MTK-FST-UIN JKT

PROJECT UTS

NATURAL LANGUAGE PROCESSING

Membuat model machine learning (Naive Bayes dan Logistic Regression)

11220940000020 M. Abdul Ghofur

11220940000028 Alif Alamsyah

KELOMPOK 4

Pendahuluan

Klasifikasi Tweet Bencana

Di era media sosial, informasi menyebar dengan sangat cepat. Sebuah tweet yang menggunakan kata "fire" (kebakaran) bisa berarti laporan bencana nyata, atau bisa juga hanya sebuah ungkapan metafora untuk sesuatu yang luar biasa ("this song is fire!"). Kemampuan untuk membedakan keduanya secara otomatis sangat krusial bagi organisasi pemantau bencana atau tim berita untuk memberikan respons yang cepat dan tepat.

Tujuan utama dari proyek ini adalah untuk membangun dan mengevaluasi beberapa model machine learning yang mampu mengklasifikasikan sebuah tweet ke dalam dua kategori:

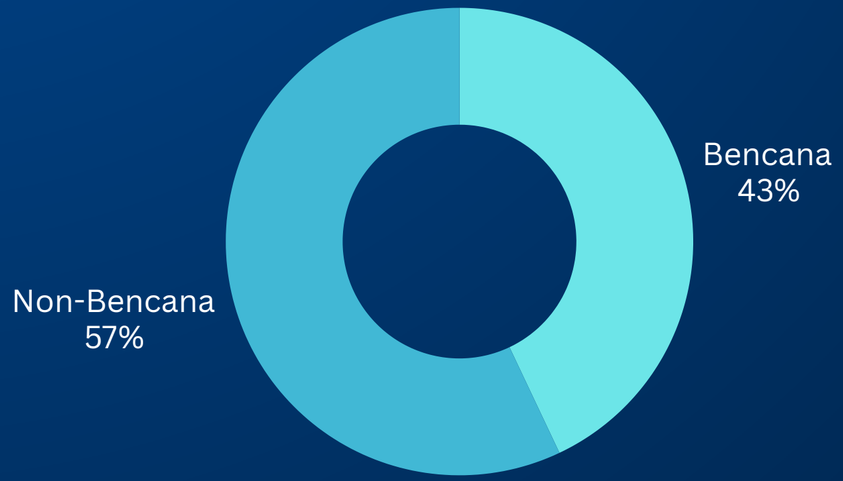
- Bencana (Target = 1): Tweet tersebut mengandung informasi tentang bencana nyata.
- Bukan Bencana (Target = 0): Tweet tersebut tidak berhubungan dengan bencana.

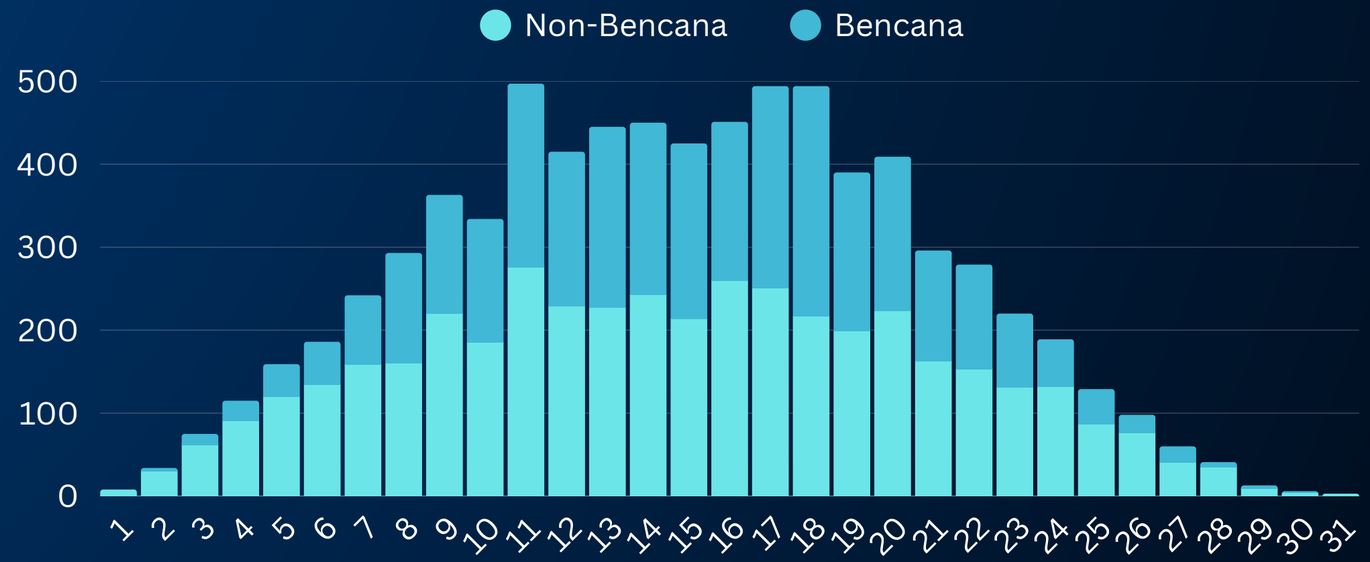
Daftar Isi

Exploratory Data Analysis

Tentang Data

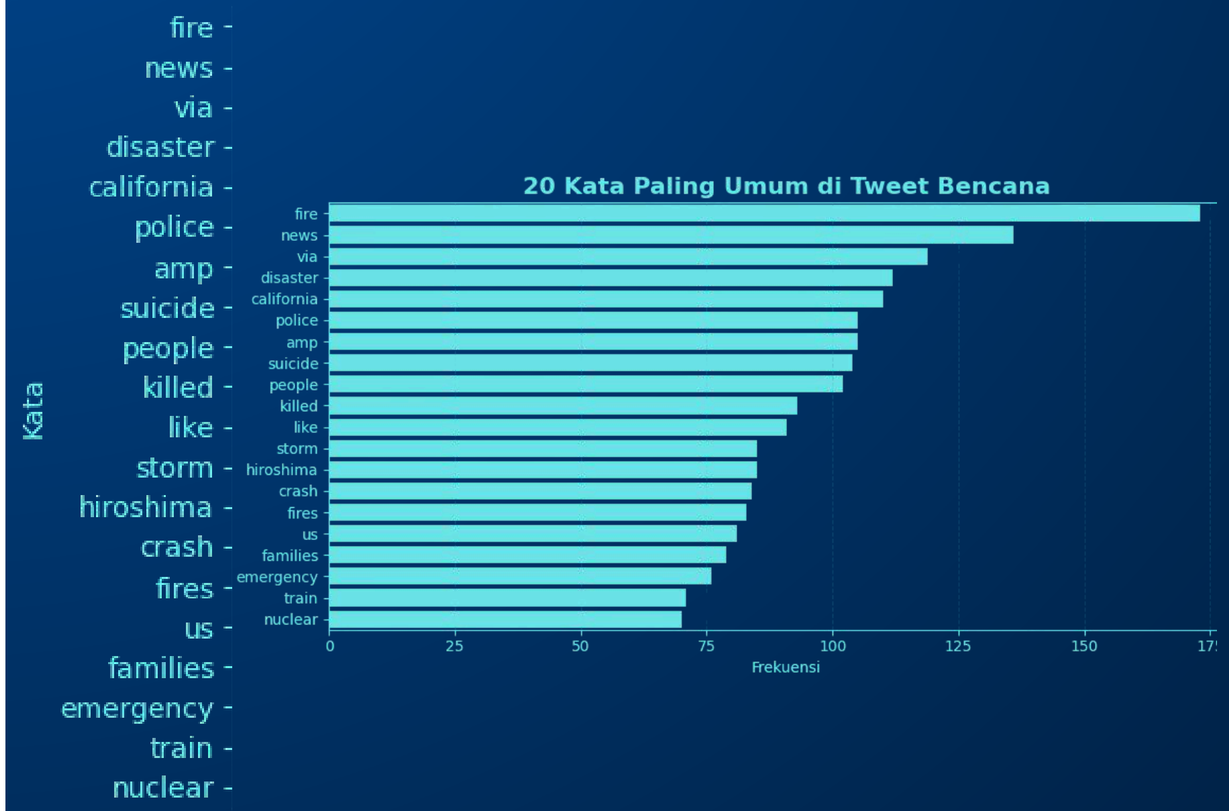
7613 Data

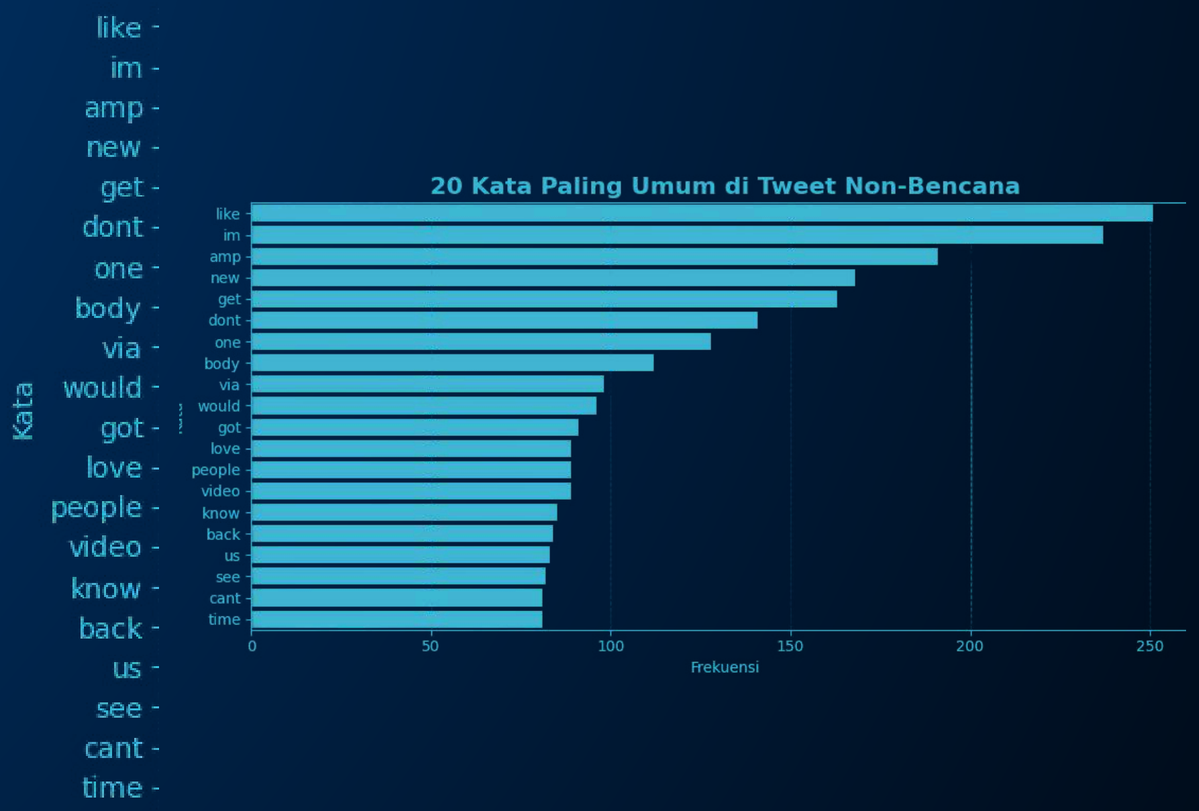




Perbedaan jumlah tweet dan distribusi kata pada setiap tweet antar kelas tidak signifikan secara visial

Exploratory Data Analysis
Tentang Data





- Perbedaan Kosakata antara tweet Bencana dan Non-bencana

Data Preprocessing

Hapus Duplikasi Data
Mencegah Data Menghafal

7613 Data → 7503 Data

Cleaning Text using taudataNlpTm

- Remove URL
- Unescape HTML
- Lower Case
- Replace newline
- Remove symbols
- Filter karakter
- Stopword removal
- Lemmatisasi (Spacy)

MTK-FST-UIN JKT

MODEL BUILDING

Text vectorization

TF-IDF vectorizer (Sklearn)

TF-IDF adalah metode untuk mengubah teks menjadi angka yang mencerminkan seberapa penting suatu kata dalam dokumen relatif terhadap seluruh koleksi dokumen.

$$tfidf(t,d) = tf(t,d) \times idf(t)$$

TF (Term Frequency)

$$tf(t,d) = count(t,d)$$

Default

$$tf(t,d) = 1 + log(count(t,d))$$

sublinear_tf=True

IDF (Inverse Document Frequency)

$$idf(t) = log\left(\frac{1+n}{1+df(t)}\right) + 1$$

Default

$$idf(t) = log\left(\frac{n}{df(t)}\right) + 1$$

smooth_idf=False

- max_df
- min_df
- ngram_range
- smooth_idf
- stop_words
- sublinear_tf

Feature Selection

SelectKBest (Sklearn)

Memilih k fitur terbaik

- X^2 test
- Anova test

Hapus Data Kosong

Menghapus baris yang semua nilainya nol.

Model Naive Bayes

Multinomial Naive Bayes

algoritma klasifikasi berdasarkan teorema Bayes yang mengasumsikan bahwa fitur input (misalnya, kata-kata dalam dokumen) mengikuti distribusi multinomial yang mewakili jumlah kemunculan kata dalam setiap kelas.

- alpha
- fit_prior

Complement Naive Bayes

variasi dari Multinomial Naive Bayes yang menghitung probabilitas berdasarkan komplemen dari kelas target, yaitu dokumen-dokumen dari kelas lain. Model ini dirancang untuk mengatasi ketidakseimbangan kelas.

- alpha
- fit_prior
- norm

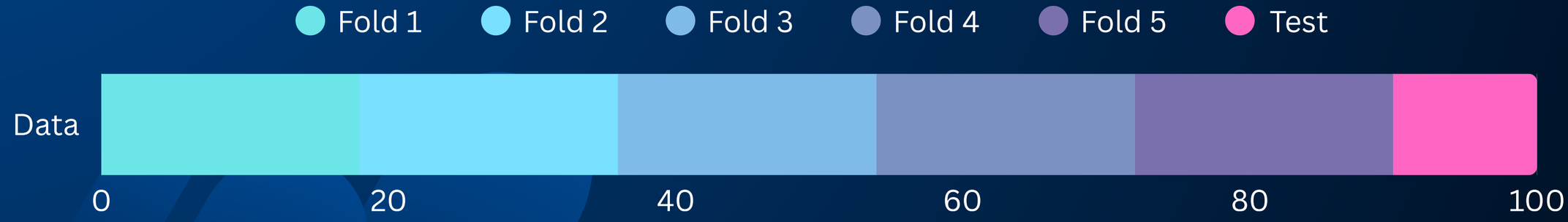
Model Logistic Regression

model statistik yang digunakan untuk memprediksi probabilitas suatu peristiwa biner (dua kelas) berdasarkan input fitur. Berbeda dengan regresi linear yang menghasilkan nilai kontinu, logistic regression menggunakan fungsi logistik (sigmoid) untuk mengubah output menjadi rentang antara 0 dan 1, sehingga cocok untuk klasifikasi.

- C
- class_weight
- max_iter
- penalty
- solver

Grid Search
Cross Validation

7503 Data



- Test = 10%
- Training = 90%
- CrossValidation pakai 5 Fold pada data train

Grid Search

Tujuan Utama Model

Sistem Peringatan Dini

Menghindari informasi Bencana terlewat → False Negatif (FN) Kecil → Recall Tinggi

Presisi Rendah → False Positif (FP) Tinggi → Banyak Informasi Bencana Palsu

Fokus Recall dengan mempertimbang kan F1 score dan akurasi

$$score = recall_{weight} * recall_1 + f1_{weight} * f1 + acc_{weight} * acc$$

Model Naive Bayes

Hyperparameter Tuning

TF-IDF vectorizer

max_df	[0.5, 0.7, 0.9]
min_df	[1, 2]
ngram_range	[(1, 1), (1, 2)]
smooth_idf	True
stop_words	english
sublinear_tf	[True, False]

Preprocess Vektor

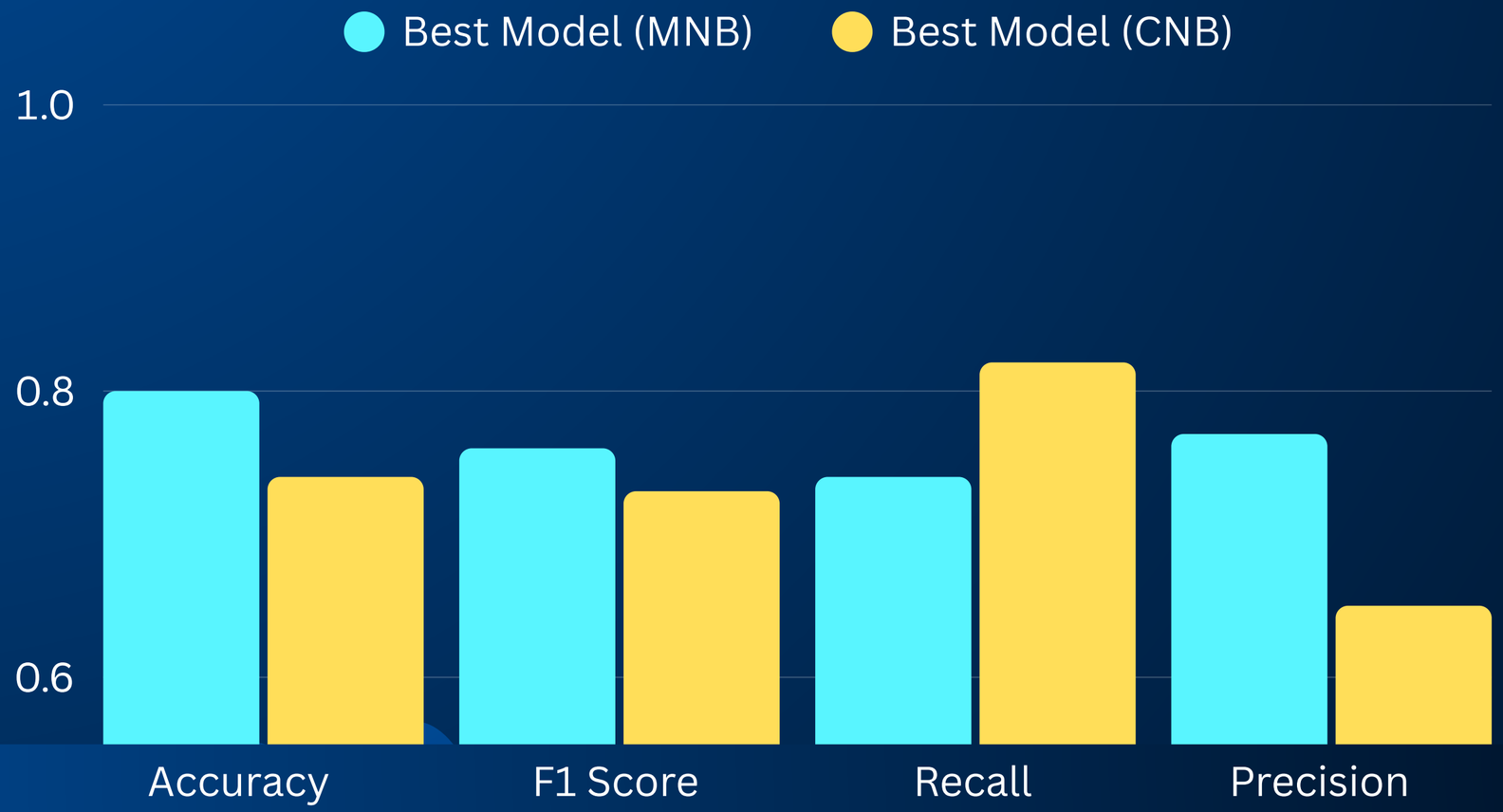
use_selectkbest	True
score_func	[X^2test, Anova test]
k_feature	[3000, 4000, 5000, 6000]
use_hapus_kosong	[True, False]

MNB & CNB parameters

alpha	[0.01, 0.1, 0.5, 1.0]
fit_prior	[True, False]
norm	[True, False]

Model Naive Bayes

By Best Recall

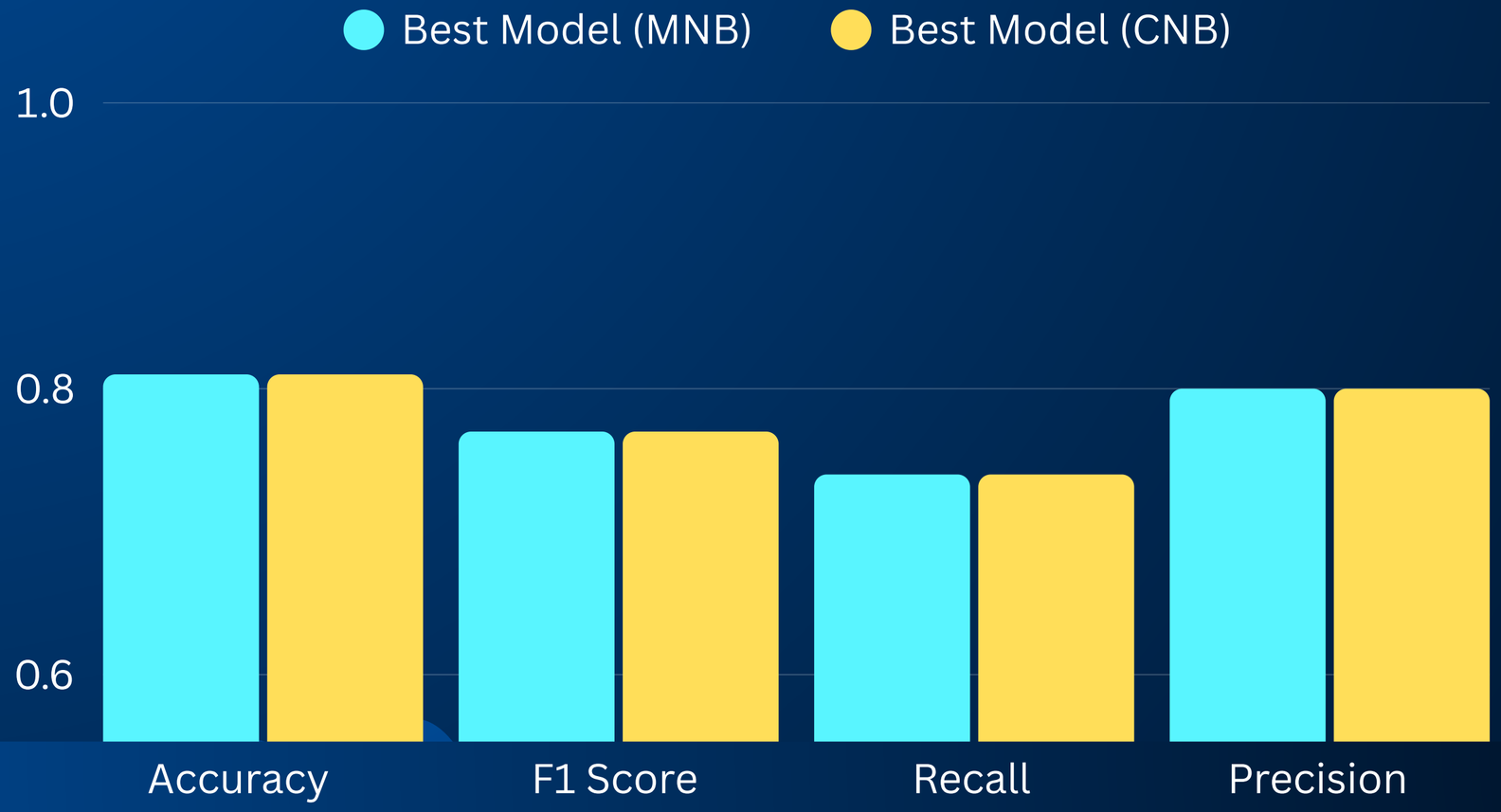


Risiko:

- Nilai Precision Jatuh

Model Naive Bayes

By Best F1

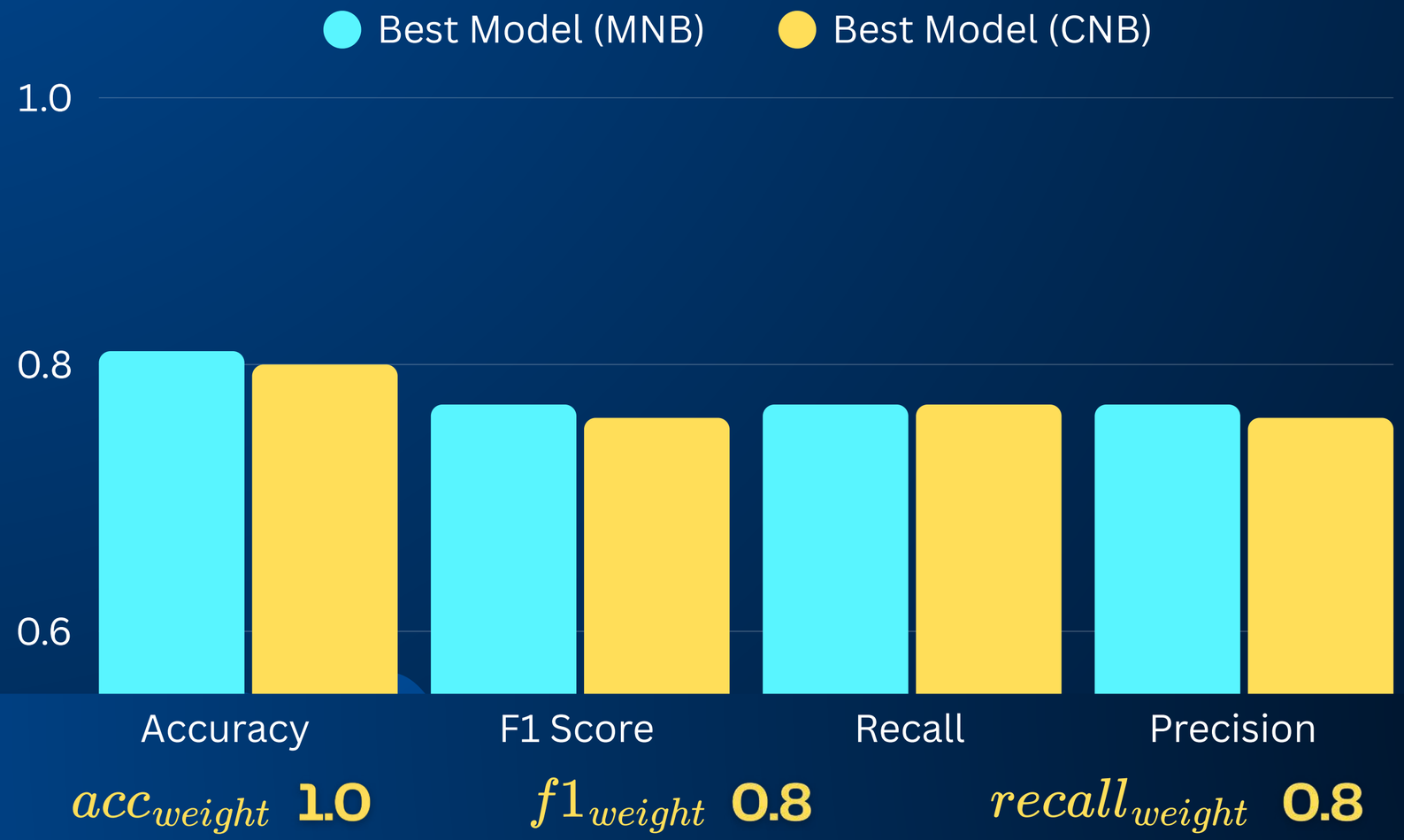


Risiko:

- Nilai Recall atau Precision rendah

Model Naive Bayes

Our Score



acc_{weight} **1.0** $f1_{weight}$ **0.8** $recall_{weight}$ **0.8**

Best Model (CNB)

$score$ **2.0217**

Best Model (MNB)

$score$ **2.0338**

- Lebih Stabil

Model Naive Bayes

Multinomial Naive Bayes

TF-IDF vectorizer

max_df	0.5
min_df	2
ngram_range	(1, 1)
smooth_idf	True
stop_words	english
sublinear_tf	True

Preprocess Vektor

use_selectkbest	True
score_func	Anova test
k_feature	3000
use_hapus_kosong	True

MNB parameters

alpha	1.0
fit_prior	False

Model Logistic Regression

Hyperparameter Tuning

TF-IDF vectorizer

max_df	0.7
min_df	2
ngram_range	[(1, 1), (1, 2)]
smooth_idf	True
stop_words	english
sublinear_tf	[True, False]

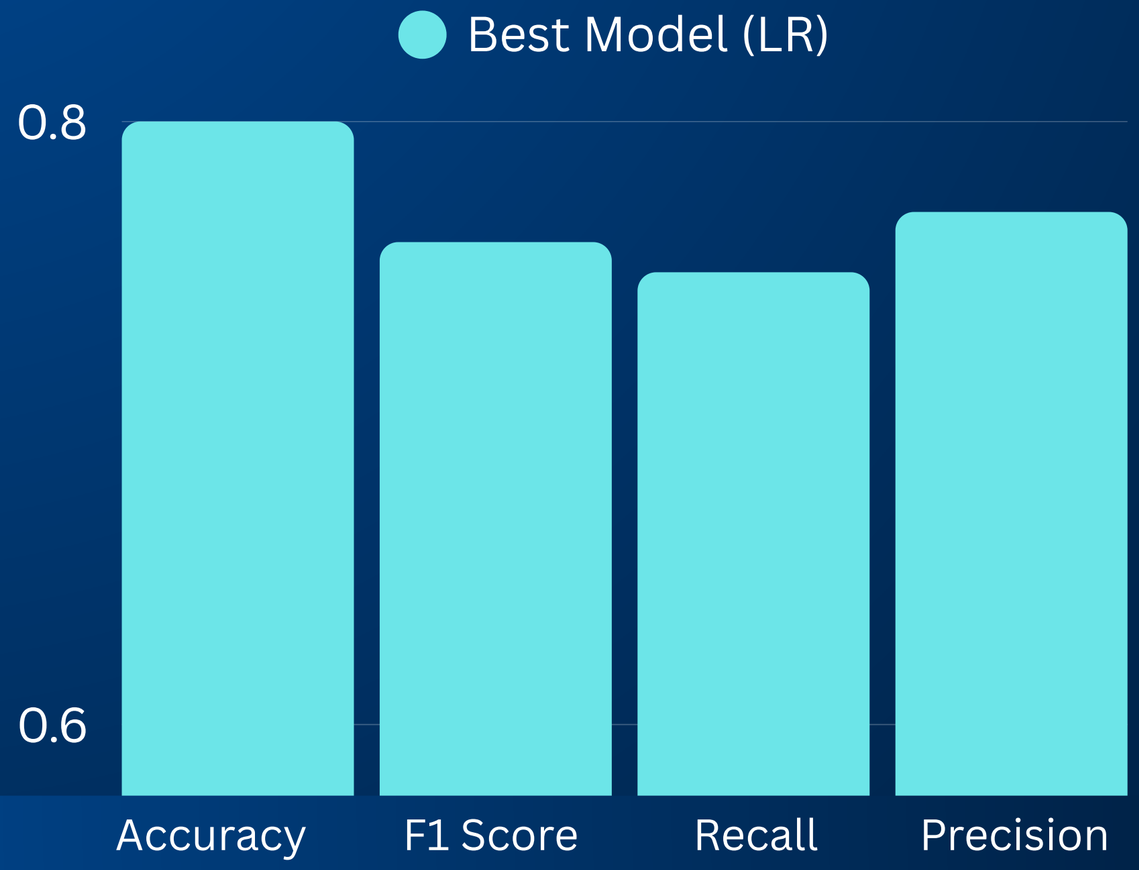
Preprocess Vektor

use_selectkbest	[True, False]
score_func	[X^2test, Anova test]
k_feature	[3000, 5000]
use_hapus_kosong	True

LR parameters

C	[1, 10, 100]
class_weight	[None, 'balanced']
max_iter	1000
penalty	['L1', 'L2', 'elasticnet']
solver	['liblinear', 'saga', 'lbfgs', 'newton-cg']

Model Logistic Regression



Best Model (LR)

$score$ **1.9937**

acc_{weight} **1.0** $f1_{weight}$ **0.8** $recall_{weight}$ **0.8**

Model Logistic Regression

TF-IDF vectorizer

max_df	0.7
min_df	2
ngram_range	(1, 2)
smooth_idf	True
stop_words	english
sublinear_tf	True

Preprocess Vektor

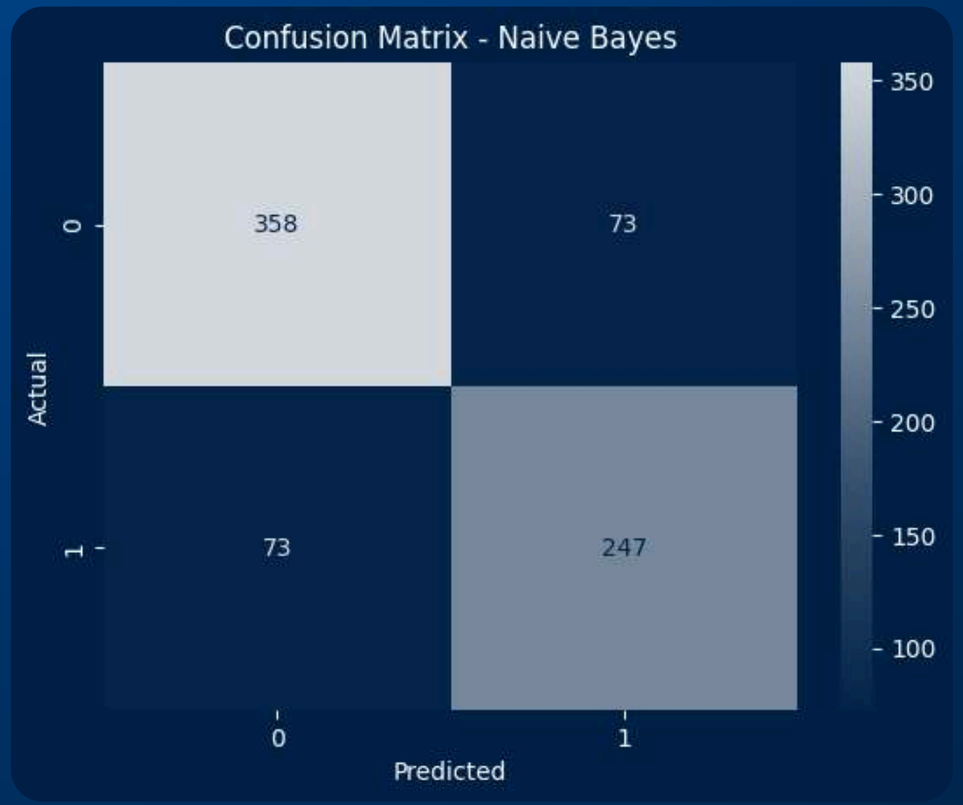
use_selectkbest	True
score_func	Anova test
k_feature	5000
use_hapus_kosong	True

LR parameters

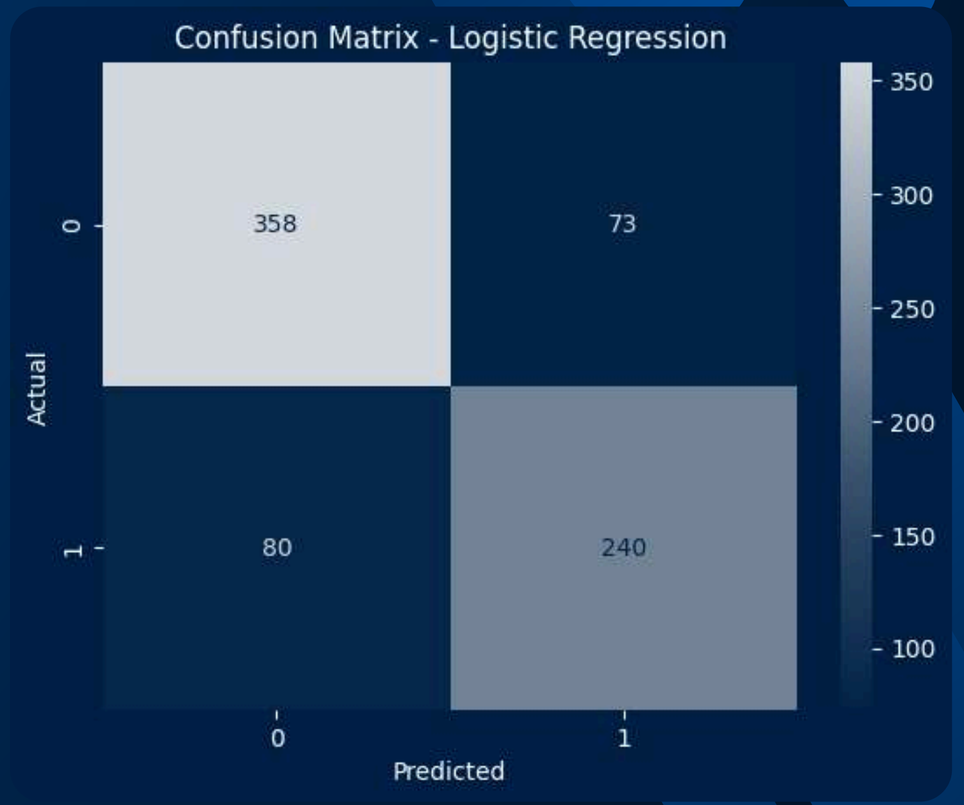
C	10
class_weight	balanced
max_iter	1000
penalty	L2
solver	saga

Confution Matrix

Naive Bayes



LogReg



MTK-FST-UIN JKT

MODEL INTERPRETABILITY

Analisis Kesalahan Prediksi

"Another white guy trying to mass murder people for no apparent reason just because let me guess he's mentally ill blah blah blah #Antioch"

white guy mass murder apparent reason guess 's mentally blah blah blah antioch

"Patient-reported outcomes in long-term survivors of metastatic colorectal cancer – British Journal of Surgery http://t.co/5Yl4DC1Tqt."

patient report outcome long term survivor metastatic colorectal cancer british journal surgery

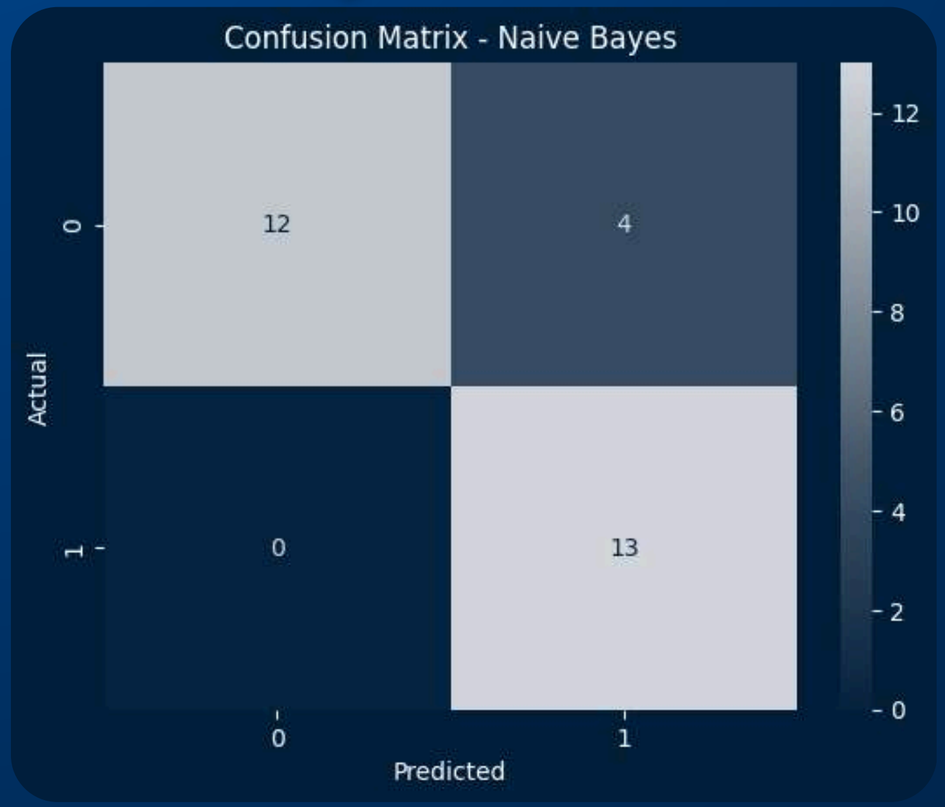
"RP said they can see smoke coming from the silo on 260th Street in Hartford but no flames."

rp smoke come silo street hartford flame

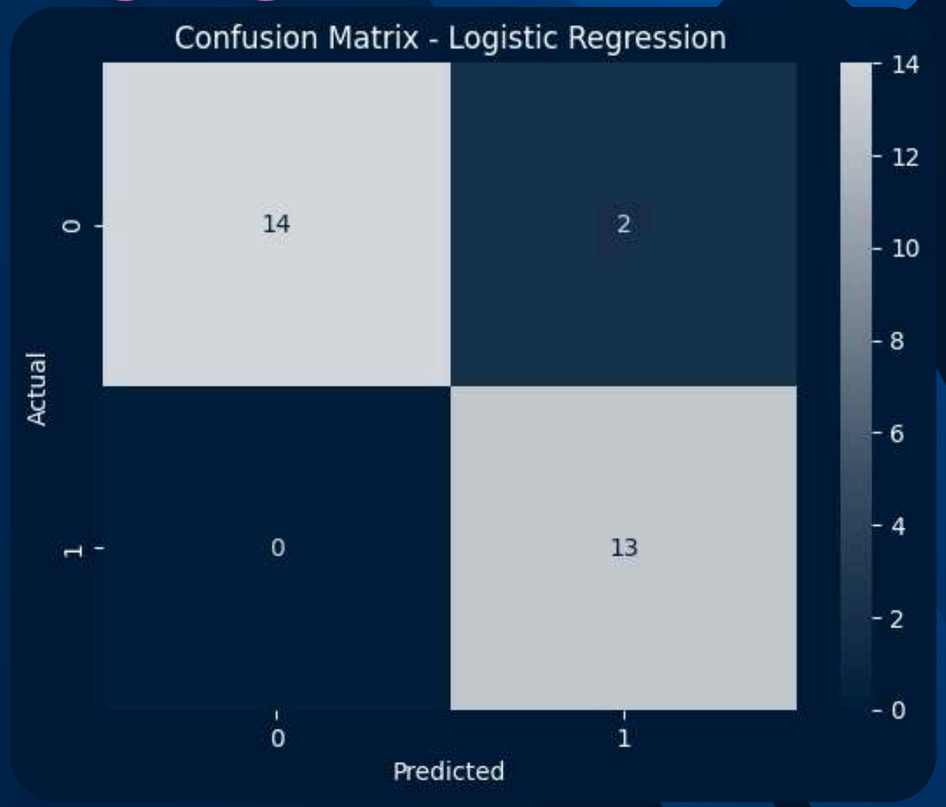
Ambiguitas kata

Confution Matrix ‘ablaze’

Naive Bayes



LogReg



Ambiguitas kata

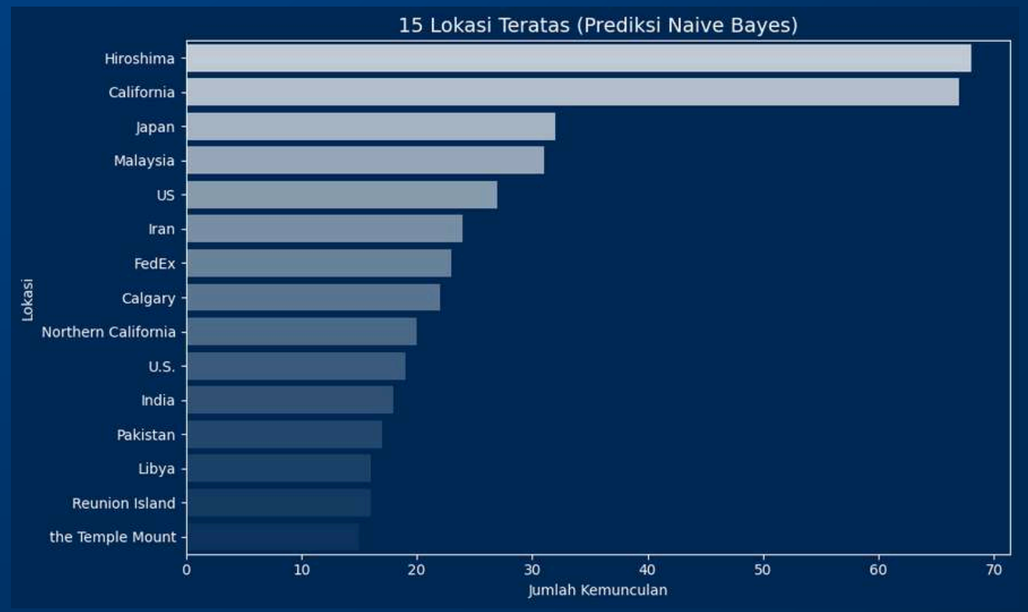
Analisis Kesalahan pada Keyword 'ablaze'

Text	Label Asli	NB	LR
-	Bencana	Non Bencana	Bencana
-	Bencana	Bencana	Non Bencana
“Crying out for more! Set me ablaze” cry set ablaze	Non Bencana	Bencana	Bencana
“on the outside you're ablaze and alive but you're dead inside” be ablaze alive be dead inside	Non Bencana	Bencana	Non Bencana

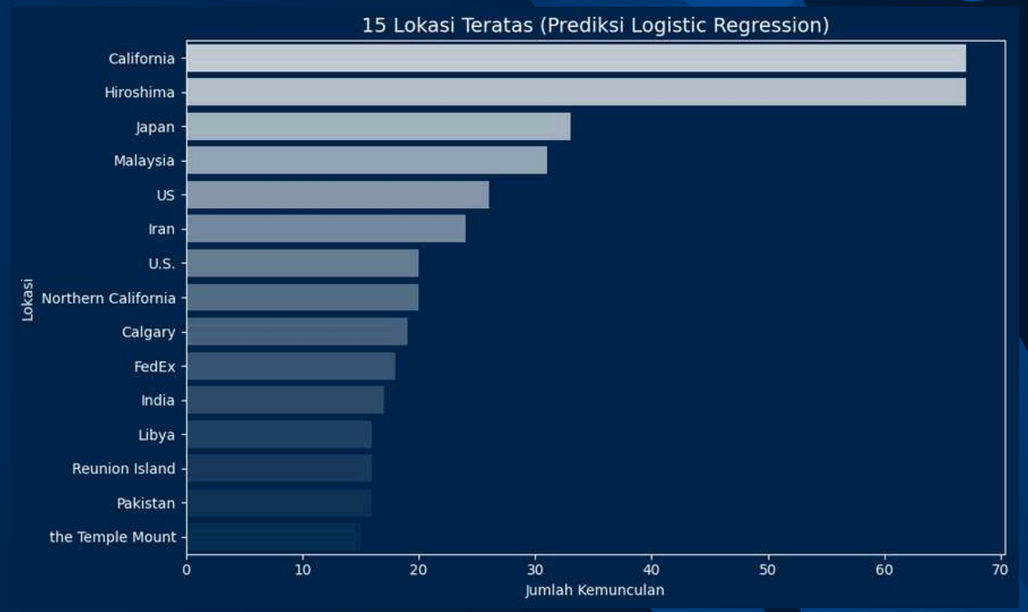
Potensi Model

Analisis Lokasi yang dapat dideteksi dari hasil model

Naive Bayes



LogReg



Kemampuan model dalam deteksi bencana dapat memberikan manfaat besar untuk melacak lokasi berdasarkan data media sosial.

NER

Kesimpulan

- Model Naive Bayes memberikan hasil yang sedikit lebih baik dibandingkan Logistic Regression, meskipun masih cukup banyak perbedaan prediksi antara kedua model.
- Untuk kata ambigu seperti “ablaze”, kedua model sudah cukup baik dalam membedakan konteks. Namun cenderung ‘memaksa’ prediksi sebagai Bencana.
- Dengan mengombinasikan bobot pada recall, F1-score, dan akurasi, model yang dihasilkan memiliki performa yang lebih stabil. Pendekatan ini mengurangi risiko kesalahan deteksi, sambil tetap mempertimbangkan prioritas metrik yang paling relevan.
- Meskipun menggunakan model Machine Learning yang relatif sederhana dan data yang terbatas, hasil prediksi yang diperoleh sudah cukup baik. Hal ini menunjukkan potensi besar pemanfaatan media sosial yang selalu terbarui sebagai sumber data untuk deteksi bencana secara real-time.

MTK-FST-UIN JKT

TERIMAKASIH

Semoga presentasi yang kami sampaikan dapat bermanfaat untuk kita semua.