PROJECT UTS

NATURAL LANGUAGE PROCESSING

Muhammad Aqlani Wafi - 11210940000070
Muhammad Fahmi Islam F - 11220940000016
Reyhan Maulana Aryaduta - 11220940000048

DATASET

```
baris =  7613 , Kolom (jumlah variabel) =  5
Tipe Variabe df =  <class 'pandas.core.frame.DataFrame'>
```

1 to 5 of 5 entries Filter

index	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation orders in California	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	1

Show 25 per page

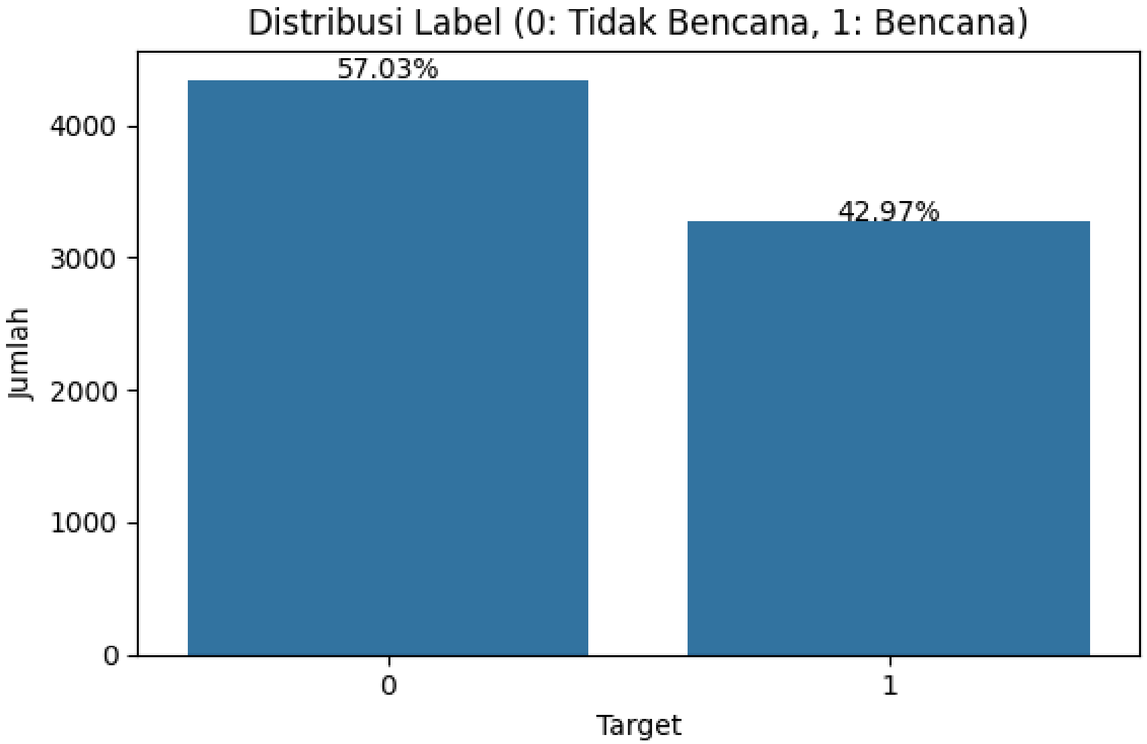
Tabel kolom:

- **text: Isi teks dari tweet.**
- **keyword: Kata kunci tertentu dari tweet (bisa kosong).**
- **location: Lokasi tempat tweet dikirim (bisa kosong).**
- **target: Label ini menandakan apakah sebuah tweet tentang**
- **bencana nyata (1) atau bukan (0).**

EXPLORATORY DATA ANALYSIS

Distribusi Label

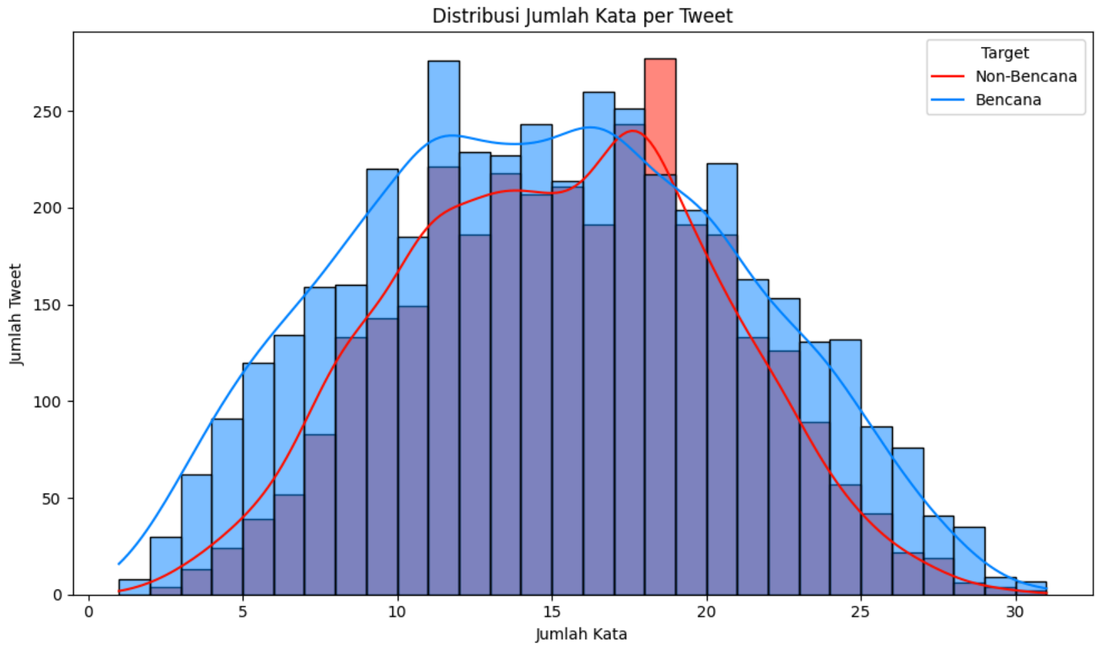
- Label 0 (Tidak Bencana) memiliki jumlah data sekitar 4000, yang jauh lebih tinggi.
- Label 1 (Bencana) memiliki jumlah data sekitar 3000, lebih sedikit dibandingkan label 0.



EXPLORATORY DATA ANALYSIS

Distribusi Jumlah Kata

Kategori "Tidak Bencana" (biru) mendominasi di sebagian besar rentang, terutama di bawah 20 kata.
Kategori "Bencana" (merah) memiliki puncak kecil di sekitar 15-20 kata, tetapi jumlahnya lebih sedikit dibandingkan "Tidak Bencana".
Kurva distribusi (garis merah dan biru) menunjukkan bahwa "Tidak Bencana" memiliki variasi lebih luas, sementara "Bencana" lebih terpusat.

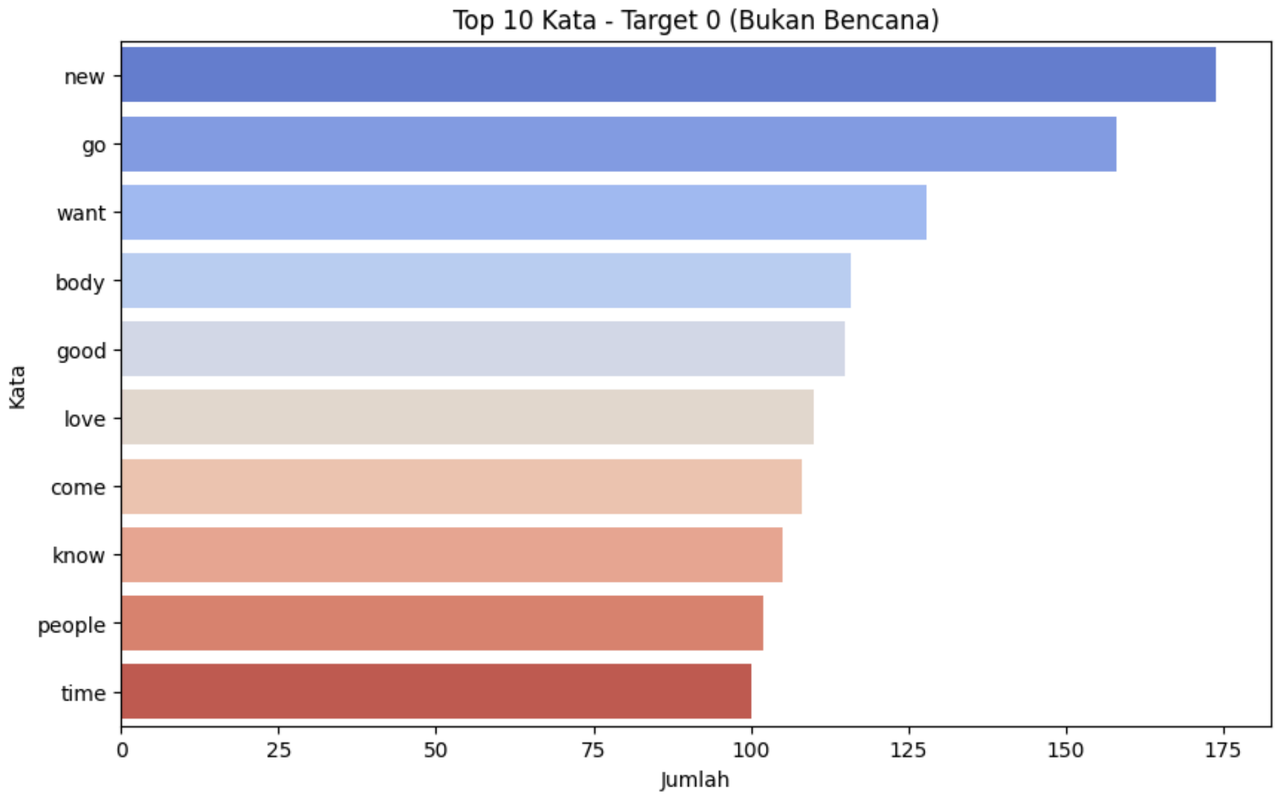


EXPLORATORY DATA ANALYSIS

Contoh Tweet Berdasarkan Label Bencana dan Tidak

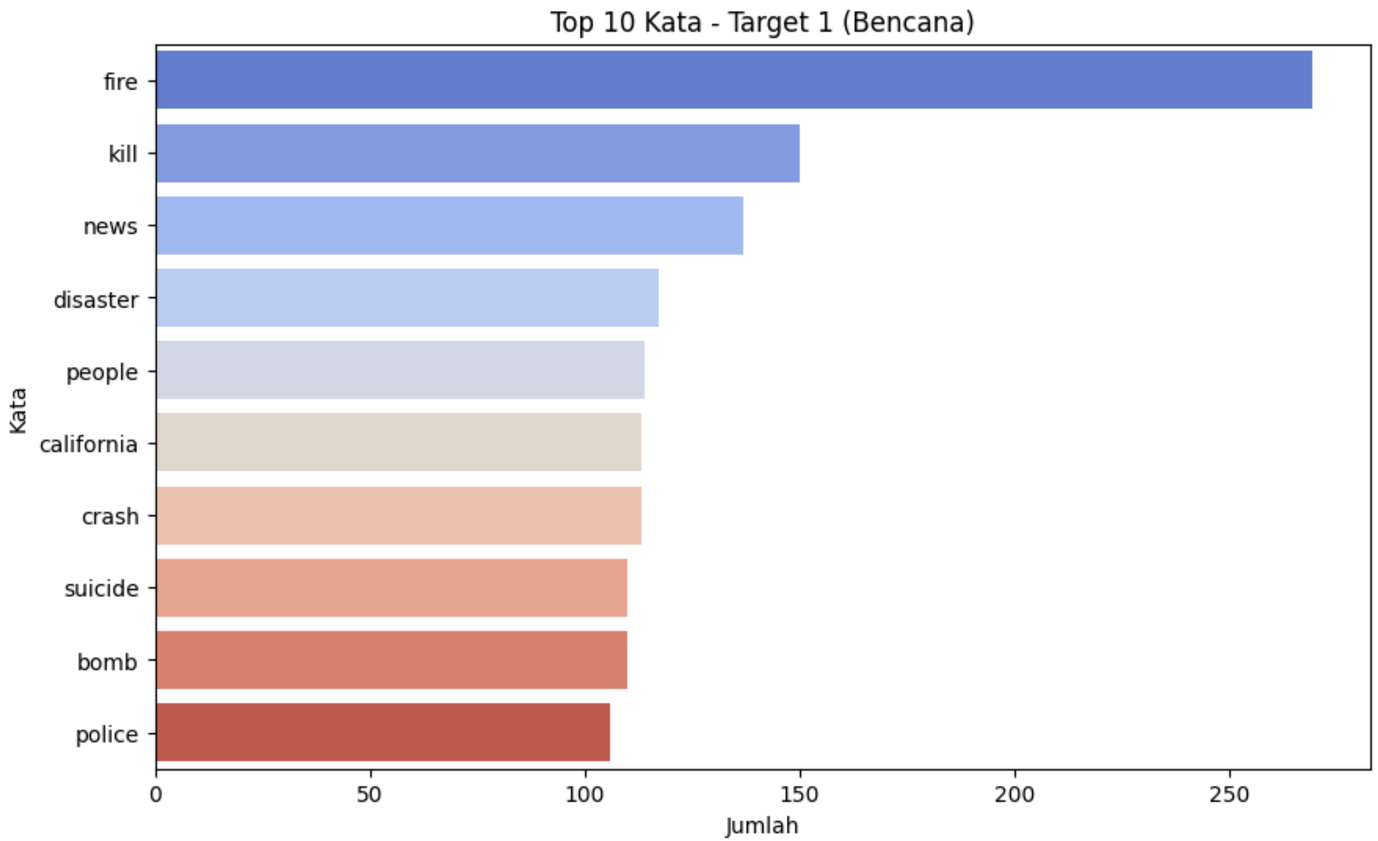
```
=== Tweet Bencana ===
1. Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all
2. Forest fire near La Ronge Sask. Canada
3. All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected

=== Tweet Bukan Bencana ===
1. What's up man?
2. I love fruits
3. Summer is lovely
```

EXPLORATORY DATA ANALYSIS



Kata-kata ini mencerminkan topik atau istilah umum yang sering muncul dalam tweet yang tidak terkait bencana, seperti hal-hal sehari-hari atau emosi positif.

EXPLORATORY DATA ANALYSIS



Kata-kata ini mencerminkan topik atau istilah yang sering dikaitkan dengan peristiwa bencana, seperti kebakaran, kekerasan, atau bencana alam, serta respon darurat seperti polisi.

EXPLORATORY DATA ANALYSIS

Data Splitting

- Pemisahan data: 80% latih, 20% uji: Data dibagi menjadi 80% untuk melatih model dan 20% untuk menguji performanya, memastikan evaluasi yang seimbang.
- Stratifikasi berdasarkan target: Distribusi kelas dalam 'target' dipertahankan sama di data latih dan uji, menghindari bias kelas.
- Random state: 42 untuk konsistensi: Seed acak diatur ke 42 agar pembagian data konsisten dan dapat direproduksi.

DATA PREPROCESSING

- Kecilkan teks: text.lower().
- Expand kontraksi: contractions.fix(text) (misalnya, "you're" jadi "you are").
- Hapus URL: re.sub(r'https?://\S+lwww\.\S+', '').
- Hapus tag HTML dan kurung siku: re.sub(r'<.*?>+', '') dan re.sub(r'\[.*?\]', '').
- Hapus tanda baca: re.sub(r'[%s]' % re.escape(string.punctuation), '').
- Ganti newline: re.sub(r'\n', ' ').
- Hapus kata berangka: re.sub(r'\w*\d\w*', '').
- Hapus karakter non-ASCII: re.sub(r'[^\x00-\x7f]', r'', text).
- Proses stopwords dan lemmatization: Gunakan nlp untuk lemmitisasi dan filter token yang bukan stopword (dari spaCy dan custom), bukan tanda baca/spasi, dan panjang > 1 huruf.
- Gabungkan token: join token bersih jadi teks.

DATA PREPROCESSING

index	text	clean_text
0	Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	deed reason may allah forgive
1	Forest fire near La Ronge Sask. Canada	forest fire near ronge sask canada
2	All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected	resident ask shelter place notify officer evacuation shelter place order expect
3	13,000 people receive #wildfires evacuation orders in California	people receive evacuation order california
4	Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	get sent photo ruby smoke pour school
5	#RockyFire Update => California Hwy. 20 closed in both directions due to Lake County fire - #CAfire #wildfires	update california hwy close direction due lake county fire
6	#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Colorado Springs areas	heavy rain cause flash flood street manitou colorado spring area
7	I'm on top of the hill and I can see a fire in the woods...	top hill see fire wood
8	There's an emergency evacuation happening now in the building across the street	emergency evacuation happen building across street
9	I'm afraid that the tornado is coming to our area...	afraid tornado come area

Tweet sebelum dan sesudah preprocessing

TEXT VECTORIZATION

Naive Bayes

Parameter yang digunakan:

- **tfidf__ngram_range: (1, 1) (hanya unigram)**
- **tfidf__min_df: 1 (batas minimum frekuensi dokumen)**
- **tfidf__max_features: 5000 (jumlah fitur maksimum)**
- **tfidf__max_df: 0.8 (batas maksimum frekuensi dokumen)**
- **algo__alpha: 1.0 (parameter regularisasi)**

TEXT VECTORIZATION

Parameter terbaik yang ditemukan adalah:

- clf__C: 1 (parameter regularisasi untuk model klasifikasi)
- clf__class_weight: 'balanced' (penyesuaian bobot kelas untuk menangani ketidakseimbangan)
- clf__penalty: 'l2' (penalti regulasi L2)
- clf__solver: 'liblinear' (algoritma optimasi)
- preprocess__tfidf__max_df: 0.85 (batas maksimum dokumen frekuensi untuk menghapus kata umum)
- preprocess__tfidf__max_features: 7000 (jumlah fitur maksimum yang digunakan)
- preprocess__tfidf__min_df: 1 (batas minimum dokumen frekuensi untuk menghapus kata langka)
- preprocess__tfidf__ngram_range: (1, 1) (hanya menggunakan unigram)

Logistik Regresi

TRAINING THE MODELS

Naive Bayes

TF-IDF

- ngram_range : [(1, 1), (1, 2)],
- max_df : [0.8, 0.9, 1.0],
- min_df : [1, 2],
- max_features : [3000, 5000, 7000],

Random Search

MULTINOMIAL DB

- alpha : [0.1, 0.5, 1.0]

TRAINING THE MODELS

Logistik Regresi

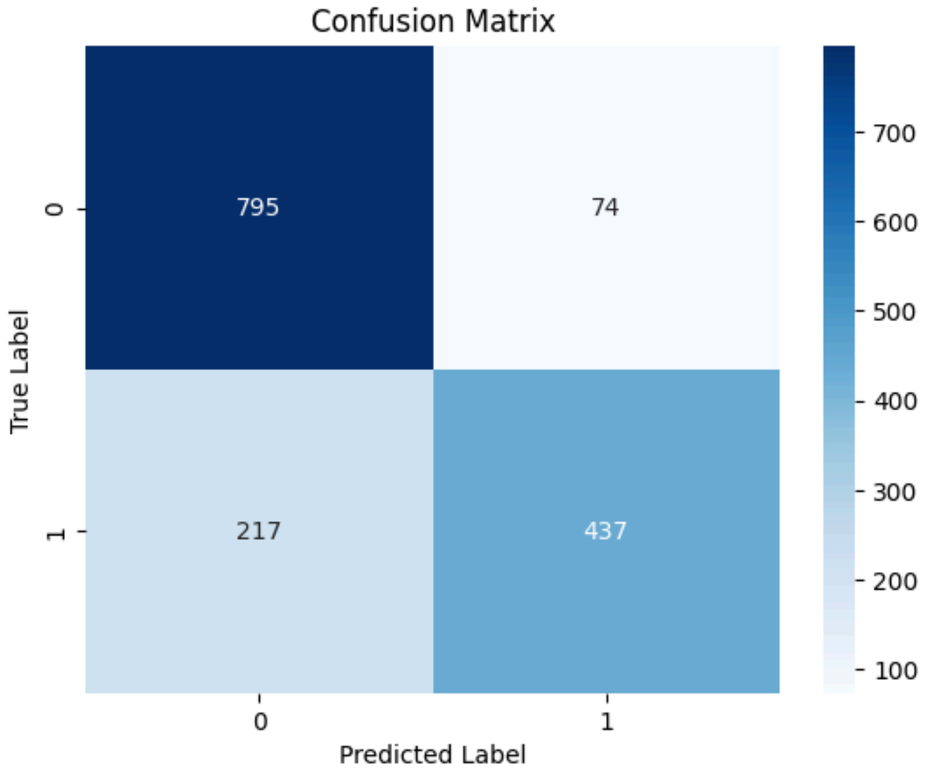
clf__C : [0.01, 0.1, 1, 10],
clf__solver : ['liblinear', 'saga'],
clf__class_weight : [None, 'balanced']

EVALUATION & VALIDATION

Naive Bayes

Evaluasi

```
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.91      0.85       869
           1       0.86      0.67      0.75       654

    accuracy                           0.81      1523
   macro avg       0.82      0.79      0.80      1523
weighted avg       0.82      0.81      0.80      1523
```



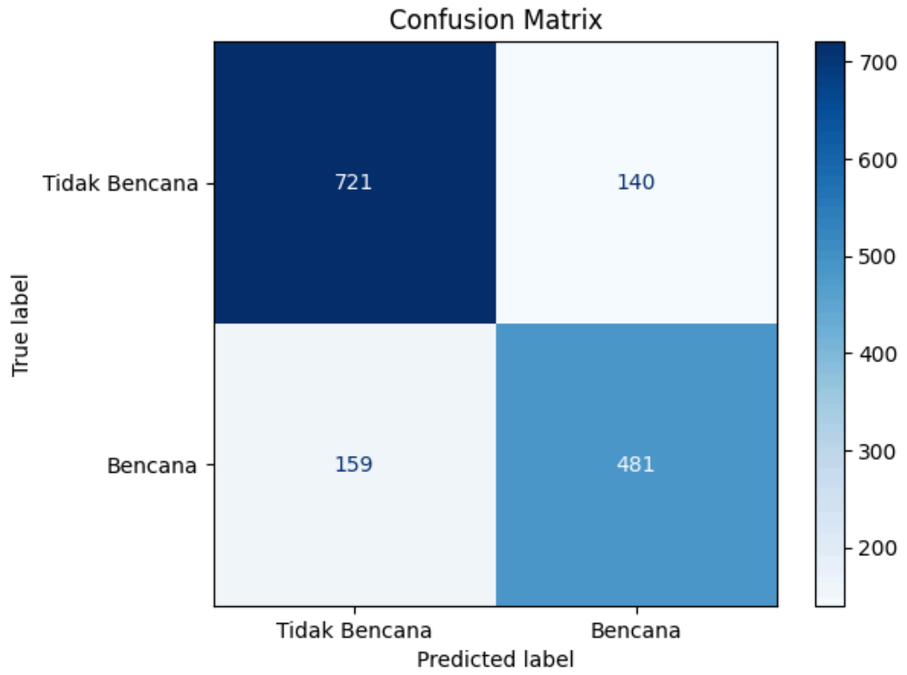
EVALUATION & VALIDATION

Regresi Logistik

EVALUASI

```
=== Evaluation on Test Set ===
Accuracy: 0.8007994670219853
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.84      0.83       861
           1       0.77      0.75      0.76       640

    accuracy                           0.80      1501
   macro avg       0.80      0.79      0.80      1501
weighted avg       0.80      0.80      0.80      1501
```

CONFUSSION MATRIX

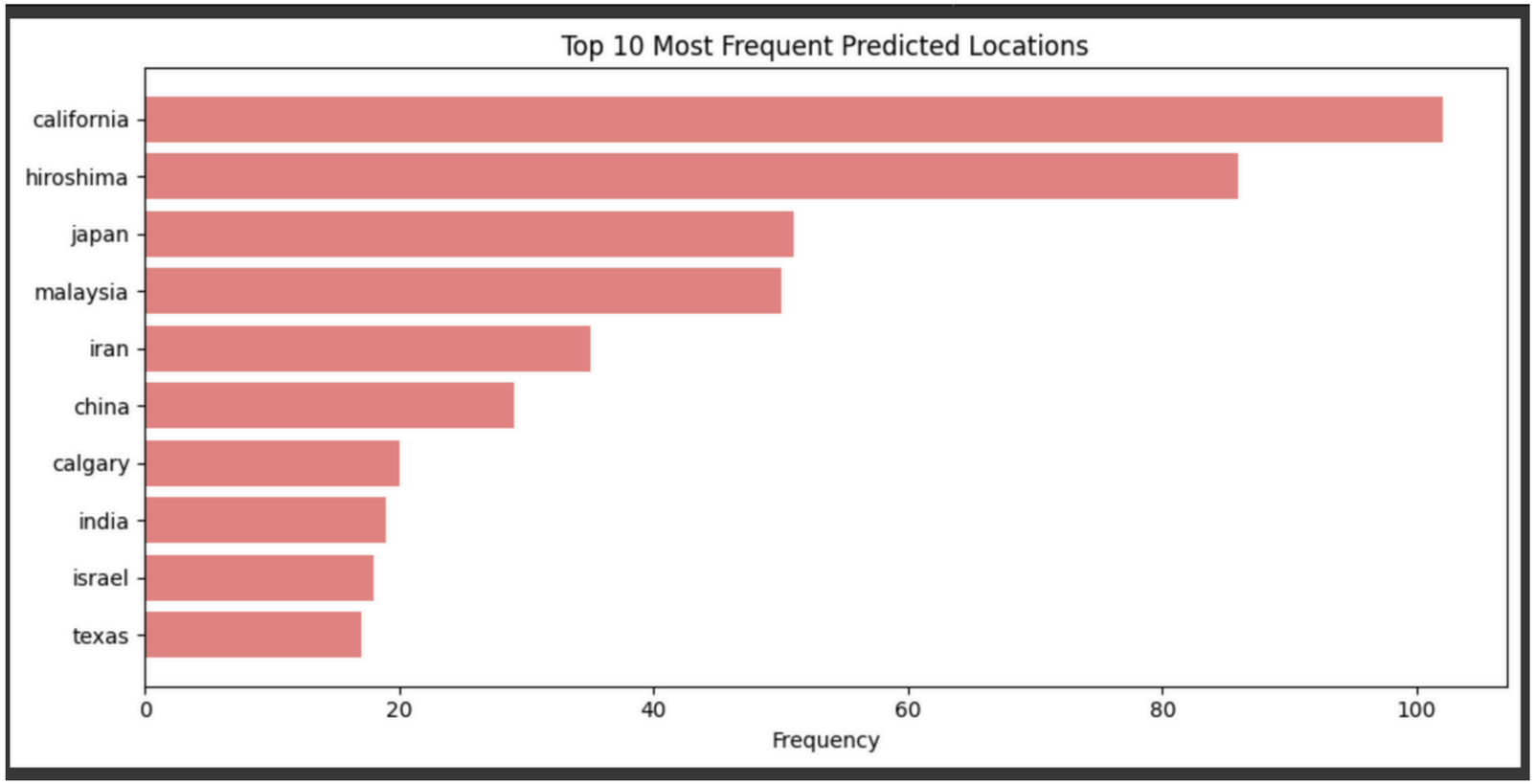


MODEL INTERPRETABILITY

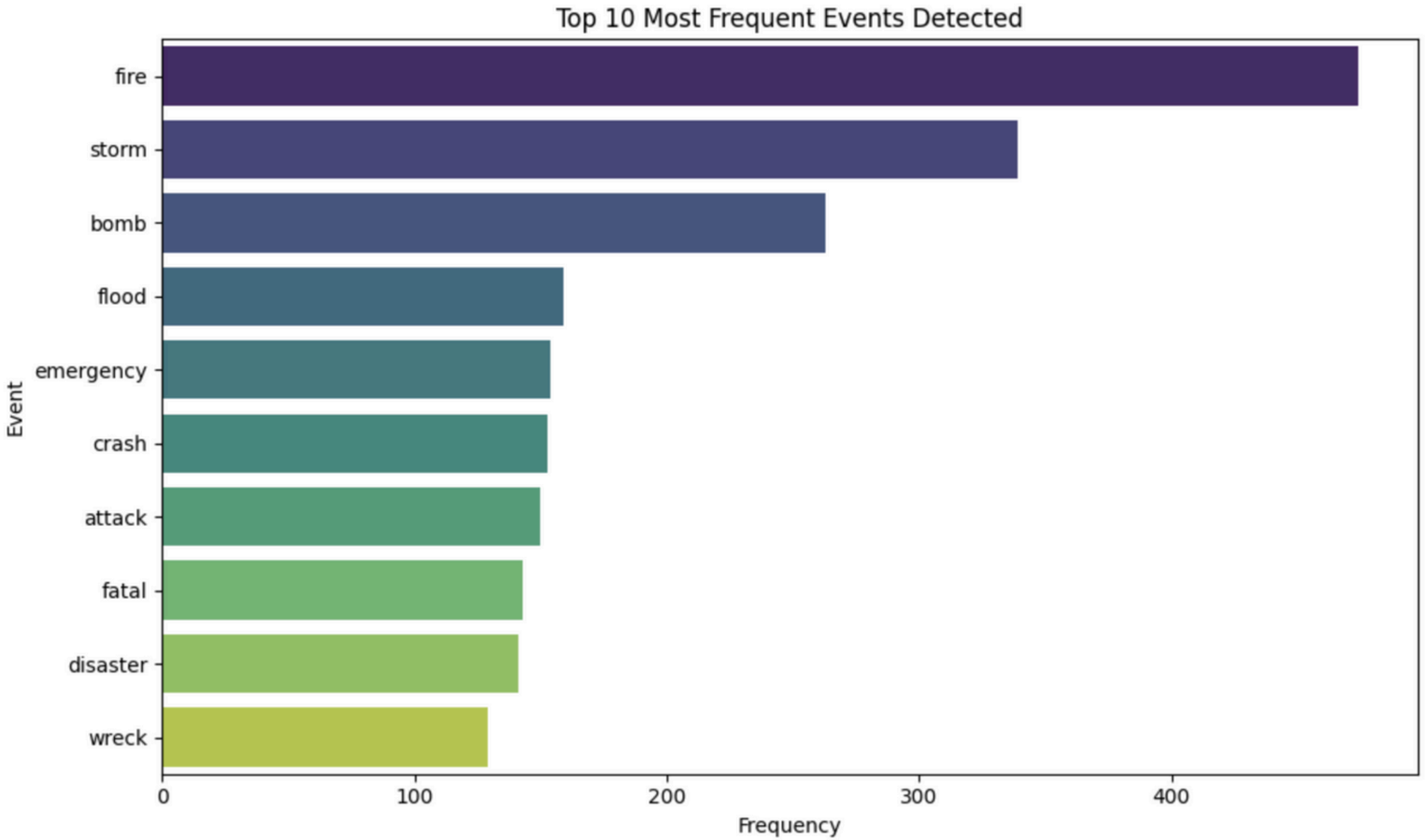
	clean_text	target	prediction_target_rl	prediction_target_nb	tagging_location	extract_event
0	deed reason earthquake allah forgive	1	1	1	[]	[earthquake]
1	forest fire near la ronge sask canada	1	1	1	[la, canada]	[fire]
2	resident ask shelter place notify officer evac...	1	1	1	[]	[evacuation]
3	people receive wildfire evacuation order calif...	1	1	1	[california]	[evacuation, fire, wildfire]
4	send photo ruby alaska smoke wildfire pour school	1	1	1	[alaska]	[fire, smoke, wildfire]
...	...	...	...	...	...	...
7608	giant crane hold bridge collapse nearby home	1	1	1	[]	[collapse]
7609	ariaahrary thetawniest control wild fire calif...	1	1	1	[california]	[fire]
7610	volcano hawaii	1	1	1	[]	[volcano]
7611	police investigate ebike collide car little po...	1	1	1	[portugal]	[collide, injury, police, threat]
7612	late home raze northern california wildfire ab...	1	1	1	[california]	[fire, wildfire]

7613 rows × 6 columns

MODEL INTERPRETABILITY



MODEL INTERPRETABILITY



KESIMPULAN

Model kami dapat mengidentifikasi lokasi dan jenis bencana berdasarkan data teks dari tweet, serta mengklasifikasikan apakah tweet tersebut membahas suatu bencana atau tidak

dengan akurasi classteringnya mencapai 81% pada kedua model (Naive Bayes & Logistic Regression)

Thank you