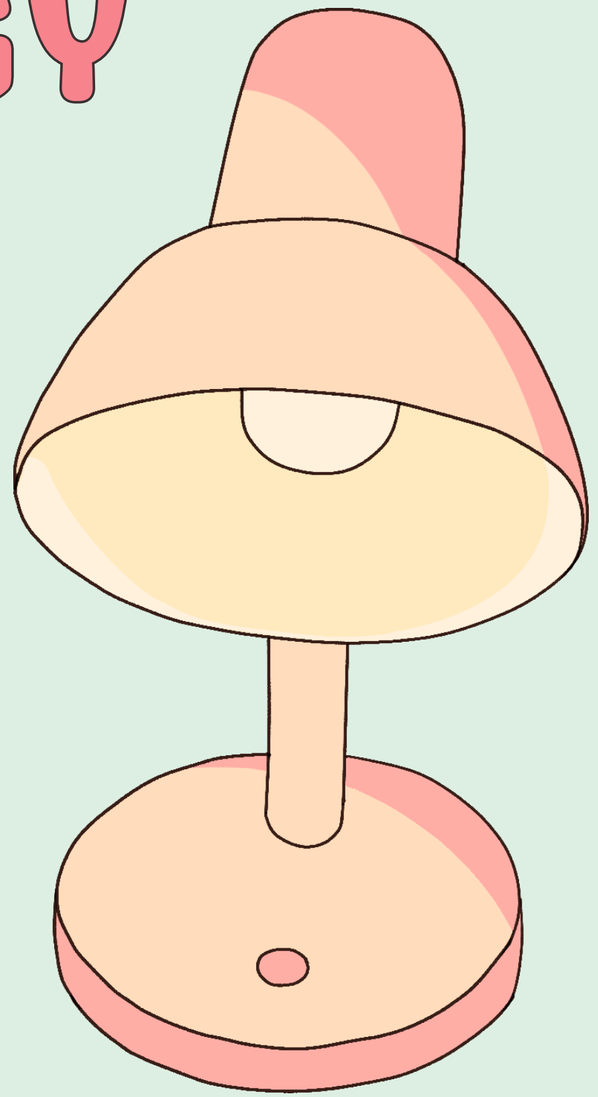
OBJECTIVE

Provide an overview of hate speech (badmouthing) phenomenon in Indonesian Twitter by its frequency, category, most common hate word, and its potential to cause social conflict.

METHODOLOGY

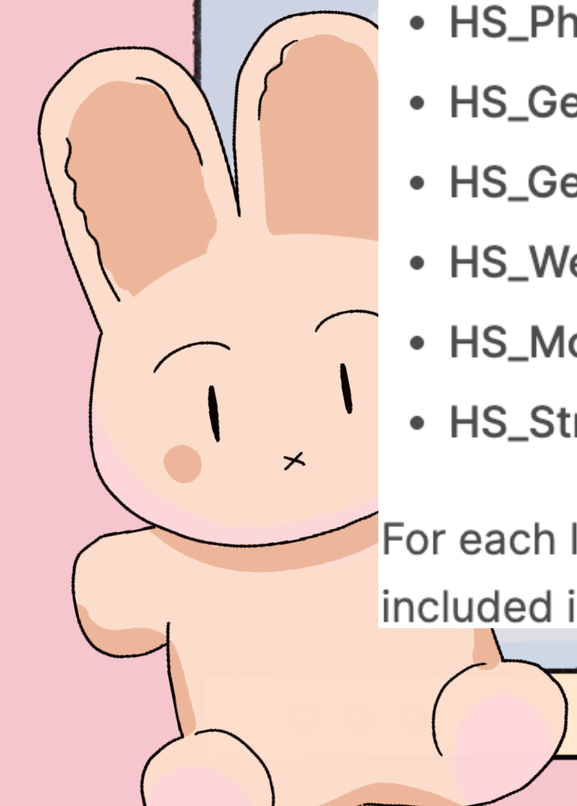
Data description

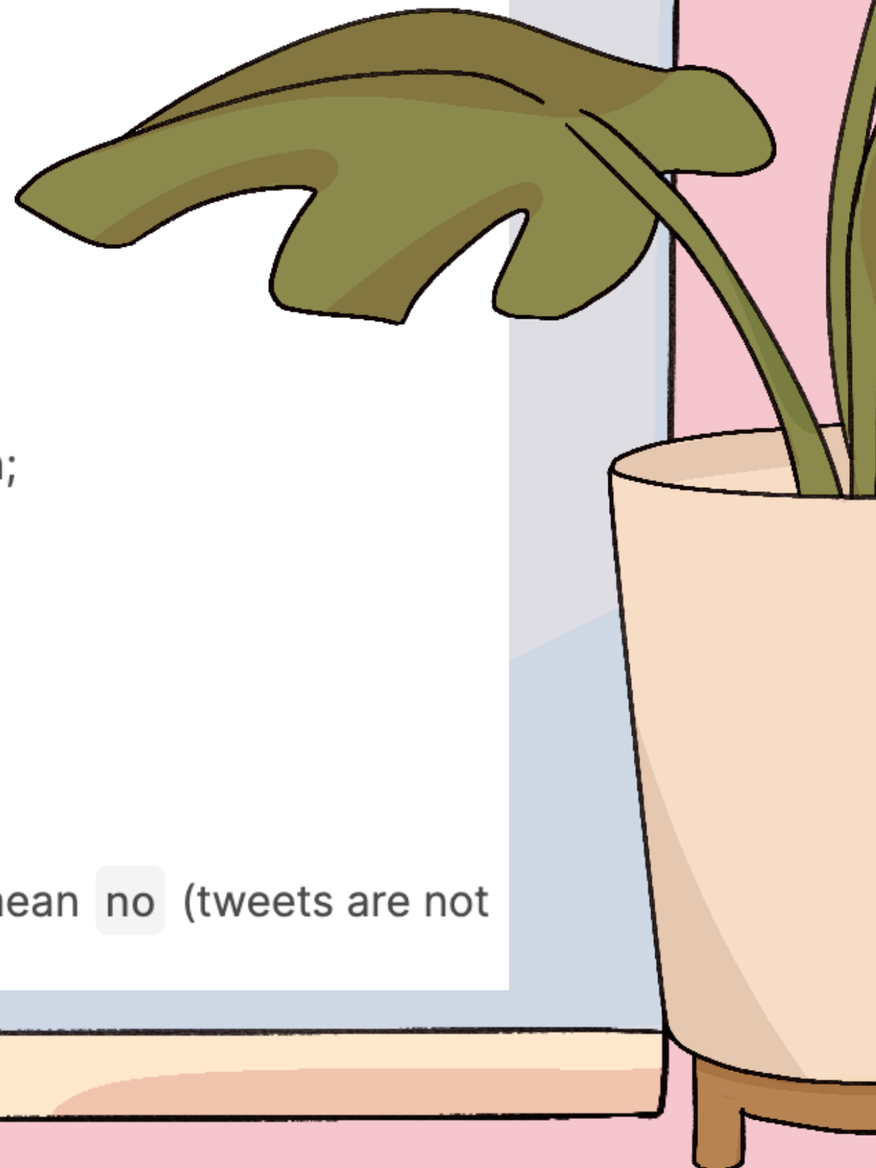
- Dataset is collected tweets contain hate speech and abusive which already group based on label.
- Data source: **https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection**
- Data sources: Twitter
- Collection methodology: Crawl using Tweepy Library
- Data type: string & boolean.



- **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_Gender** : hate related to other invective/slander;
- **HS_Weak** : weak hate speech;
- **HS_Moderate** : moderate hate speech;
- **HS_Strong** : strong hate speech.

For each label, `1` means `yes` (tweets including that label), `0` mean `no` (tweets are not included in that label).





Preprocessing (before: 13169, after: 12901)

- Case folding
- Removing symbol/punc/emoticon using re
- Adding Stopwords
- Drop duplicates

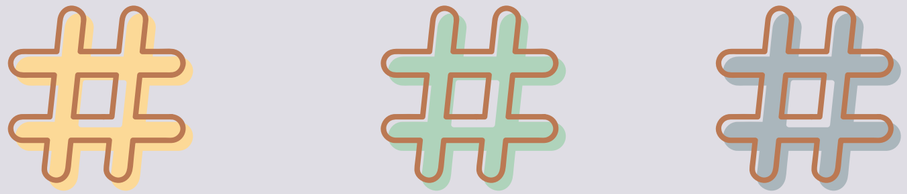
	Tweet
0	- disaat semua cowok berusaha melacak perhatia...
1	RT USER: USER siapa yang telat ngasih tau elu?...
2	41. Kadang aku berfikir,

	Tweet
0	cowok berusaha melacak perhatian lantas remehk...
1	telat edan sarap bergaul cigax jifla calis licew
2	kadang berpikir percaya tuhan

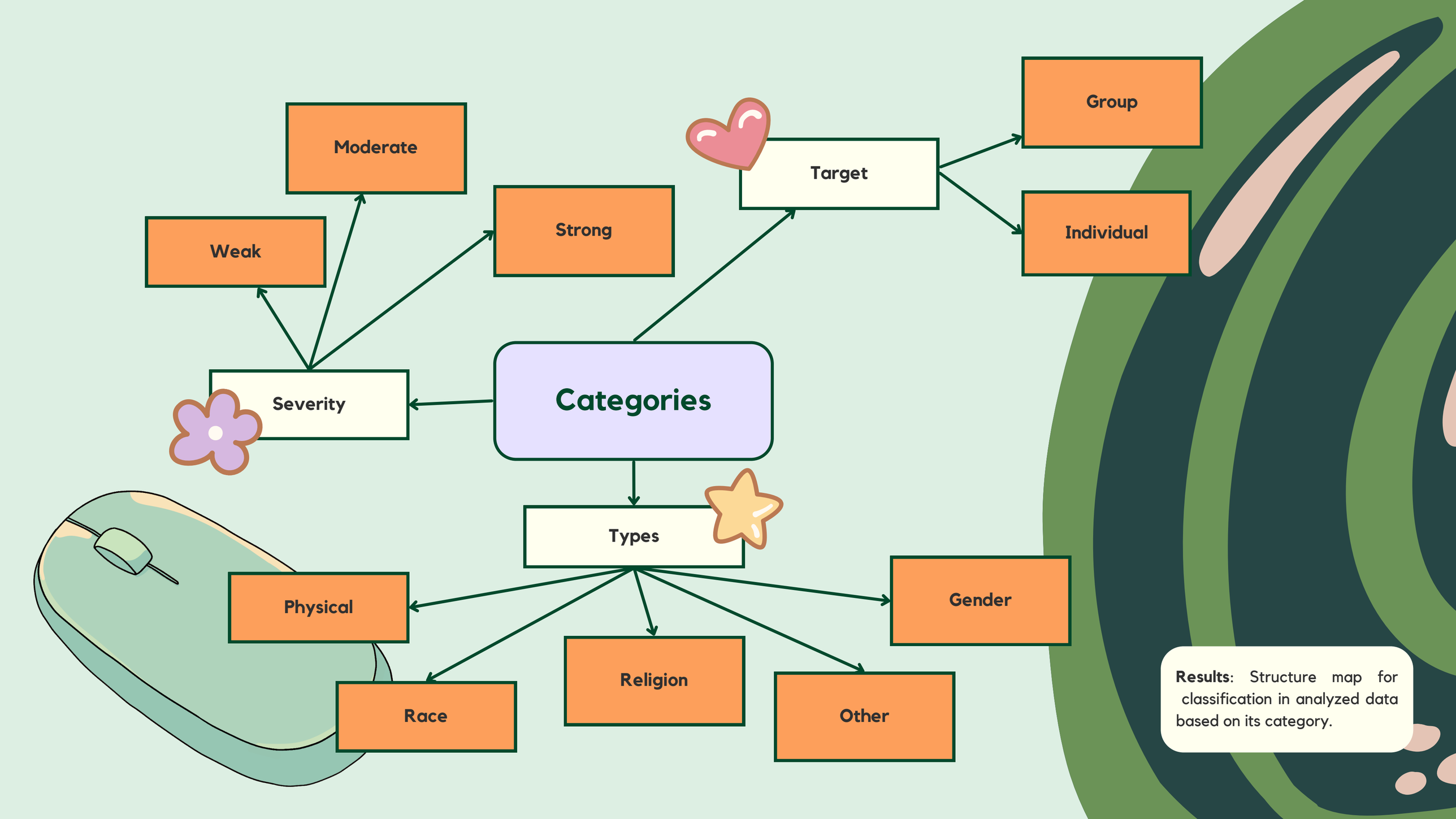
Dataset	Size	Text Genre	Topic	Labels	Annotation Level	#of Annotators	Agreement Measurement
HateMotiv	5000	Twitter	Hate crimes and their motivation	Hate crime types: - Physical assault - Verbal abuse - Incitement to hatred - Other Motivation: - Racism - Religion - Disability - Sexism - Unknown	mention	2 annotators	F-score 0.66 for hate crimes type 0.71 for the motivation of hate crimes
Waseem and Hovy [23]	16,914	Twitter	Hate speech	Racism Sexism Both Neither	Tweet	Crowdsource workers	Cohen's kappa = 0.57
Davidson et al. [15]	25,000	Twitter	Hate speech	Hate Offensive Neither	Tweet	Crowdsource workers	InterCoder-agreement score 92%
HatEval [25]	19,600	Twitter	Hate speech against immigrants or women	Hate Not hate Aggressive Not aggressive	Tweet	Crowdsource workers	Average IAA at 0.75

RELEVANT FINDINGS

Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus



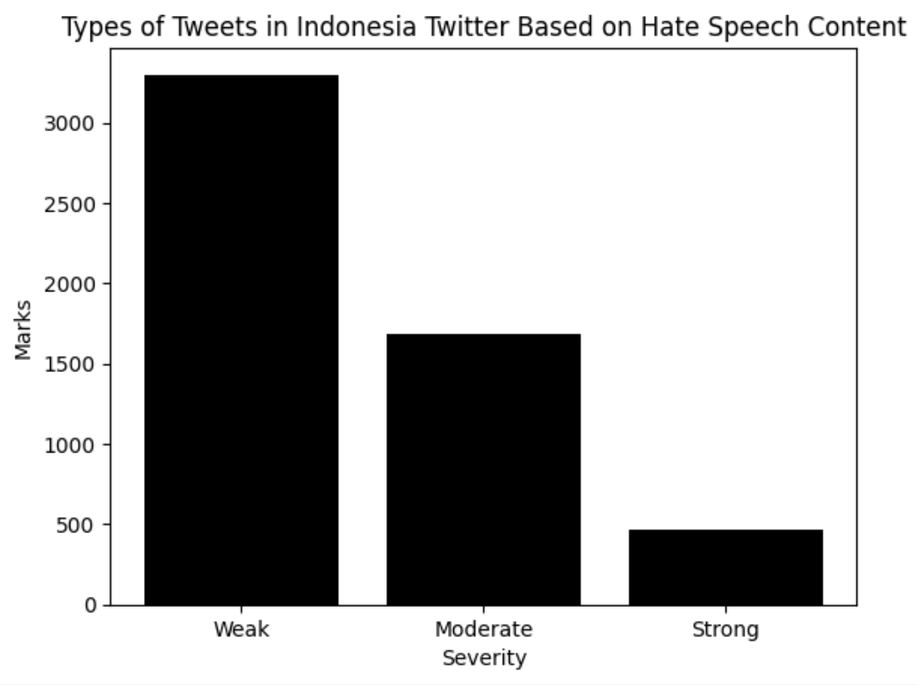
RESULTS & SUMMARY

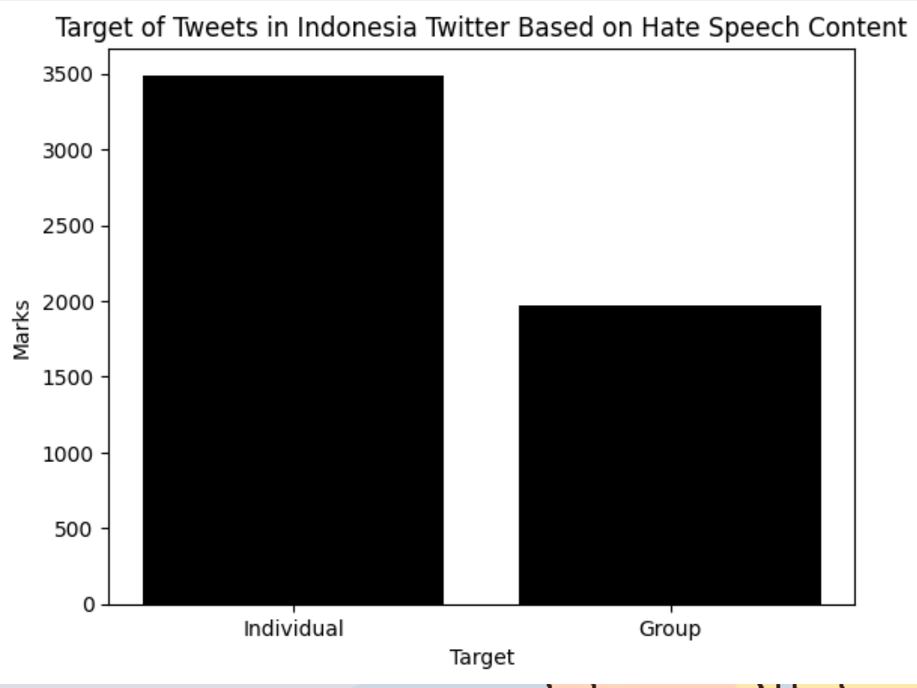


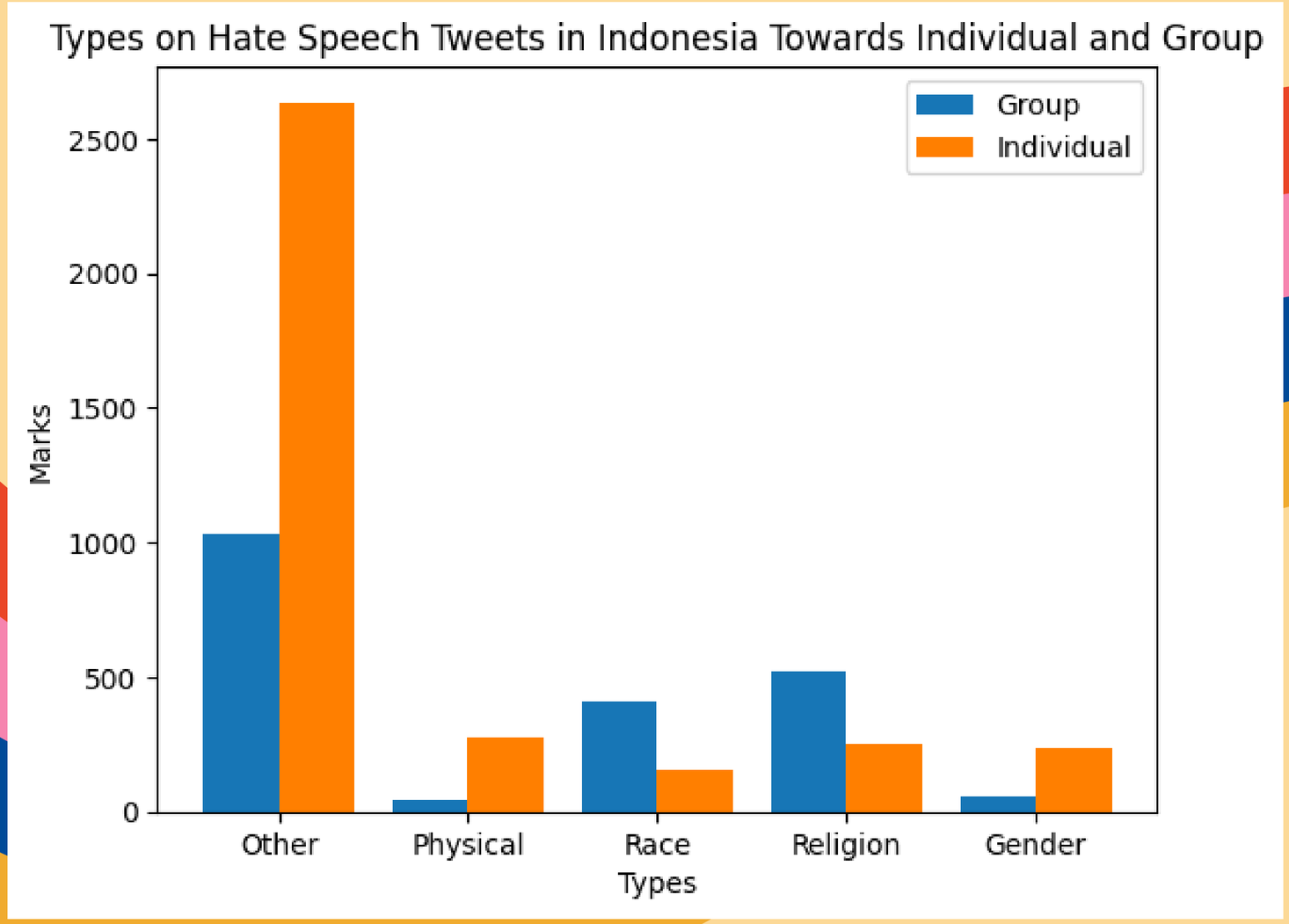
Results: Structure map for classification in analyzed data based on its category.

RESULTS

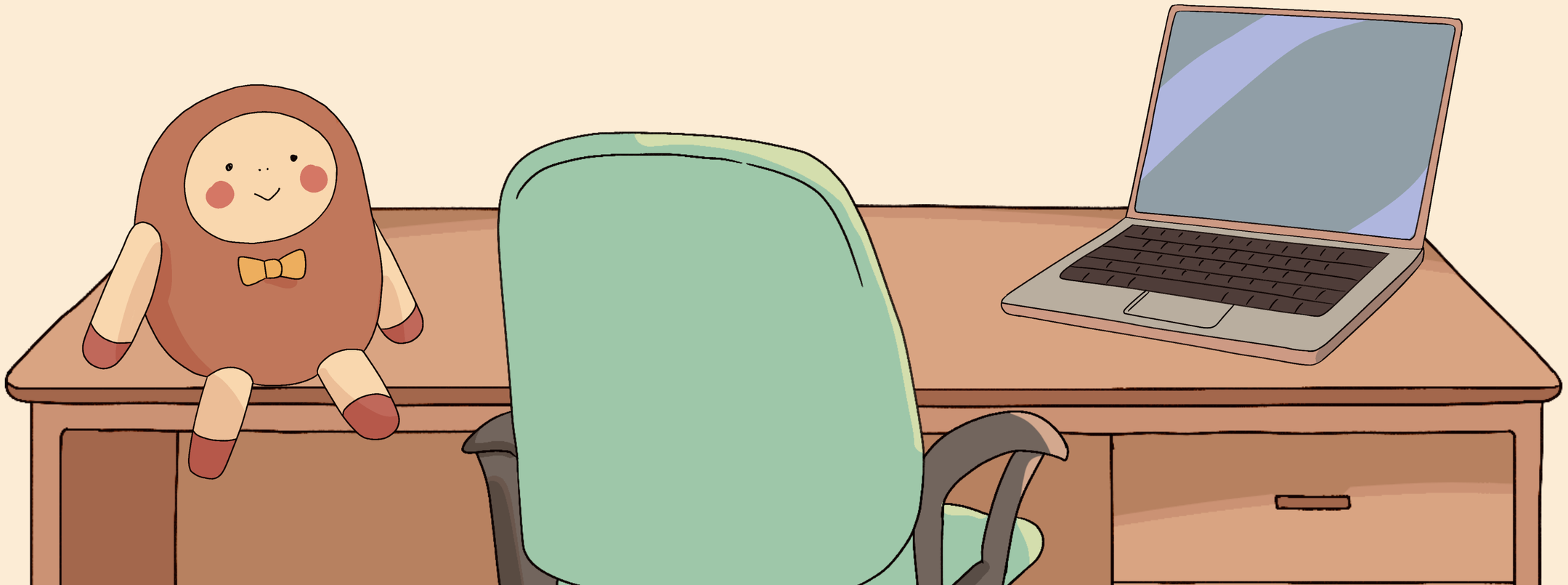
Among 12.901 preprocessed tweets, 5.457 tweets contain hate speech.






Types on Hate Speech Tweets in Indonesia Towards Individual and Group
Group
Individual
2500
2000
1500
1000
500
0
Marks
Other
Physical
Race
Religion
Gender
Types

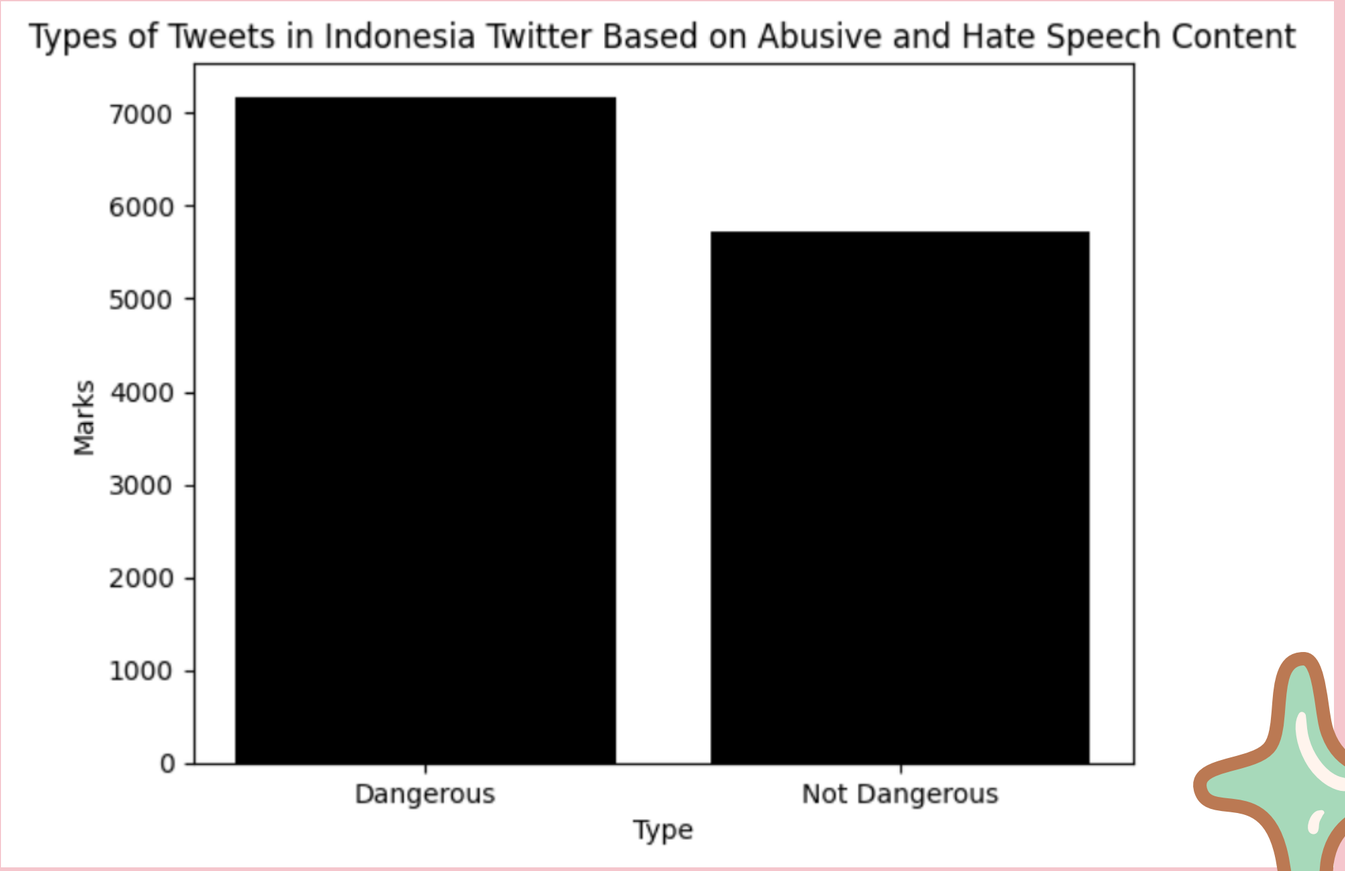
MOST COMMON WORDS




anjing
dasar
kontol
bencong
kafir
banci
banget
sok
cewek
manusia
homo
antek komunis
rezim
partai komunis
usir
negara
kafir
komunis
cina
ganyang partai
cina
partai komunis
rakyat
muka
mata
bolot
cacat otak
idiot
gila
budek
picek
dasar
ganti
cebong
rezim
lengserkan
cebong
rakyat
dewan perwakilan
ganti
goblok
banget
perwakilan rakyat
dasar
kristen
allah subhanahu
islam
anti islam
subhanahu wa
kafir
agama
wa taala
muslim
penista agama

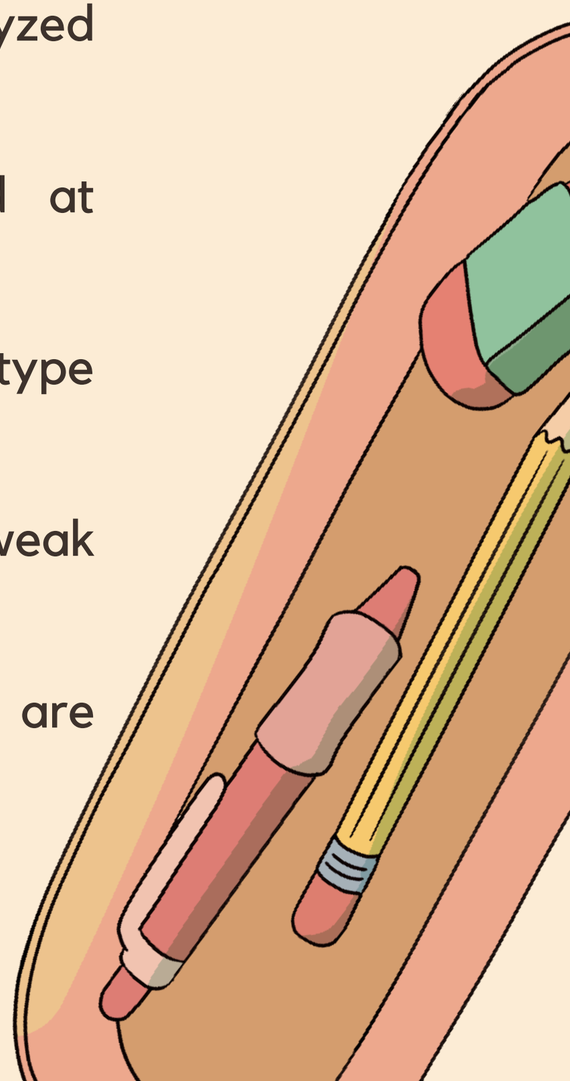
RESULTS

- **Hate speech** that **combine** with **abusive words** often **cause** the happening of **social conflict** because abusive words triggers emotions that not only impact one person but in many cases, certain groups.
- Graph shows 7176 tweets contain hate speech also have abusive words which labeled as dangerous because it could cause social conflict.



SUMMARY

- The percentage of tweets contain hate speech in analyzed data is 42.2%.
- Tweets contain hate speech were mainly directed at individuals instead of groups of people.
- Other type of hate speech was found to be the most type used in tweets contain hate speech.
- Most tweets contain hate speech severity level are weak followed by moderate and strong.
- Most common word in tweets contain hate speech are "cebong", "partai komunis", "cina", "rakyat", and "ganti".
- 55.6% analyzed tweets potentially cause social conflict.



RECOMMENDATIONS

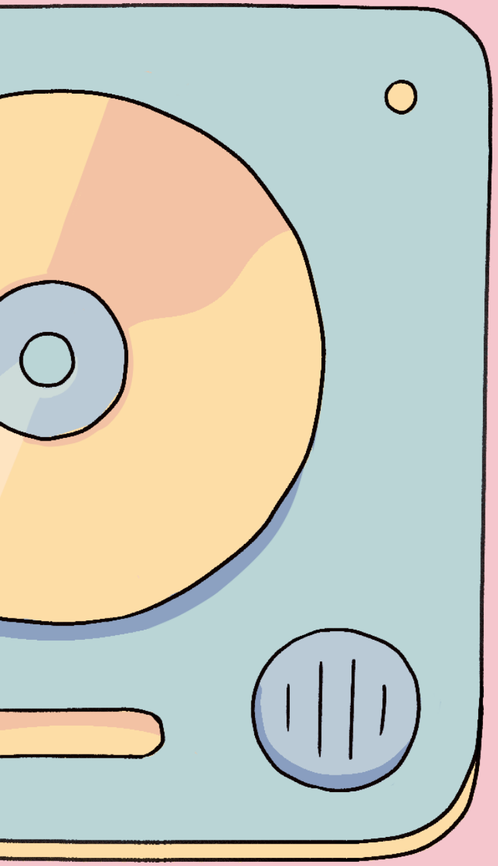
For Internet/Twitter User:

- Don't be the perpetrators of making and spreading hate speech.
- Activate block chain/block party to perform a positive algorithm.
- Realizing social media is a public space not a personal space, so behave and respect others.
- Remember anything shared will automatically become a digital track record.

For Stakeholders:

- Policies towards online free speech need to be re-evaluate and upgrade.
- System improvement such as filtering or banning bad words before it reach a greater mass or certain subject.

REFERENCES



- 25+ Online hate crime statistics and facts. (2022, July 20). Retrieved from comparitech: https://www.comparitech.com/blog/information-security/online-hate-crime-statistics/
- Alnazzawi, N. (2022). Using Twitter to Detect Hate Crimes and Their Motivations: The HateMotiv Corpus. Knowledge Extraction from Data Using Machine Learning, 69.
- Hate speech becomes Indonesia's most frequently reported online crime. (2017, March 26). Retrieved from digwatch: https://dig.watch/updates/hate-speech-becomes-indonesias-most-frequently-reported-online-crime
- Dharmawan, D. A., Sumarwan, P. U., & P Lubis, D. D. (2021, July 26). IPB University Communication Expert: Indonesia's Digital Civility Index is the Worst in Southeast Asia. Retrieved from IPB: https://ipb.ac.id/news/index/2021/07/pakar-komunikasi-ipb-university-digital-civility-index-indonesia-terburuk-di-asia-tenggara/

THANK YOU FOR LISTENING!