



ANALISIS SENTIMEN KOMENTAR YOUTUBE TERHADAP BENCANA DI SUMATRA

Raihan Pratama



**Turut berduka cita atas kejadian yang terjadi di provinsi
Sumatra belakangan ini, semoga semua bisa secepatnya
kembali ke kondisi normal, aamiin**



Table of Content

Problem & Latar Belakang

Tools

Scraping & Preprocessing

Memberikan Label Data

Exploratory Data Analysis (EDA)

Train & Tuning Model

Example & Deployment

Conclusion

Problem & Latar Belakang



Bencana yang belakangan ini terjadi di provinsi Sumatra memicu berbagai reaksi publik di media sosial, khususnya di YouTube. Namun, opini masyarakat tersebar dalam ribuan komentar yang sulit dianalisis secara manual. Project ini bertujuan untuk menganalisis sentimen masyarakat terhadap bencana banjir di Sumatra melalui komentar YouTube untuk memahami respons publik secara emosional (positif, negatif, netral) menggunakan pendekatan Natural Language Processing (NLP)

TOOLS



seaborn



pandas



Hugging Face



gradio

Scraping dan Preprocessing

Informasi Data

- Channel : Nama saluran youtube
- Likes : Jumlah like di suatu komentar
- Comment: Komentar

	channel	likes	comment
0	Ferry Irwandi	4K	Lekas pulih Bumi Andalaskul!
1	Ferry Irwandi	0	Prabowo dari dulu jg punya lahan sawit d Sumat...
2	Ferry Irwandi	0	Lahir di Jambi asal payuhkumbauh
3	Ferry Irwandi	0	👉👉👉
4	Ferry Irwandi	0	PROVOKATOR SESUNGGUHNYA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27433 entries, 0 to 27432
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   channel   27433 non-null   object  
 1   likes     27433 non-null   object  
 2   comment   27433 non-null   object  
dtypes: object(3)
memory usage: 643.1+ KB
```

	channel	count
0	CURHAT BANG Denny Sumargo	14761
1	Ferry Irwandi	6154
2	Sepulang Sekolah	3674
3	Kamar JERI	2844

Dataset didapatkan menggunakan Apify

Dataset ini berisi informasi nama channel, komentar dan jumlah like, data diambil dari 4 jenis channel berbeda yang membahas kenapa Sumatra tidak ditetapkan sebagai bencana nasional, total data yang berhasil diambil sekitar 27.433 baris

Data duplikat dan data hilang

setelah dilakukan pengecekan terdapat 1137 data duplikat dan tidak ada data yang hilang, data duplikat diatasi dengan cara menghapus data tersebut karena jumlah nya yang termasuk kecil dibandingkan total data

```
# lihat komentar yang duplikat
df['comment'].duplicated().sum()

np.int64(1137)
```

```
# buang data duplikat
df.drop_duplicates(subset=['comment'], keep='first', inplace=True)
df['comment'].duplicated().sum()

np.int64(0)

df.isnull().sum()

0
channel 0
likes 0
comment 0
```

Preprocessing

Cleaning

Membersihkan data komentar dari url, username, emoji, simbol dan angka

Case Folding

Mengubah kalimat menjadi huruf kecil semua agar konsisten dan meningkatkan pemahaman konteks

Tokenization

Memecah kalimat menjadi kata-kata atau token individual untuk diproses lebih lanjut oleh model.

Remove Stopword

Menghapus kata umum yang tidak membawa makna penting seperti 'dan', 'yang', 'di'

Normalisasi

Mengubah kata yang tidak baku, gaul atau slang menjadi kata baku agar mudah dipahami

Stemming

Cara untuk memotong imbuhan (awalan, akhiran) dari kata-kata agar kembali ke bentuk dasarnya (kata dasar) contoh "berlari", "pelari", "larinya" menjadi "lari"

Preprocessing

Sebelum Stopword diterapkan, kata yang tidak membawa makna yang berarti terlihat dominan tapi setelah Stopword dilakukan hasil menunjukan kata yang muncul hanya kata yang benar benar memiliki makna

Stopword



Preprocessing

Stopword dilakukan 2 kali agar lebih bersih yaitu pertama dilakukan menggunakan sastrawi dan dilakukan lagi menggunakan sastrawi + custom, terlihat bahwa setelah dilakukan stopword yang kedua kata 'kalau', 'semua', 'sama' dan kata yang tidak membawa makna lainnya sudah terhapus

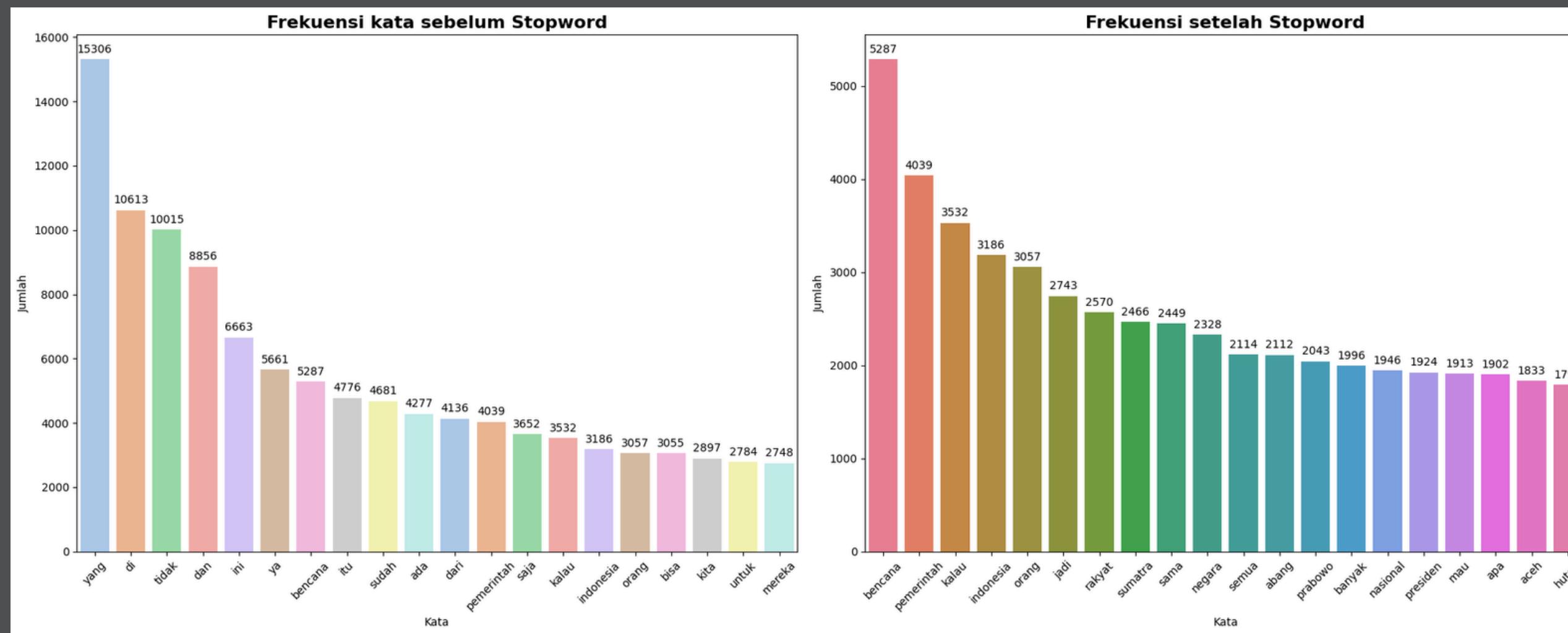
Stopword + Custom



Preprocessing

Hasil juga terlihat setelah di visualisasikan frekuensi kata sebelum dan sesudah stopword

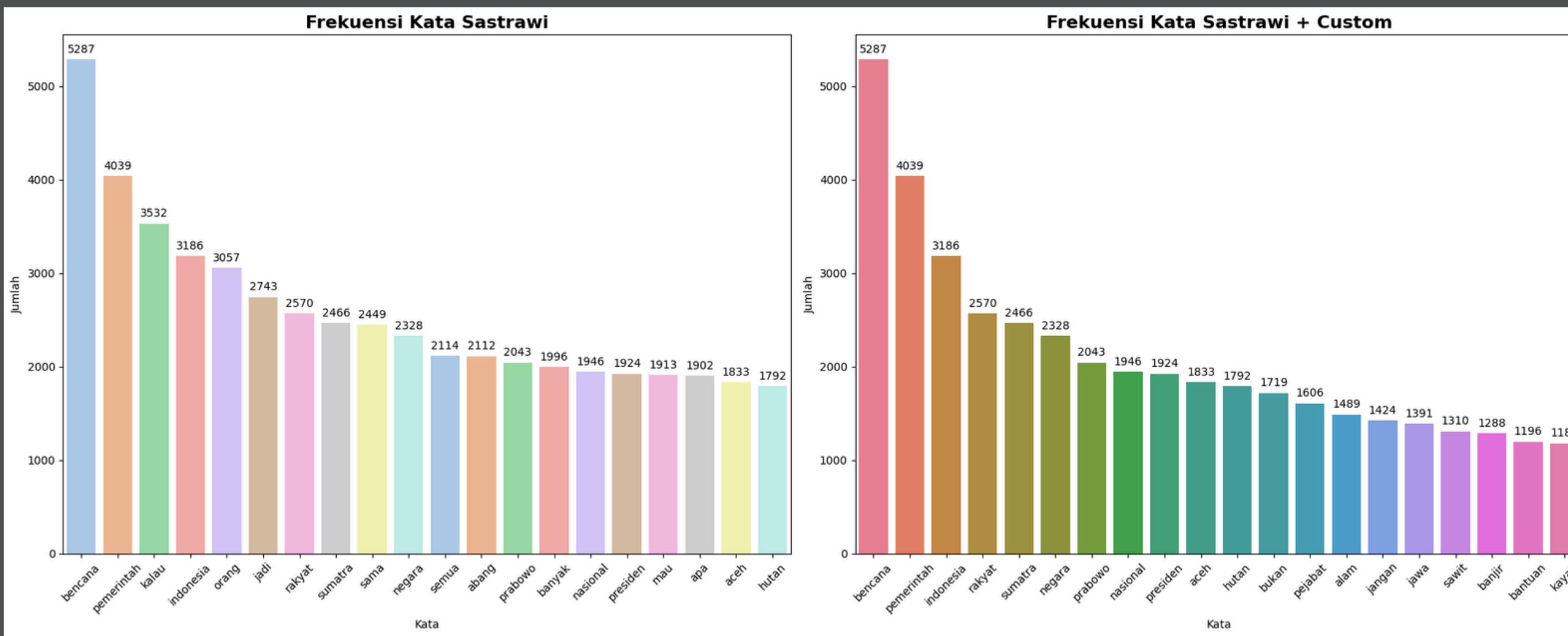
Stopword



Preprocessing

Hasil juga terlihat setelah di visualisasikan frekuensi kata stopword biasa dan stopword + custom

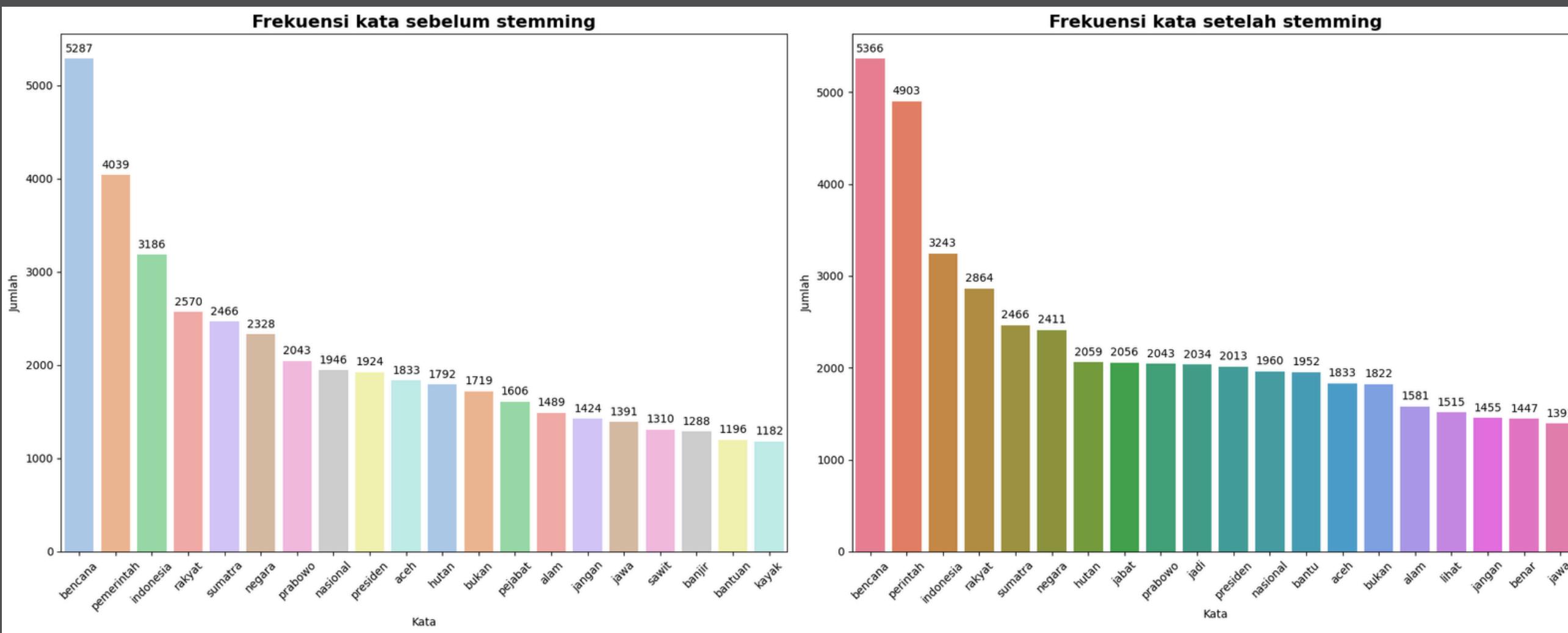
Stopword + Custom



Preprocessing

Hasil juga terlihat setelah di visualisasikan frekuensi kata sebelum stemming dan sesudah stemming

Stemming



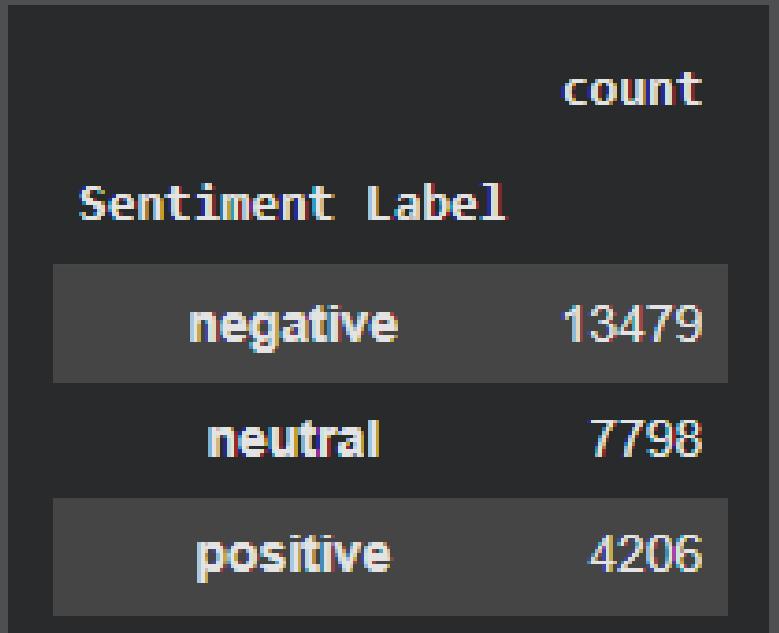
Memberikan Label Data

Memberi Label

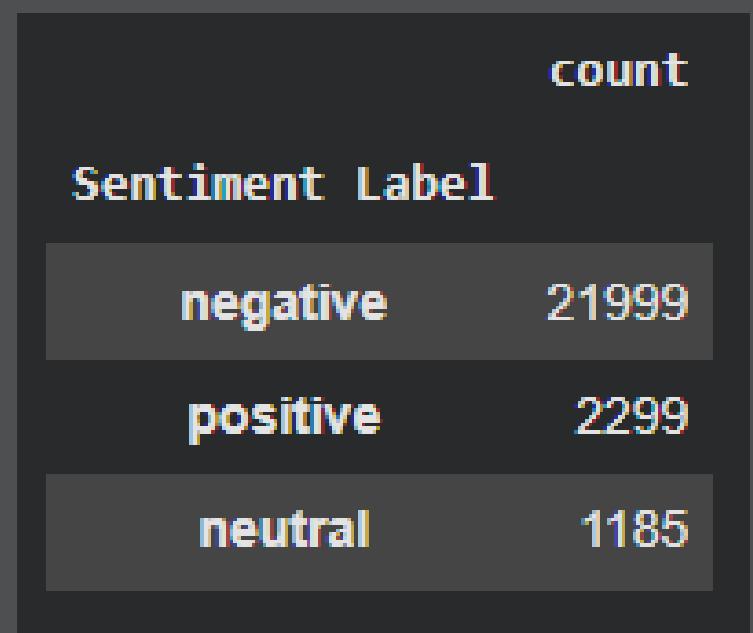
Pelabelan data menggunakan IndoBERT, lebih tepatnya w11wo/indonesian-roberta-base-sentiment-classifier menggunakan 2 kolom yaitu stemming dan sebelum stemming

karena sebelum stemming semua kata tidak berubah kembali ke bentuk dasar maka model akan banyak memberi label negatif lalu jika stemming semua kata akan kembali ke bentuk dasarnya maka hasil label nya akan lebih seimbang

Stemming



Sebelum Stemming



saya memutuskan untuk memakai hasil label stemming karena akan mendapatkan hasil yang lebih seimbang

Exploratory Data Analysis (EDA)

Hapus baris sentimen score rendah

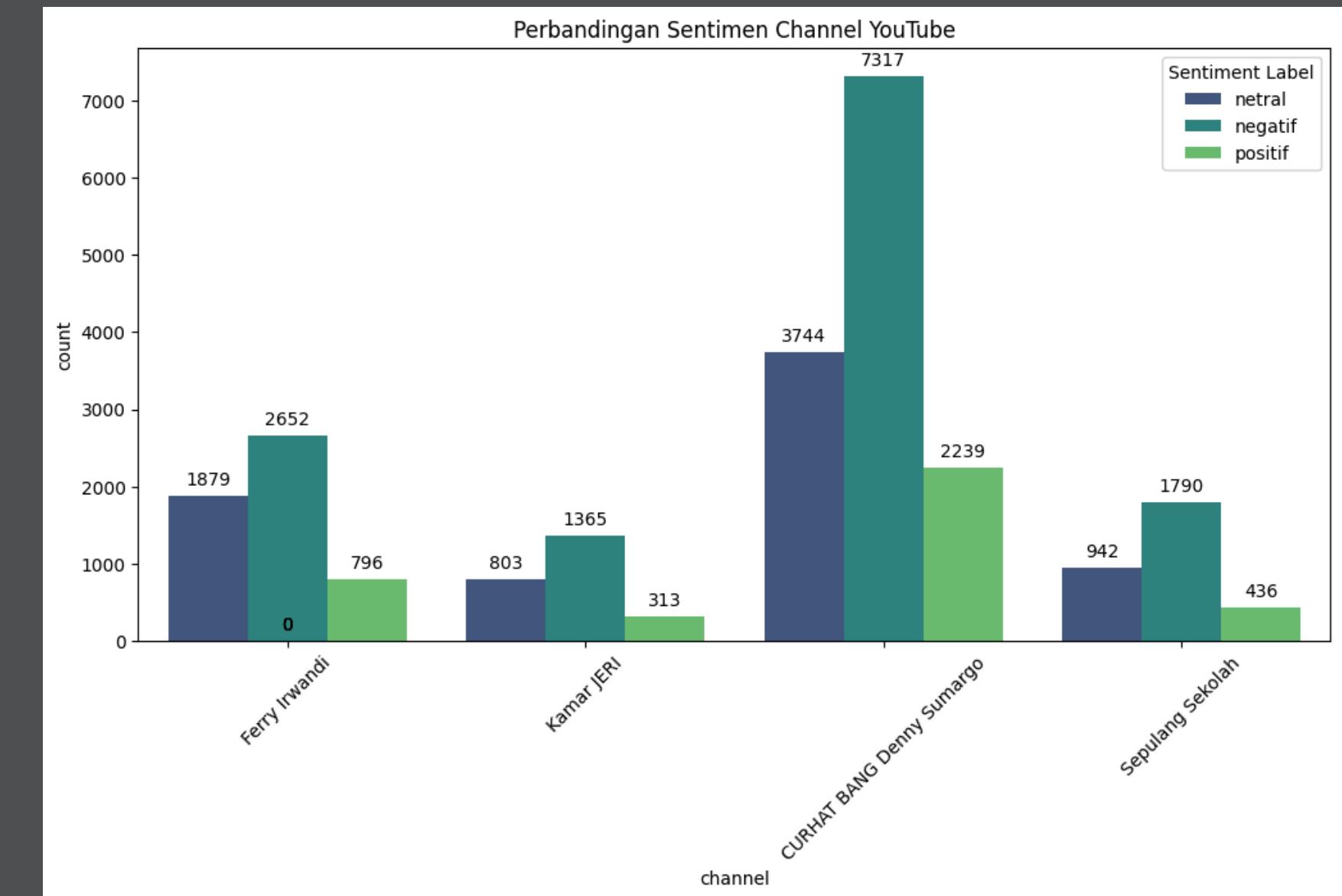
Berikut terdapat 1.207 baris data yang sentimen score nya rendah yaitu dibawah 60%

Setelah di deteksi baris yang memiliki sentimen score dibawah 60% lalu diputuskan untuk dihapus saja karena secara hasil memang benar model tetap memberikan label di baris data tersebut tapi model tidak cukup yakin bahwa data di baris tersebut label yang dia berikan benar

Jumlah data awal:	25483
Jumlah data setelah filter:	24276
Data yang dibuang:	1207
Sebaran Label Setelah Filter:	
count	
Sentiment Label	
negatif	13124
netral	7368
positif	3784

Visualisasi

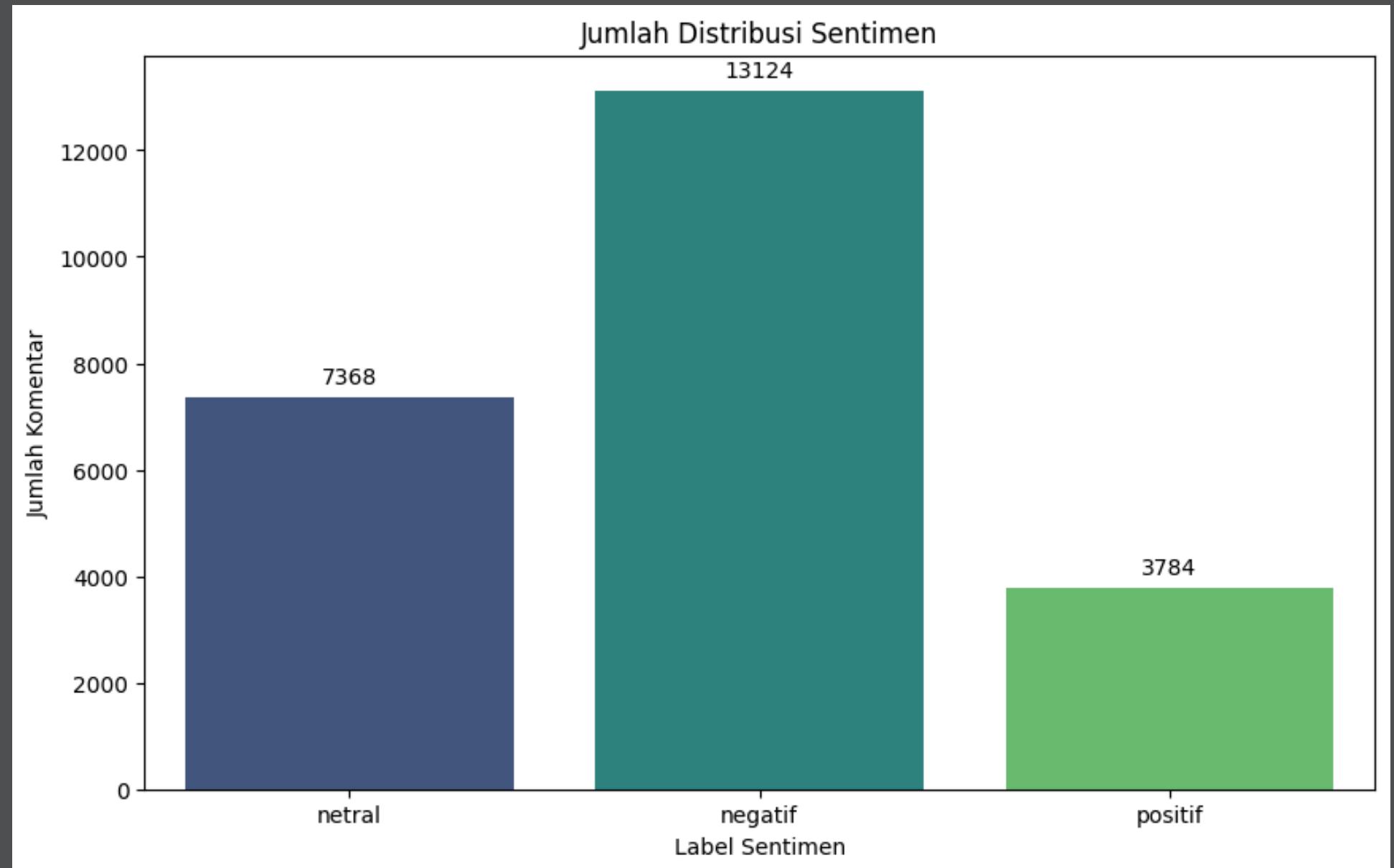
Selanjutnya kita melihat perbandingan hasil sentimen di 4 channel berbeda, terlihat bahwa channel CURHAT BANG Denny Sumargo memiliki jumlah data yang paling banyak dan paling banyak juga sentimen negatif nya diikuti dengan channel Ferry Irwandi dan 2 channel sisa nya yang punya data mirip mirip



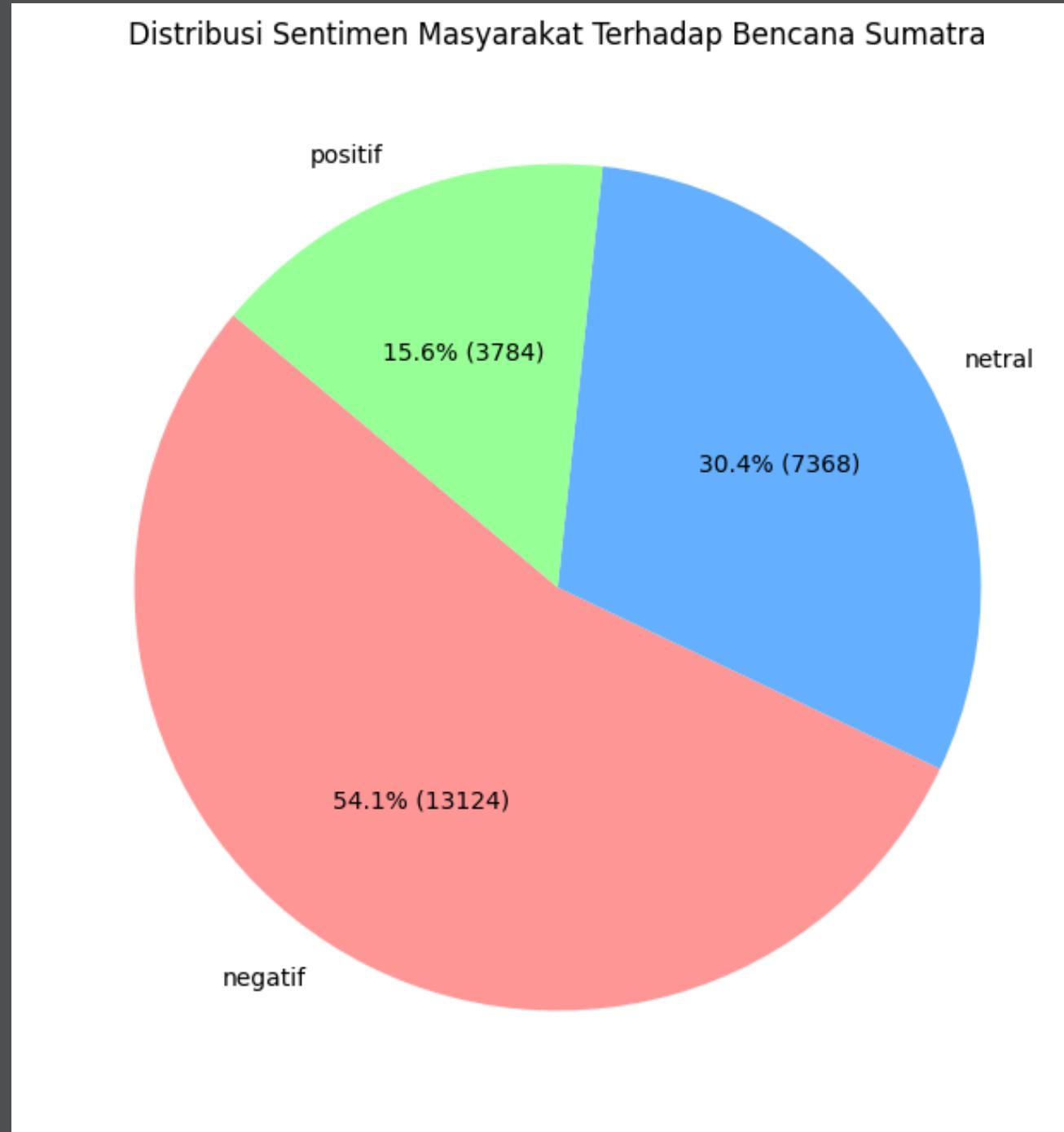
Visualisasi

Selanjutnya kita melihat perbandingan distribusi sentimen dari 4 channel tersebut yang mana hasilnya lebih banyak sentimen negatif dari masyarakat lalu total sentimen negatif hampir 2 kali lipat sentimen netral dan hampir 4 kali lipat sentimen positif

ini menunjukan bahwa banyak masyarakat yang merasa kecewa, marah atau kritis yang langsung diluapkan dengan berkomentar

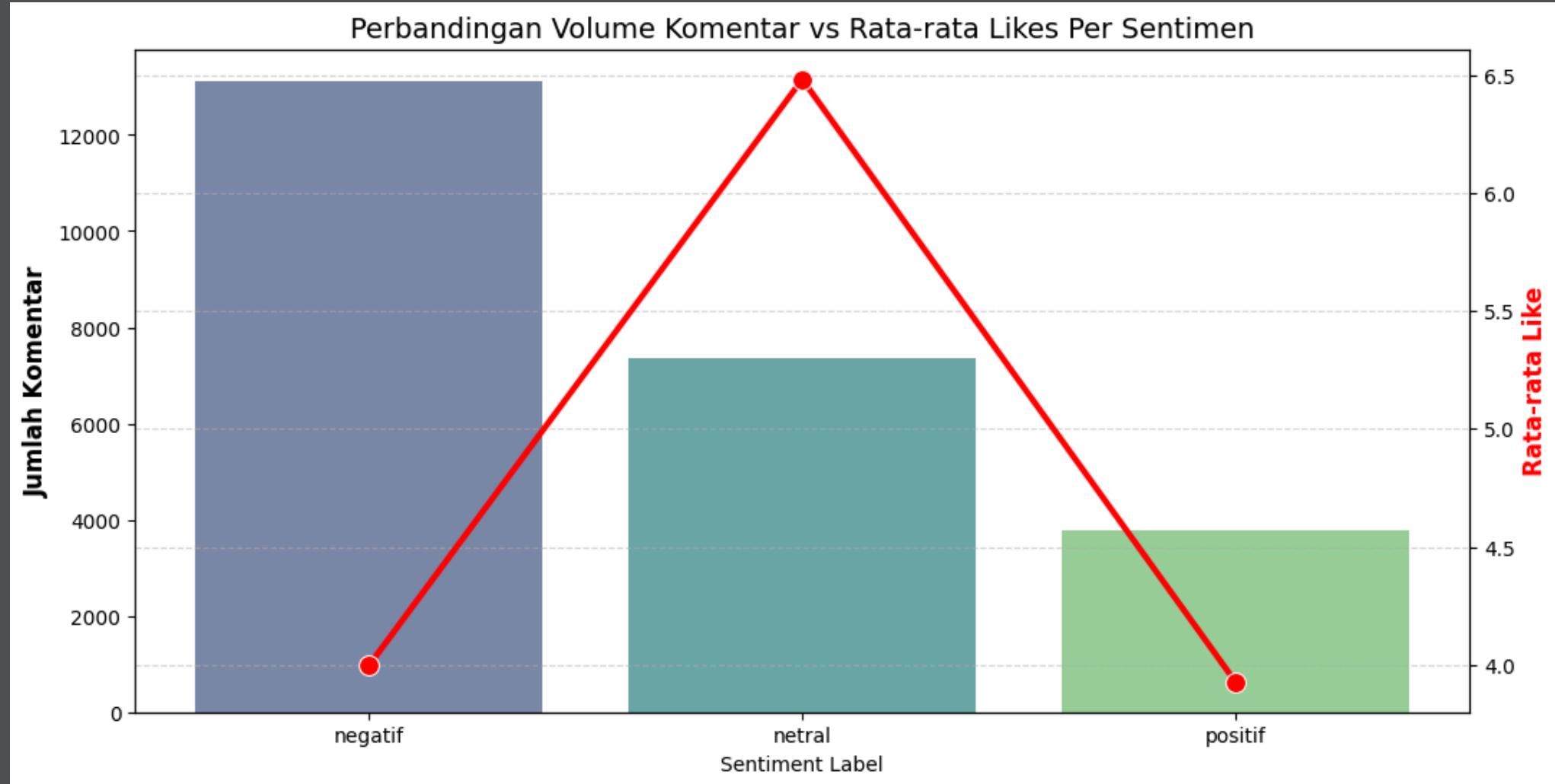


Visualisasi



Kemudian jika kita melihat perbandingan persentase lebih dari setengah komentar di 4 channel tersebut bernilai negatif yaitu 54% diikuti dengan netral 30% dan positif 15%

Visualisasi

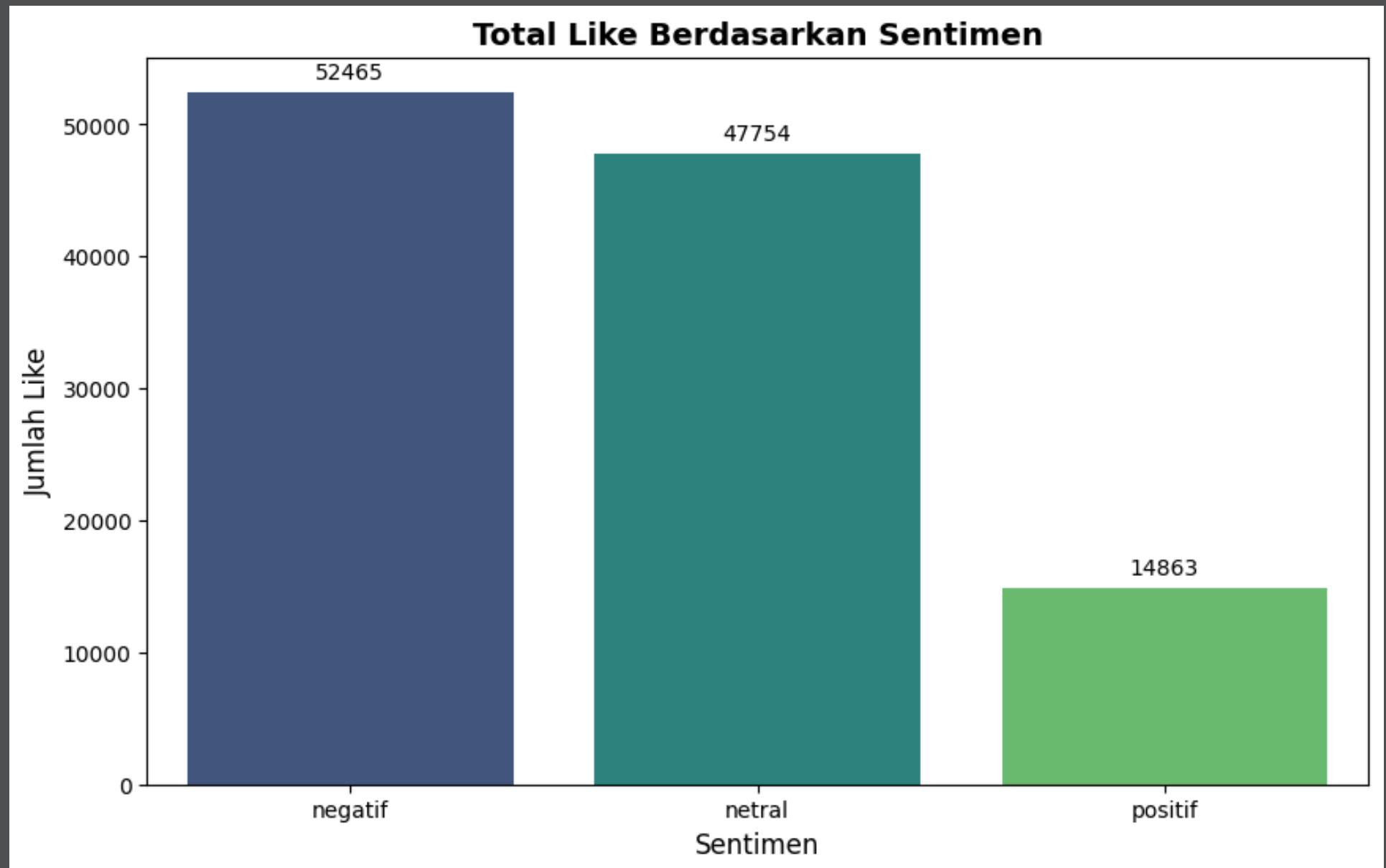


Selanjutnya kita coba lihat perbandingan jumlah komentar di masing masing sentimen dengan rata rata like, terlihat bahwa meskipun komentar sentimen negatif memiliki jumlah yang paling banyak tapi sentimen komentar netral ternyata memiliki rata rata like lebih banyak

ini menunjukan bahwa meskipun banyak masyarakat yang merasa kecewa, marah atau kritis yang langsung diluapkan dengan berkomentar tapi ada segelintir orang yang tidak ikut menulis komentar tapi meraka setuju dengan komentar yang bersifat netral

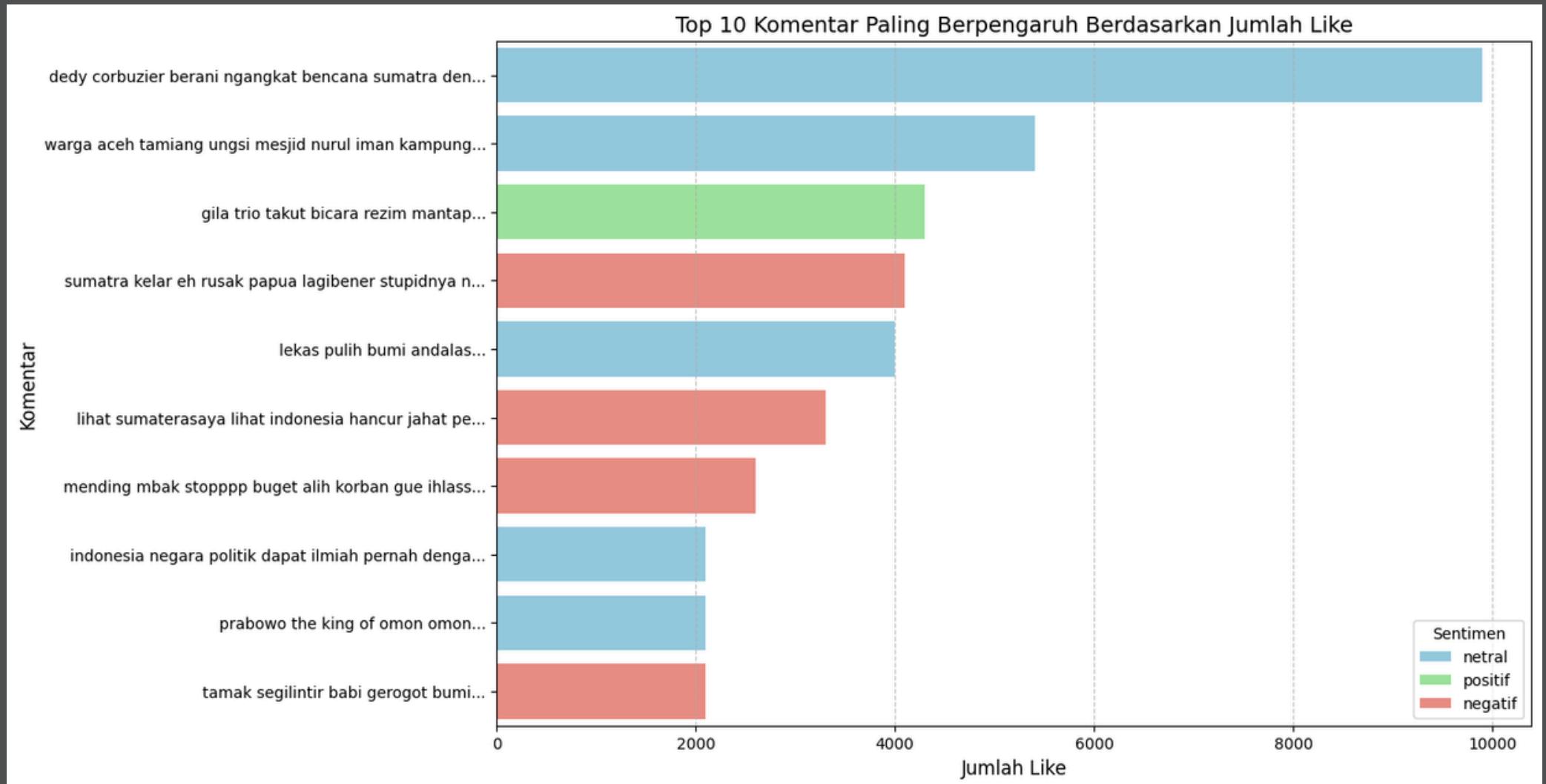
Visualisasi

Selanjutnya kita coba lihat perbandingan jumlah like di masing masing sentimen, terlihat bahwa sentimen negatif memiliki like terbanyak karena komentar dengan sentimen negatif tidak semua nya di like oleh masyarakat tapi banyak like yang menumpuk di beberapa komen saja diikuti dengan netral dan positif

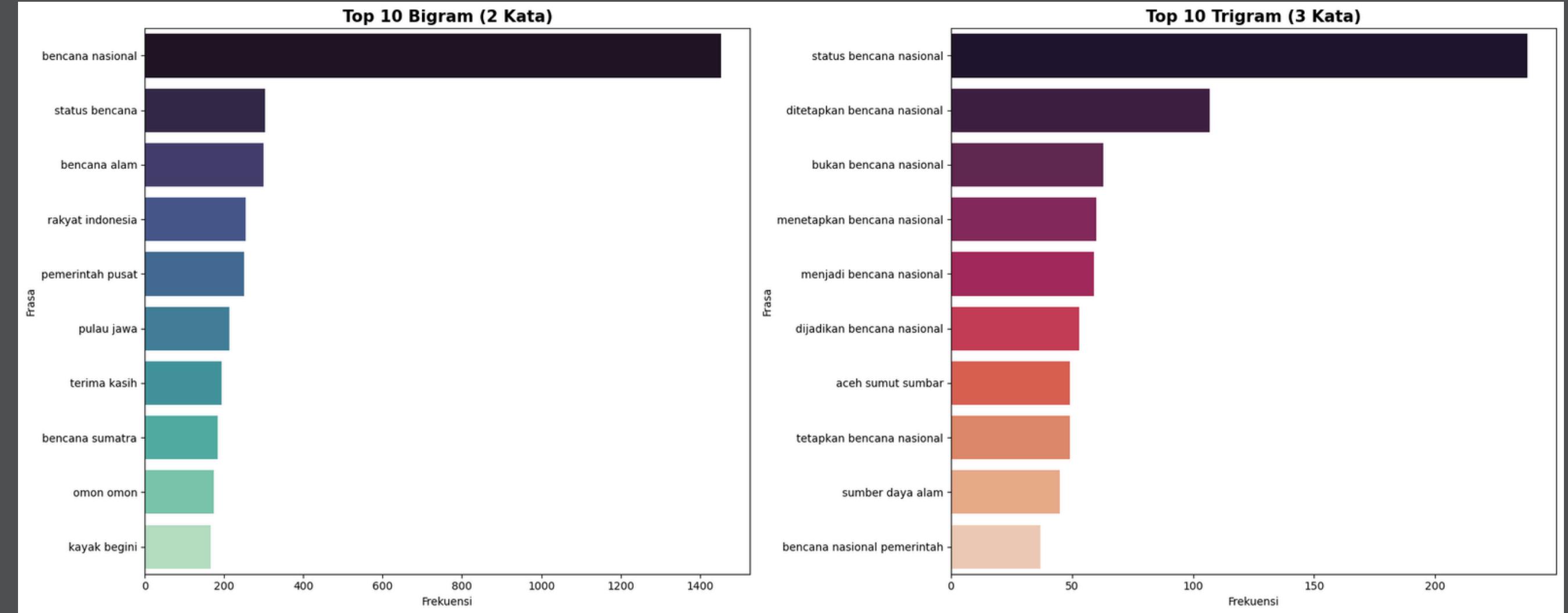


Visualisasi

Selanjutnya kita coba lihat top 10 komentar apa saja di masing masing sentimen yang memiliki like terbanyak, terlihat bahwa 2 sentimen netral menduduki peringkat 1 dan 2 diikuti sentimen 1 positif dan 1 negatif



Visualisasi



Selanjutnya kita coba lihat Bigram (2 kata) dan Trigram (3 kata) yang paling sering muncul, terlihat status bencana di Bigram dan status bencana nasional di Trigram menduduki peringkat teratas dengan gap yang cukup signifikan ke nomor selanjutnya, lalu di Bigram peringkat 2 dan 3 masih tentang bencana kemudian diikuti dengan kata lain, selanjutnya untuk Trigram 6 peringkat teratas masih berisi kata bencana nasional diikuti dengan kata lainnya dan jika ditotal top 10 Trigram hampir semua nya ada kata bencana nasional

Ini menandakan bahwa desakan dari masyarakat tentang bencana sumatra ini untuk segera dijadikan bencana nasional semakin masif, lambat nya bantuan yang bisa sampai ke lokasi bencana dan parahnya dampak bencana yang ditimbulkan yang kemungkinan besar menjadi penyebab banyak nya komentar tersebut

Train & Tuning Model

Jenis Model

Logistic Regression

Random Forest

GRU (Gated Recurrent Unit)

LSTM (Long Short-Term Memory)

SVC

Naive Bayes

Hasil

Untuk hasil yang lebih balance diputuskan untuk melakukan downsampling ke jumlah data paling sedikit

Sentiment Label	count
negatif	13124
netral	7368
positif	3784

Sentiment Label	count
negatif	3784
netral	3784
positif	3784

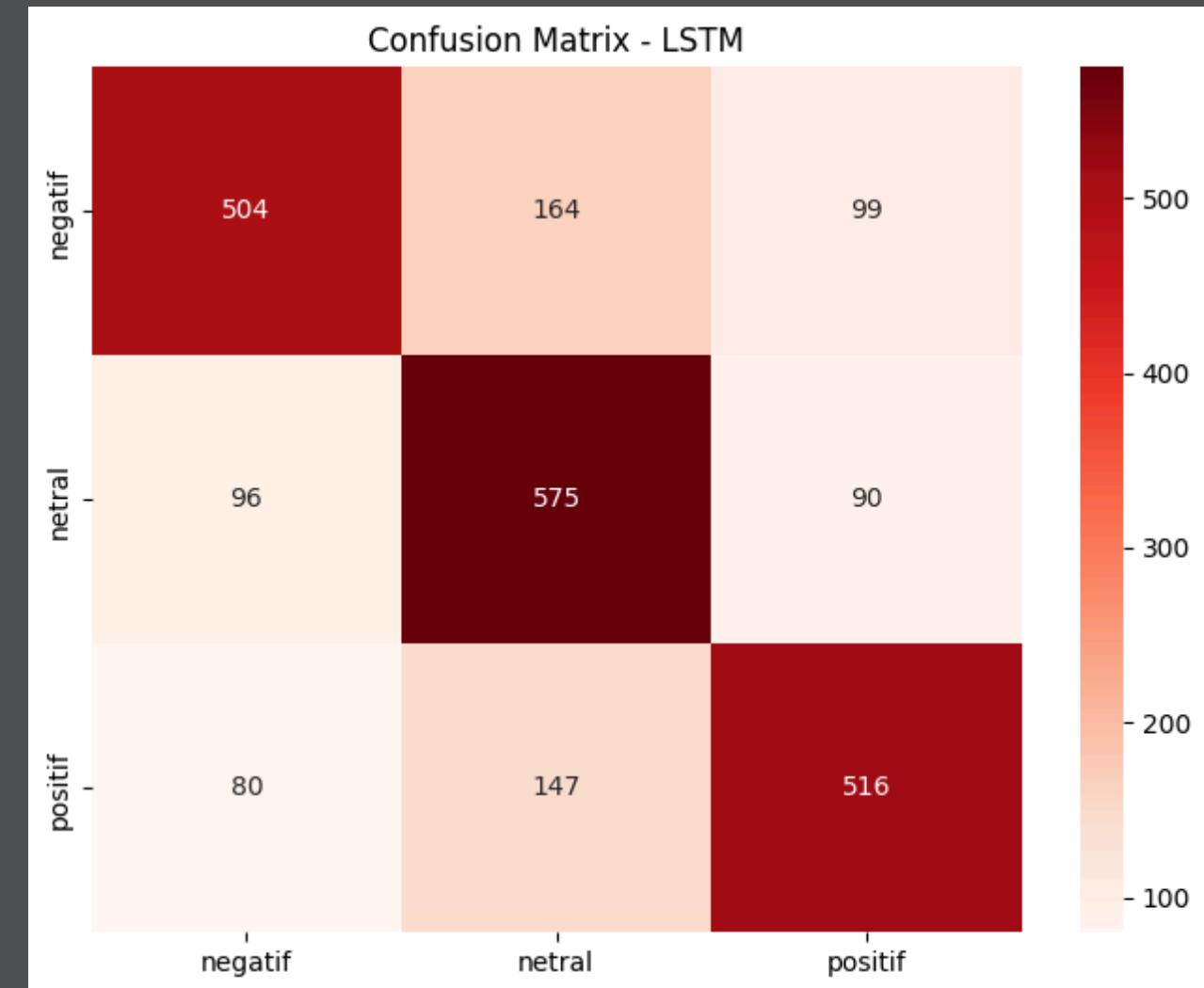
Hasil

Perbandingan sebelum dan sesudah tuning:

Model	Status	F1-score	Accuracy
LSTM	Before Tuning	0.65	0.69
LSTM	After Tuning	0.7	0.7
GRU	Before Tuning	0.68	0.71
GRU	After Tuning	0.67	0.68
SVC	Before Tuning	0.69	0.7
SVC	After Tuning	0.7	0.7
Naive Bayes	Before Tuning	0.7	0.68
Naive Bayes	After Tuning	0.69	0.68
Logistic Regression	Before Tuning	0.7	0.69
Logistic Regression	After Tuning	0.7	0.69
Random Forest	Before Tuning	0.68	0.67
Random Forest	After Tuning	0.68	0.67

Setelah dilakukan training dan tuning terlihat bahwa ada beberapa model yang hasil evaluasi tuning nya sama persis dengan training atau bahwa menurun

Ini menunjukan bahwa Training saja sudah cukup untuk mendapatkan hasil evaluasi terbaik dari model



Hasil akhir diputuskan untuk menyimpan model LSTM sebelum tuning karena model ini memiliki skor evaluasi paling stabil dan merata diantara model lain

Example & Deployment

Perbandingan & Contoh

Disini terlihat bahwa hasil prediksi LSTM lebih banyak ke negatif diikuti netral lalu positif

```
Distribusi Prediksi pada Data Test:  
Predicted Sentiment (LSTM)  
negatif      811  
netral       746  
positif      714  
Name: count, dtype: int64
```

```
Data test actual  
negatif      767  
netral       761  
positif      743  
Name: count, dtype: int64
```

Perbandingan & Contoh

Disini terlihat bahwa hasil prediksi LSTM ada beberapa yang salah prediksi namun ada sedikit baris yang kelihatannya lebih masuk akal prediksi LSTM daripada IndoBERT lalu ada juga hasil IndoBERT yang lebih masuk akal dibanding hasil LSTM, ini semua kemungkinan disebabkan beberapa alasan

	Actual Sentiment	Predicted Sentiment (LSTM)	Comment
0	negatif	negatif	pohon sawit sama hutan lindung dasar lansia
1	positif	positif	untung aku pilih wkwk
2	positif	positif	kira tuhan serta beri kuat tiap sungguh cinta ...
3	netral	positif	bagus aceh merdeka kaya alam rampok jabat maling
4	positif	netral	zi hidup dekengane pusat
5	netral	netral	prabowo milik lahan sumatra luas tiga kali sin...
6	negatif	negatif	sedih kasihan indonesia alam manusia siksa gar...
7	positif	positif	betul turun langsung lapang
8	positif	positif	mantap abang masuk
9	netral	netral	gubernur sumut hehehe
10	netral	positif	undang gus sumbing bang wkwkwkwk
11	netral	netral	naik bencana nasional bajet perintah pusat dae...
12	positif	negatif	gila begini wni
13	negatif	negatif	tahun merdeka negara serasa tak pernah benar b...
14	negatif	negatif	aduh kasihan saudara ku sana tni al anggota se...
15	negatif	positif	mundur amin
16	netral	positif	indonesia pulau jawa
17	negatif	negatif	mirip pakai bpjs parah klaim korban dianggap b...
18	positif	positif	benerrr jangan pernah tunggang penting lawan p...
19	positif	positif	presiden gemoy sayang sawit lingkung rakyat

Perbandingan & Contoh

Alasanya adalah karena

1. Perbedaan cara kerja IndoBERT dan LSTM
2. LSTM bisa menangani kata slang/tidak baku lebih baik karena dilatih dengan komentar langsung dibandingkan IndoBERT yang dilatih korpus bahasa Indonesia yang relatif lebih formal
3. IndoBERT adalah model bahasa pra-terlatih (*pre-trained language model*) berbasis arsitektur Transformer yang dikhususkan untuk bahasa Indonesia dan sudah dilatih menggunakan bahasa indonesia yang jumlah nya jutaan kata

	Actual Sentiment	Predicted Sentiment (LSTM)	Comment
2251	negatif	negatif	malu bukan kultur jabat indonesia negara asia ...
2252	netral	netral	korban ribu status bencana nasional aktif
2253	positif	positif	gila berani betul kawal koi cok
2254	negatif	netral	cari versi konspirasi min
2255	positif	positif	kapitalis suka referendum kayak suka referendum...
2256	netral	positif	dedi corbucer dukung rezim
2257	negatif	negatif	salah siapa salah pilih masa pemiu
2258	netral	netral	wowo peduli citra dunia kirim pasu aman pbb lu...
2259	negatif	positif	sulit bahlil tum partai bpk lu jabat apa parpo...
2260	netral	positif	selamat bagi dosa
2261	positif	netral	partai hahahaha mimpi doa mu
2262	positif	positif	bukan berani awal kurang aktivitas podcast mun...
2263	negatif	negatif	perintah munafik benar kas kosong program jela...
2264	negatif	negatif	lahir kayak lihat susah empati
2265	negatif	positif	pusat enak rakyat bencana mending merdeka makm...
2266	negatif	negatif	betul kata bennix kokori sibuk negara orgpemer...
2267	netral	netral	tuju tuntut tuntas hukum berat berat libat seb...
2268	negatif	positif	trio yakin lu satu radikal bacot medsosudh kay...
2269	netral	netral	teori presiden prabowo pimpin nyata lapang jok...
2270	negatif	netral	dedy tahi budak politik

Deployment

Deployment dilakukan menggunakan Gradio, link bisa diakses disini
(<https://aaf110cf09f562638f.gradio.live/>)

Analisis Sentimen Komentar YouTube - Model LSTM

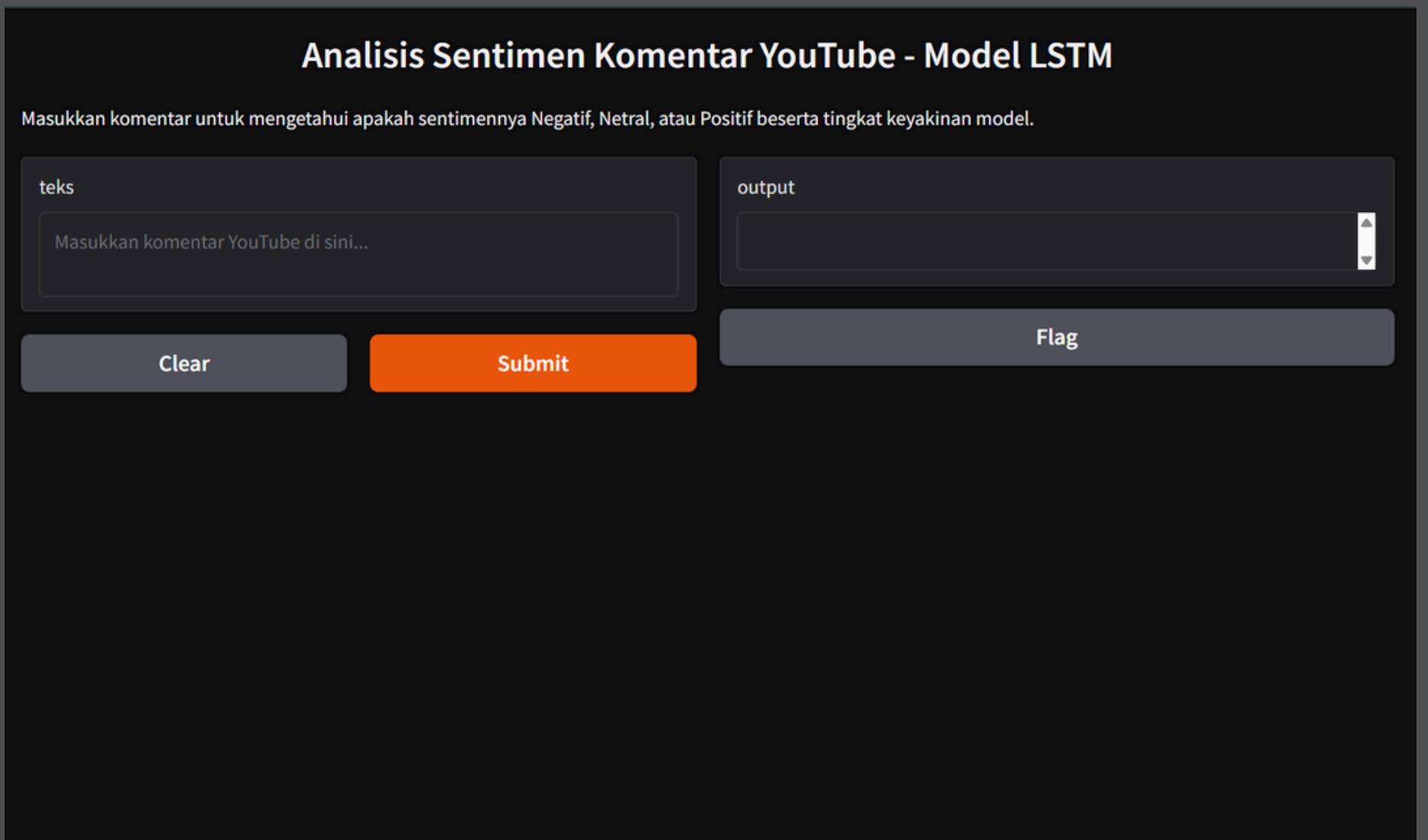
Masukkan komentar untuk mengetahui apakah sentimennya Negatif, Netral, atau Positif beserta tingkat keyakinan model.

teks

output

Flag

ClearSubmit

A screenshot of a web-based sentiment analysis application. The title is "Analisis Sentimen Komentar YouTube - Model LSTM". Below it is a subtitle: "Masukkan komentar untuk mengetahui apakah sentimennya Negatif, Netral, atau Positif beserta tingkat keyakinan model.". There are two main input fields: "teks" (text) and "output". The "teks" field contains placeholder text "Masukkan komentar YouTube di sini...". Below these fields are two buttons: "Clear" and "Submit" (highlighted in orange). To the right of the "output" field is a "Flag" button. The entire interface is set against a dark background.

Conclusion

Conclusion

1 Akurasi dan stabilitas model terbaik

Model LSTM (Bidirectional) yang digunakan berhasil mencapai akurasi sebesar 71% dengan nilai evaluasi (F1-score) yang stabil di angka 0.70 hingga 0.72 untuk ketiga kelas sentimen

2 Keunggulan model LSTM terhadap bahasa tidak formal

Model LSTM terbukti lebih unggul dalam menangani komentar YouTube yang mengandung kata-kata slang atau bahasa tidak baku karena model ini dilatih langsung menggunakan data komentar masyarakat, hal ini membuat beberapa baris prediksi LSTM terasa lebih masuk akal dibandingkan model IndoBERT yang cenderung dilatih dengan korpus bahasa Indonesia formal

3 Distribusi sentimen publik terhadap bencana

Hasil analisis menunjukkan bahwa reaksi publik terhadap bencana banjir di Sumatra didominasi oleh sentimen negatif, yang kemudian diikuti oleh sentimen netral dan positif, lambat nya bantuan yang bisa sampai ke lokasi bencana dan parahnya dampak bencana yang ditimbulkan yang kemungkinan besar menjadi penyebab banyak nya komentar dengan sentimen negatif

4 Keberhasilan implementasi dan deployment

Proyek ini telah berhasil mencapai tahap akhir dengan dilakukannya deployment menggunakan platform Gradio yang dapat diakses secara publik melalui link tertentu

5 Evaluasi project

Hasil label oleh IndoBERT tidak semua nya akurat malah ada beberapa yang lebih masuk akal hasil prediksi model ML klasik, jadi model canggih juga perlu di cek kembali hasil label nya agar lebih akurat atau masuk akal

THANK YOU

By : Raihan Pratama

