

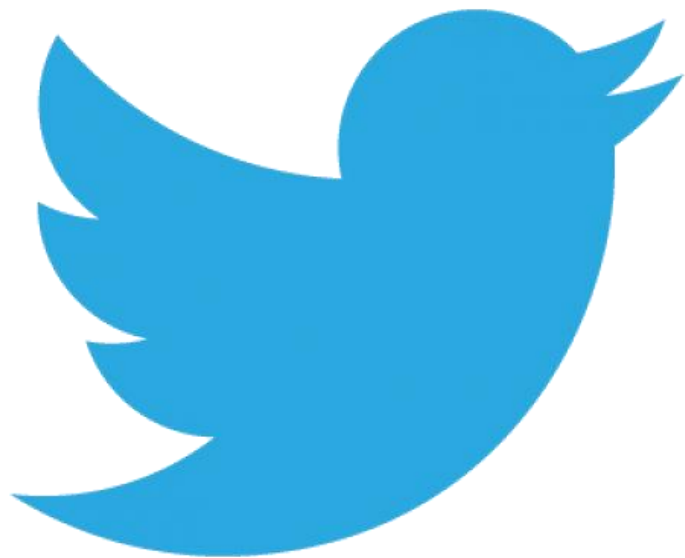


INDONESIAN ABUSIVE AND HATE SPEECH TWEET

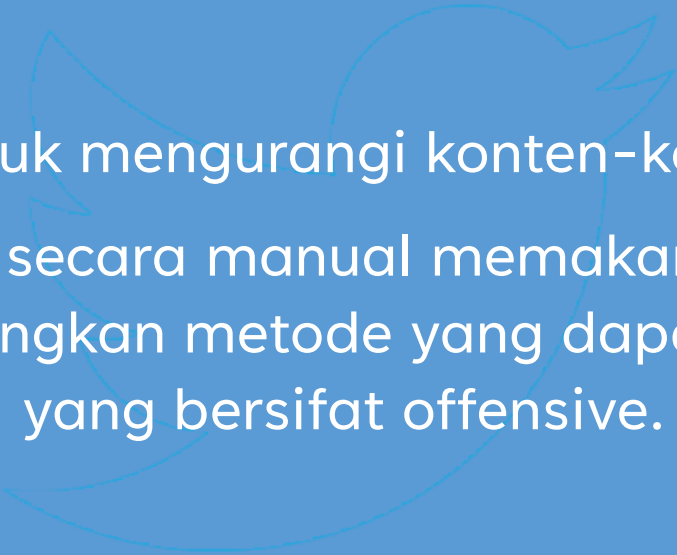
Muhafidz Ahmad Halim

14 Maret 2023

TWITTER



- Tempat dimana semua orang bisa membuat konten berupa tulisan dan membagikannya secara luas. Dari konten edukatif hingga konten hiburan.
- Namun, tidak sedikit orang-orang membuat konten yang bersifat offensive terhadap sesuatu, seperti hate speech.
- Konten tersebut dapat memicu banyak hal negatif lainnya, bahkan bisa menular pada orang lain.

- 
- Diperlukan tindakan untuk mengurangi konten-konten hate speech.
 - Identifikasi hate speech secara manual memakan waktu yang lama, sehingga perlu dikembangkan metode yang dapat secara otomatis menarik konten-konten yang bersifat offensive.

kaggle



ILHAM FIRDAUSI PUTRA · UPDATED 3 YEARS AGO



57

New Notebook

Download (833 kB)



Indonesian Abusive and Hate Speech Twitter Text

Multi-Labeled Hate Speech and Abusive Indonesian Twitter Text by okkyibrohim



<https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text>

13023

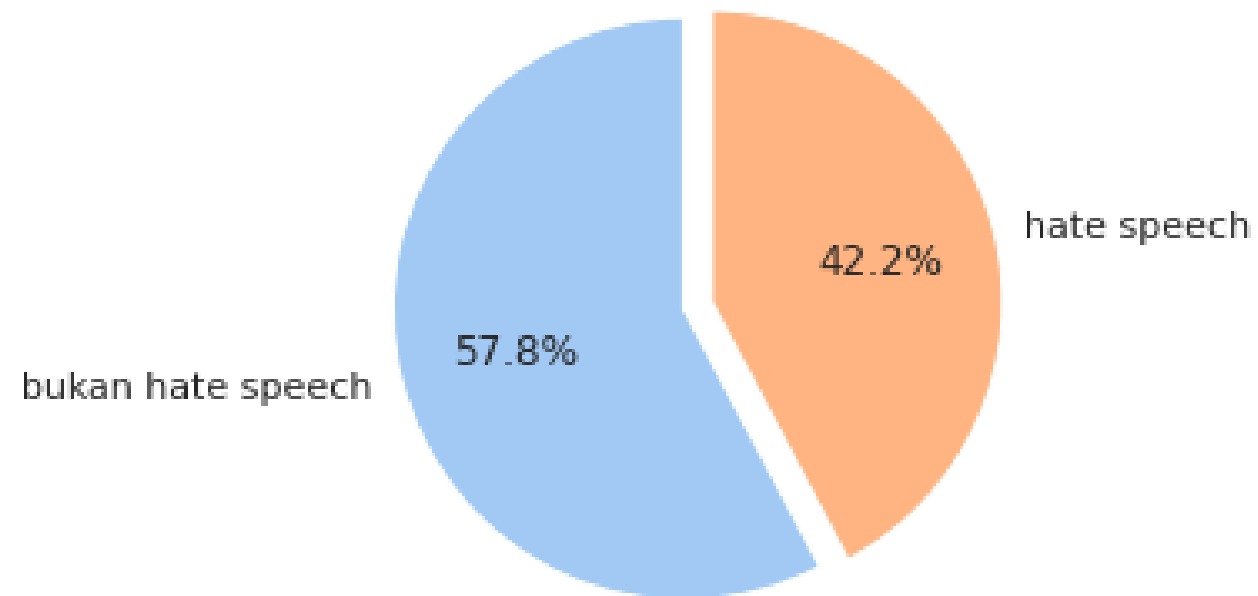
TWEET

*setelah tweet duplikat dihapus

DATA

- Terdiri dari 13 kolom yang terdiri dari:
 - Kolom teks **tweet**,
 - kolom yang menyatakan apakah tweet tersebut termasuk **hate speech atau tidak**,
 - kolom yang menyatakan apakah tweet tersebut termasuk **abusive atau tidak**, dan
 - 10 kolom **jenis hate speech**
- Data Preprocessing:
 - Case folding
 - Menghapus karakter tidak penting
 - Menghapus tanda baca
 - Menghapus stopwords
 - Mengganti kata-kata alay

HATE SPEECH



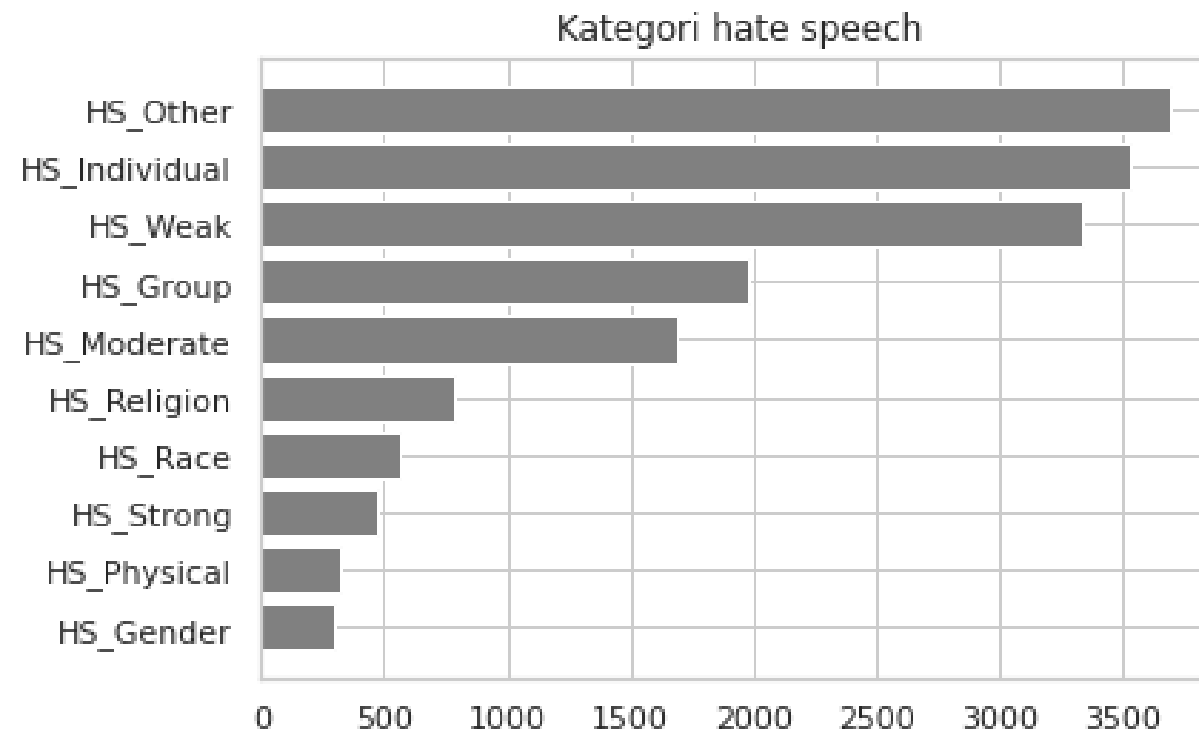
Not HS



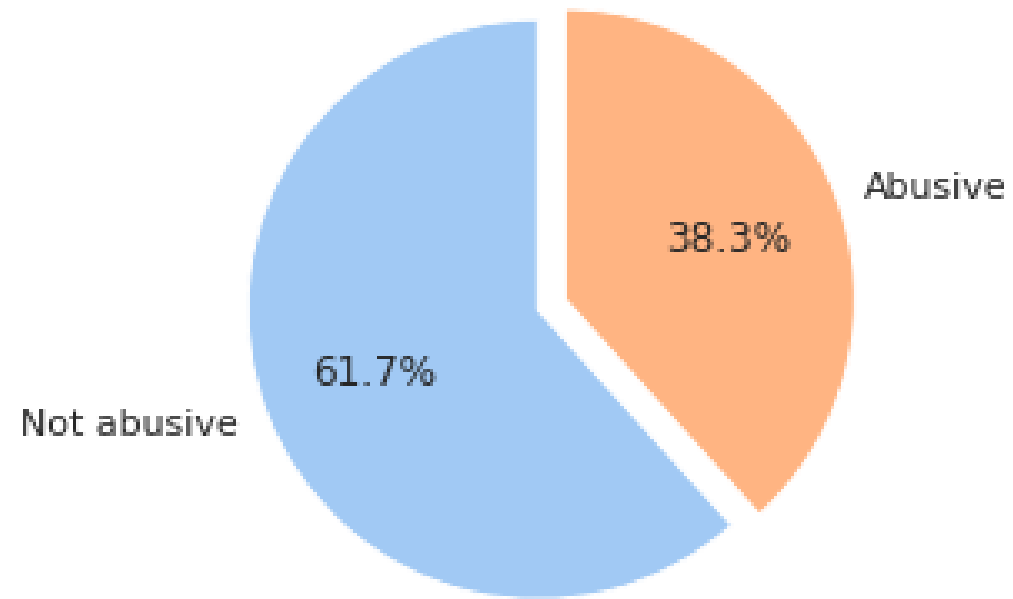
HS



HATE SPEECH

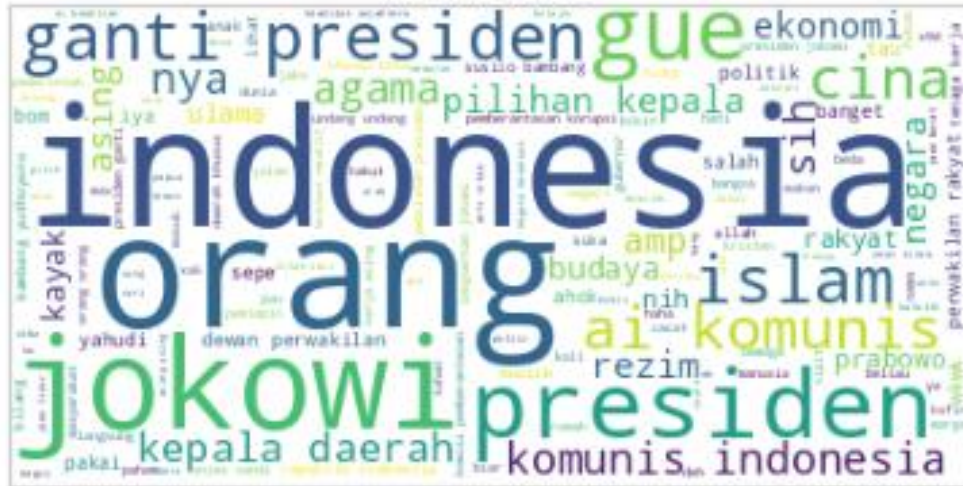


ABUSIVE



ABUSIVE

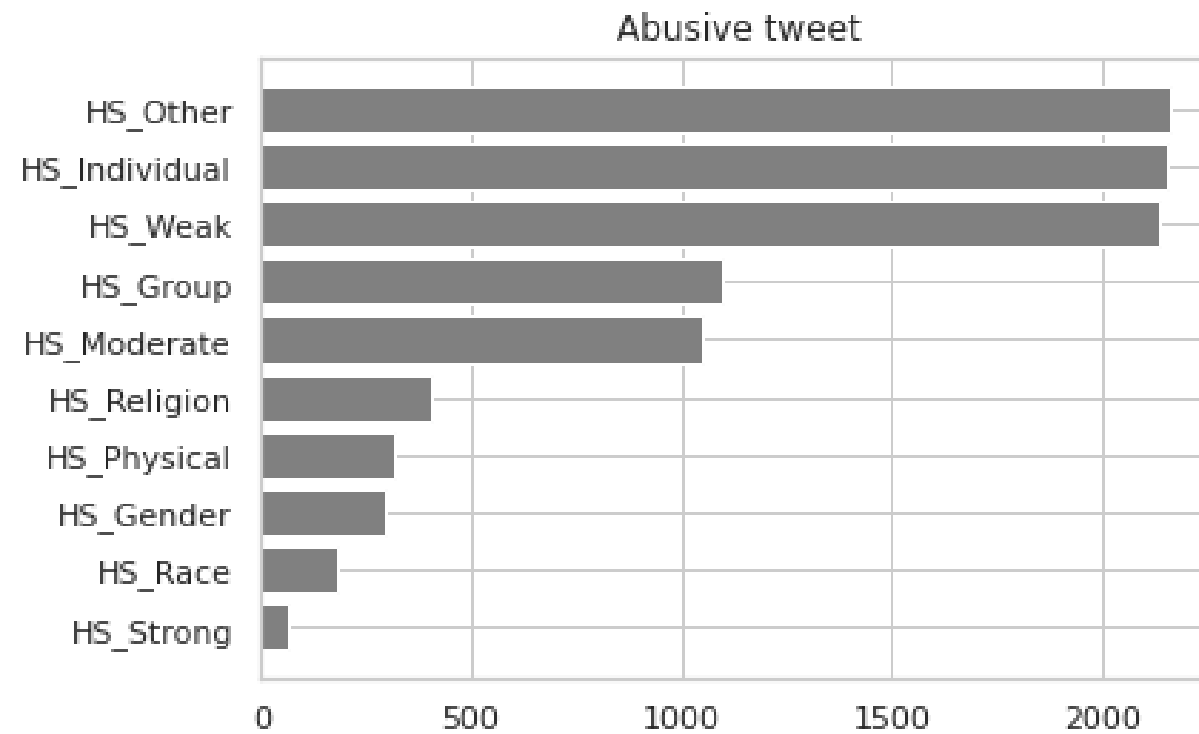
Not Abusive



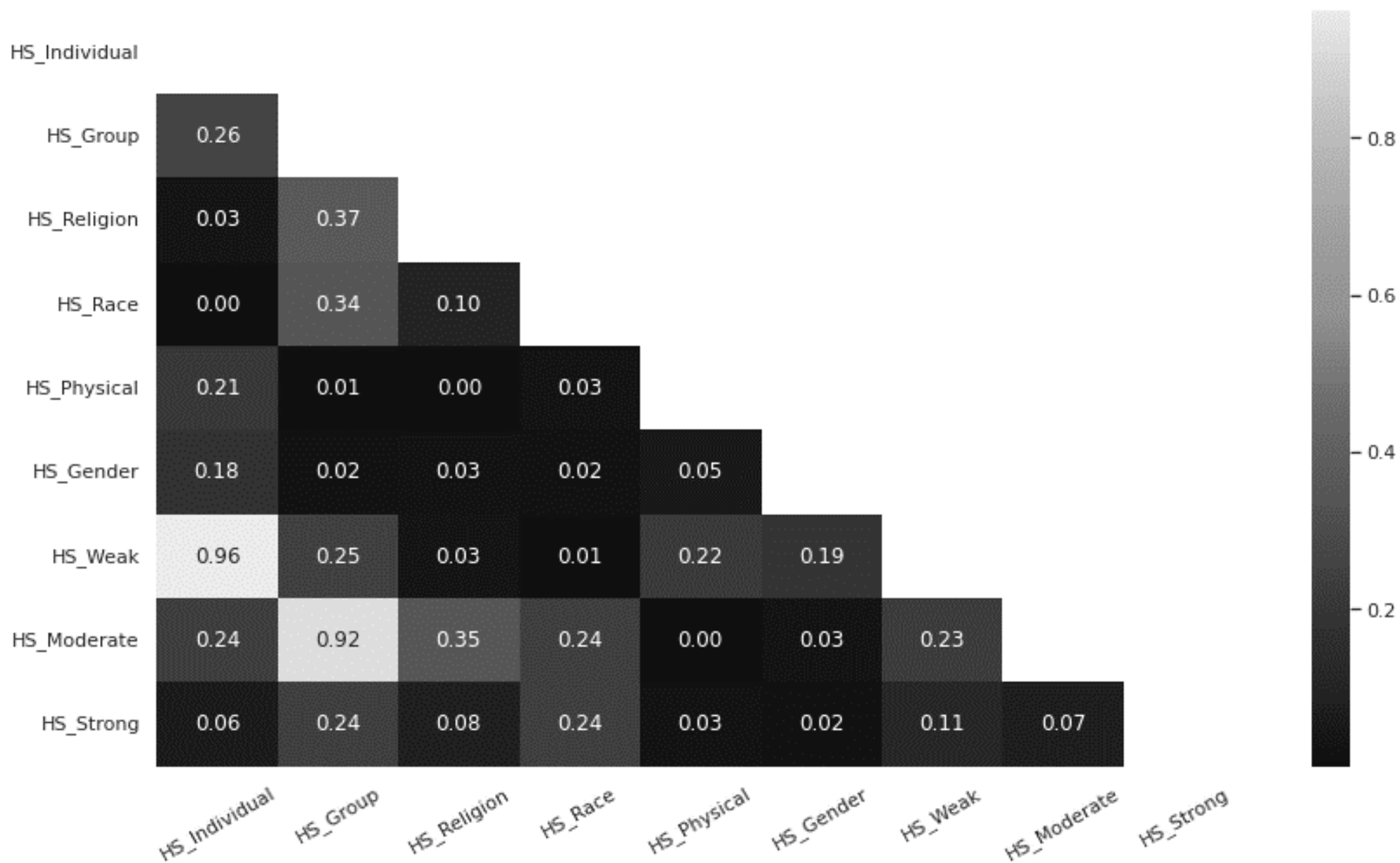
Abusive



ABUSIVE



CORRELATION

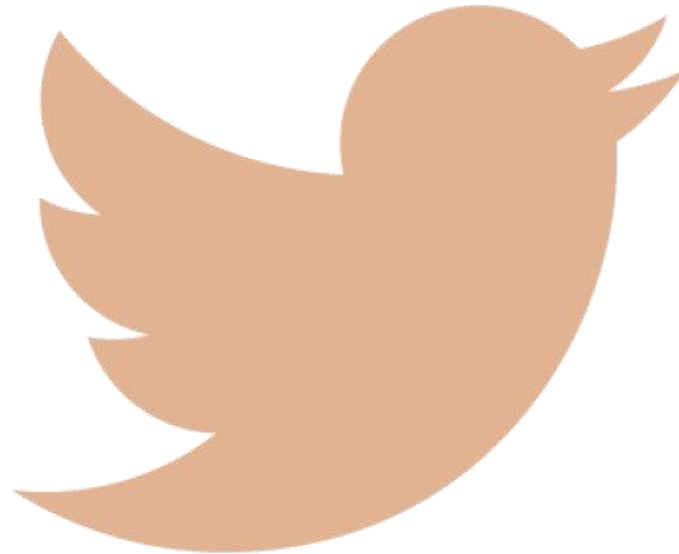


SENTIMENT ANALYSIS

Hate speech and abusive tweet



**Hate speech but
NOT abusive tweet**

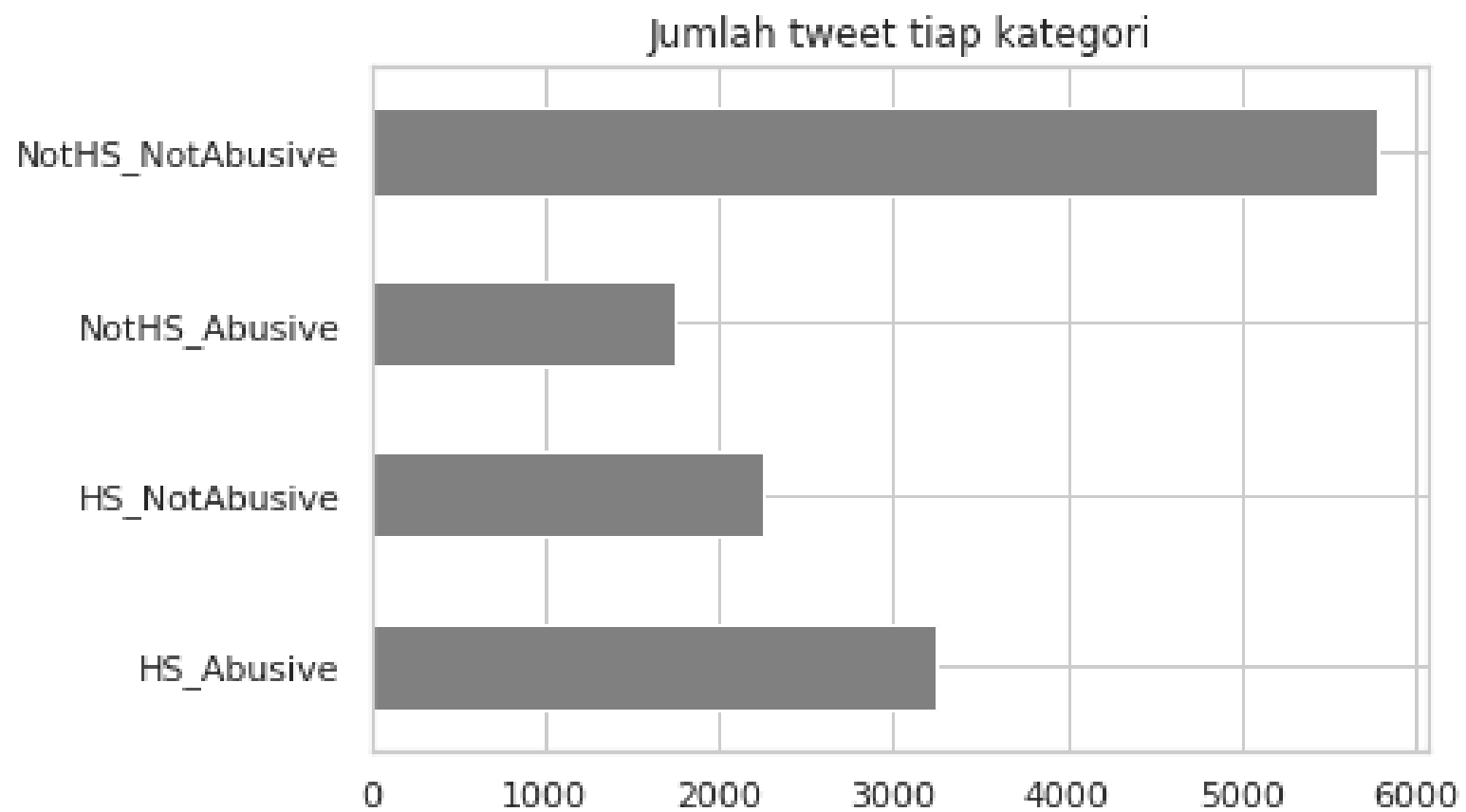


**NOT hate speech but
abusive tweet**



NOT hate speech and NOT abusive tweet





MODELLING

- 1 BiLSTM Layer dengan jumlah hidden state 32 dan dropout 0,5
- 1 Dense Layer dengan jumlah hidden state 16 dan regularisasi l2 0,5 serta fungsi aktivasi ReLU
- 1 Dropout Layer 0,5
- Output Layer dengan fungsi aktivasi softmax
- Optimizer Adam
- Learning rate = 0,001
- Epochs = 45, namun jika akurasi validasi telah mencapai lebih dari 90% atau dalam 5 epoch terakhir tidak ada penurunan nilai loss yang signifikan, proses pelatihan akan langsung berhenti
- Batch size = 256

83,74%

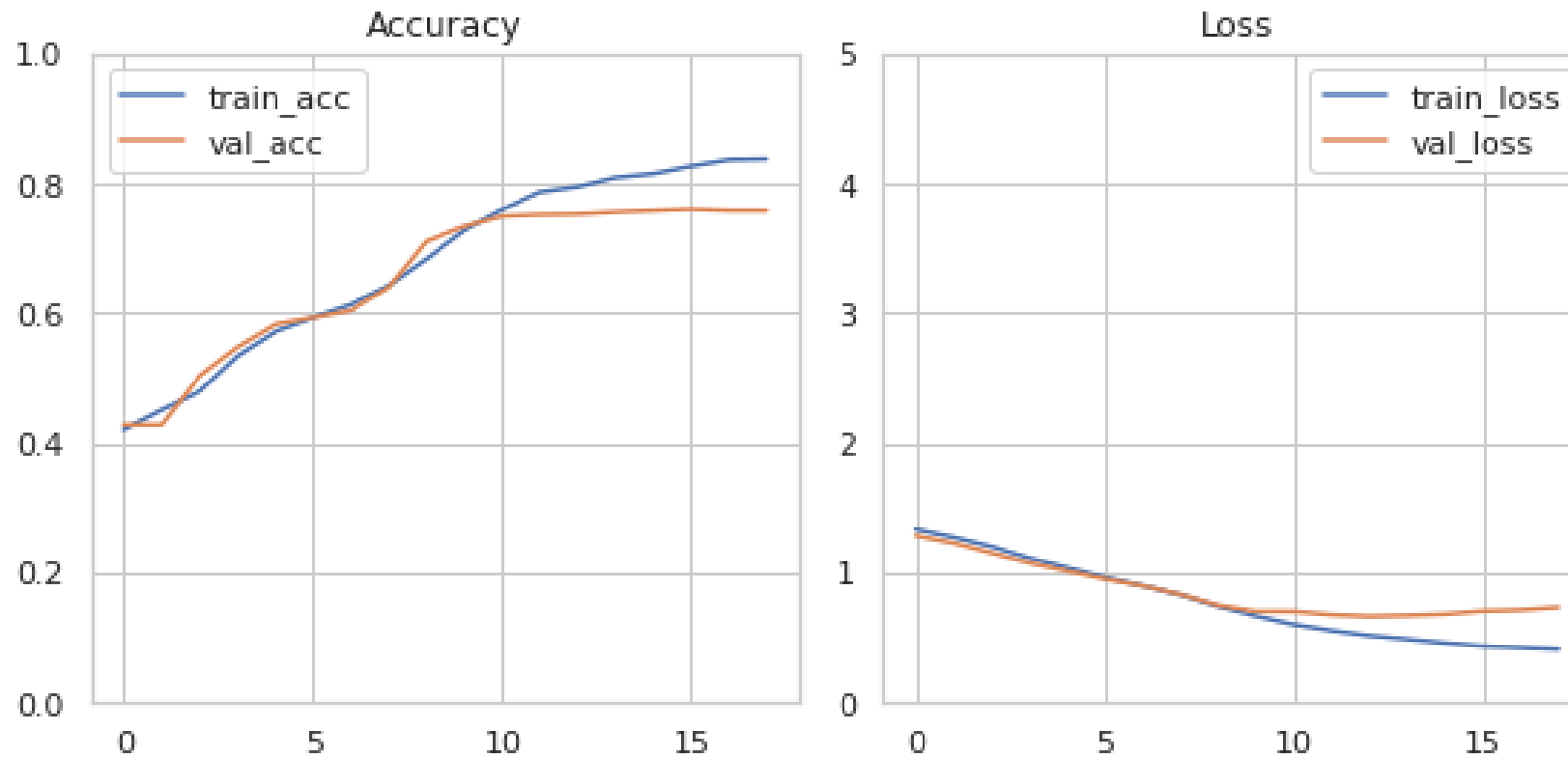
**TRAINING
ACCURACY**

75,78%

**VALIDATION
ACCURACY**

*proses training berhenti di epoch ke-18

EVALUASI TRAINING



CONFUSION MATRIX DATA TEST

HS_Abusive	352	50	55	39
HS_NotAbusive	21	252	1	83
NotHS_Abusive	57	6	162	40
NotHS_NotAbusive	29	71	19	717
	HS_Abusive	HS_NotAbusive	NotHS_Abusive	NotHS_NotAbusive

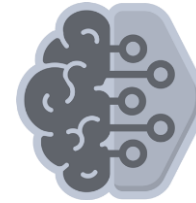
EVALUASI HASIL TEST

	PRECISION	RECALL	F1-SCORE
HS_Abusive	77%	71%	74%
HS_NotAbusive	66%	71%	68%
NotHS_Abusive	68%	61%	65%
NotHS_Not_abusive	82%	86%	84%
Macro_avg	73%	72%	73%
Weighted_avg	76%	76%	76%
ACCURACY	76%		

CONTOH IMPLEMENTASI



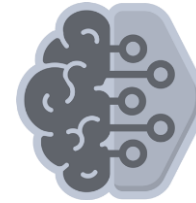
*"ANYS BONEKA AMERIKA JADI KHILAFAH BUATAN AMERICA PAHAM DRUN !!
BUKAN KHILAFAH AJARAN ISLAM"*



Hate Speech but Abusive Tweet



banyak yang bikin ripiw soal oshibe arab2an kemaren, dikarenakan jiwa fomo ini, saya ingin juga meripiw hal teresebut terlepas dari pro&cons nya para wota woti



NOT Hate Speech and NOT Abusive Tweet

KESIMPULAN

- Dari dataset yang digunakan, terdapat hingga 42,2% tweet yang berisi ujaran kebencian. Kata yang paling sering muncul dalam hate speech tersebut terdiri dari “Jokowi”, “cebong”, “komunis”, dan “ganti presiden”.
- 38,3% tweet yang berkonotasi kasar (abusive tweet) menampilkan beberapa kata kasar yang sering, seperti “g*blok”, “anj*ng”, dan “d*ngu”.
- Hate speech tweet dan abusive tweet paling tinggi berada pada kategori *other* dan individu, yang artinya banyak tweet yang menyerang personal orang lain.
- Hate speech individu memiliki korelasi yang kuat dengan hate speech weak (seperti menyindir). Sedangkan hate speech yang ditujukan pada suatu grup memiliki korelasi yang kuat dengan hate speech moderate (seperti mengancam, fitnah, provokasi).
- Model machine learning dapat mendeteksi 76% sentiment tweet dengan benar pada data test, yang sebetulnya masih perlu ditingkatkan lagi.

REKOMENDASI

- Memberikan edukasi yang lebih gencar kepada masyarakat mengenai dampak negatif dan ke-tidak-bermanfaatannya hate speech.
- Menerapkan model machine learning untuk menyaring hate speech tweet. Dengan model yang telah dibuat, setidaknya dapat menyaring 70% tweet yang ada.
- Terus meningkatkan performa model machine learning dengan hyperparameter tuning lebih lanjut dan memperkaya data hingga dapat diperoleh data yang seimbang.
- Memantau secara berkala model machine learning yang telah diterapkan, dan juga memantau tren dan perubahan dalam penggunaan bahasa dan perilaku netizen terkait hate speech.



SEKIAN, TERIMA KASIH

Muhafidz Ahmad Halim

muhafidz.ahmad@gmail.com

<https://www.linkedin.com/in/muhafidz-ahmad-halim/>