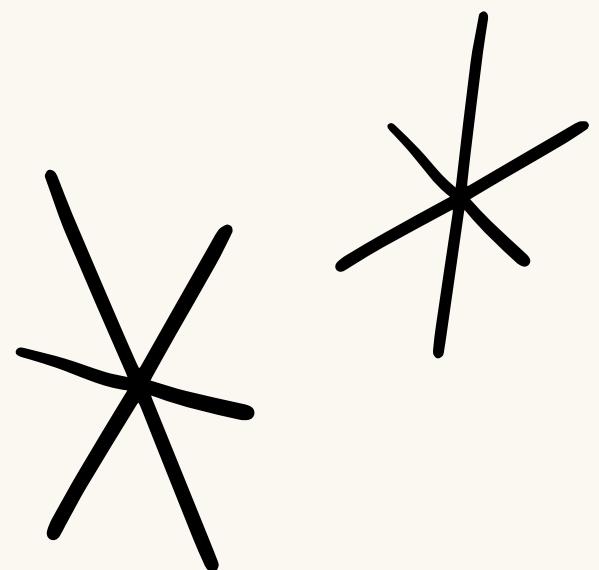


SENTIMEN



MOBILISTRIK

Final Data Competition ISFEST 2024



Tim Dataism



Daniel Wicaksono Nugroho

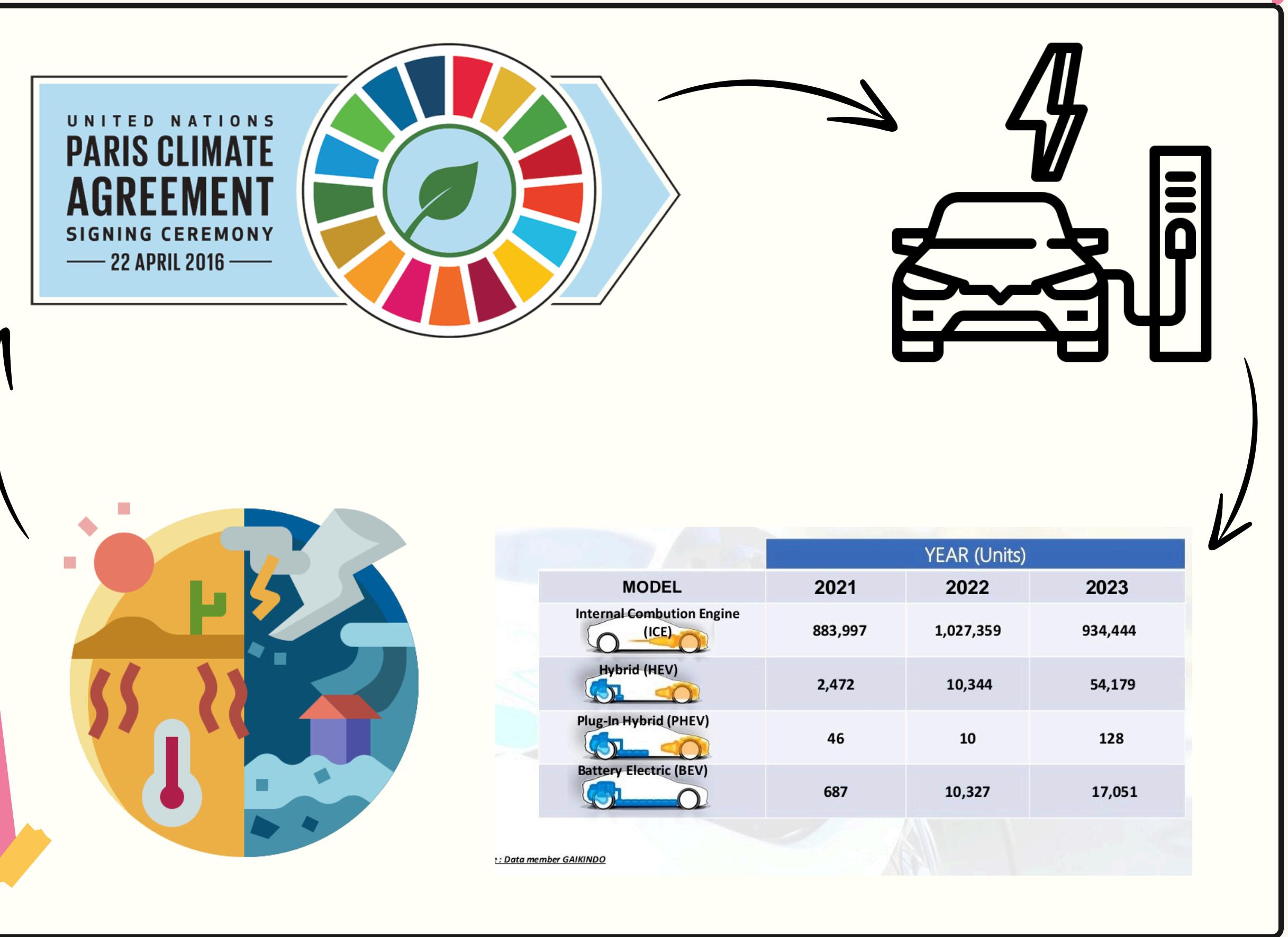


Muhammad Rafli Nugrahasyach



Nadia Salsabila Yasmin

BUSINESS UNDERSTANDING



DATA UNDERSTANDING



- Data berupa daftar opini masyarakat Indonesia tentang mobil listrik
- Data didapatkan dari kolom komentar YouTube
- Data terdiri 1517 baris & 5 kolom

No	Nama Kolom	Keterangan	Tipe Data	Jumlah Missing Value
1	id_komentar	ID unik komentar	Objek	0
2	nama_akun	Nama pengguna	Objek	1
3	tanggal	Waktu komentar dibuat	Objek	0
4	text_cleaning	Komentar opini mengenai mobil listrik	Objek	2
5	sentimen	Klasifikasi sentimen dari opini	Objek	0

Data PREPARATION

I Konversi Tipe Data

- Tipe data 'tanggal' menjadi datetime

II Penanganan Missing Value

- Hapus data kosong tersisa 1514 baris

III Labeling

- Negatif = 0
- Netral = 1
- Positif = 2

I
Konversi
Tipe Data

II
Penanganan
Missing Value

III
Labeling

IV
Cleansing
Data



Data PREPARATION

IV Cleansing Data

1 CASE FOLDING

- Ubah data teks menjadi huruf kecil (lowercase)

2 REMOVE SYMBOL

- Pembersihan data
- Hapus simbol-simbol

3 REMOVAL STOPWORDS

- Hapus kata hubung
- Tetapi susunan teks tetap
- Tidak mengubah maksud sentimen

5 STEMMING

- Ubah kata sesuai Bahasa Indonesia baku

6 VECTORIZER: TF-IDF

- Tentukan bobot setiap kata
- Hitung dengan rumus $W(dt)$
- IDF = Inverse Document Frequency
- TF = Term Frequency

$$W_{dt} = tf_{dt} \times IDF_t$$

I Konversi
Tipe Data

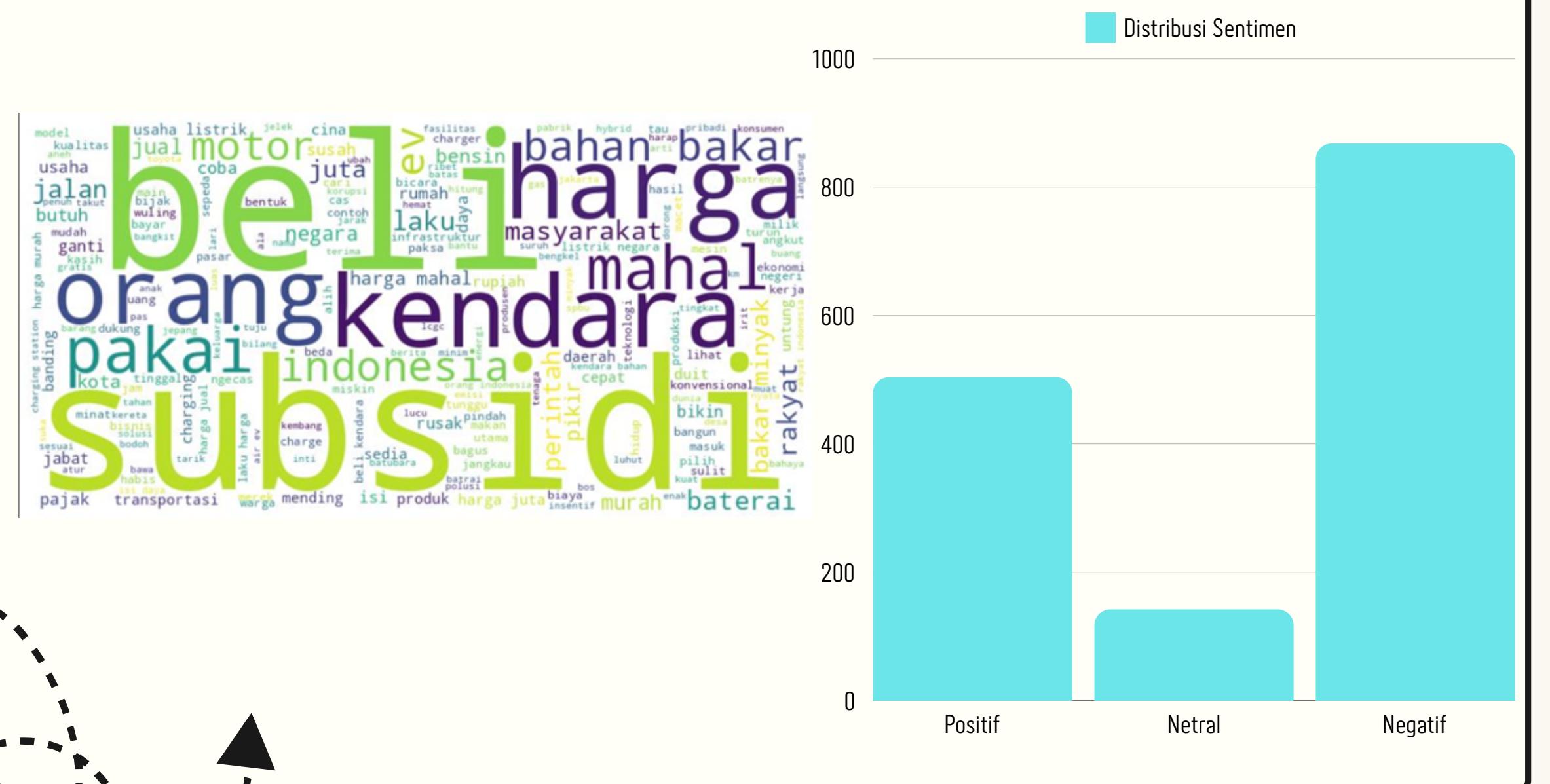
II Penanganan
Missing Value

III Labeling

IV Cleansing
Data



PREDICTION MODEL AND EVALUATION



EXPLORATORY DATA ANALYSIS

- Terjadi ketidakseimbangan distribusi data dengan data sentimen negatif yang dominan
- Walaupun cukup normal karena sentimen netral dengan jumlah sedikit dapat menjadi anomali
- Penanganan moedling data dapat dilakukan dengan metode Synthetic Minority Oversampling Technique (SMOTE)
- Karakteristik data komentar dapat dilihat pada visualisasi wordcloud

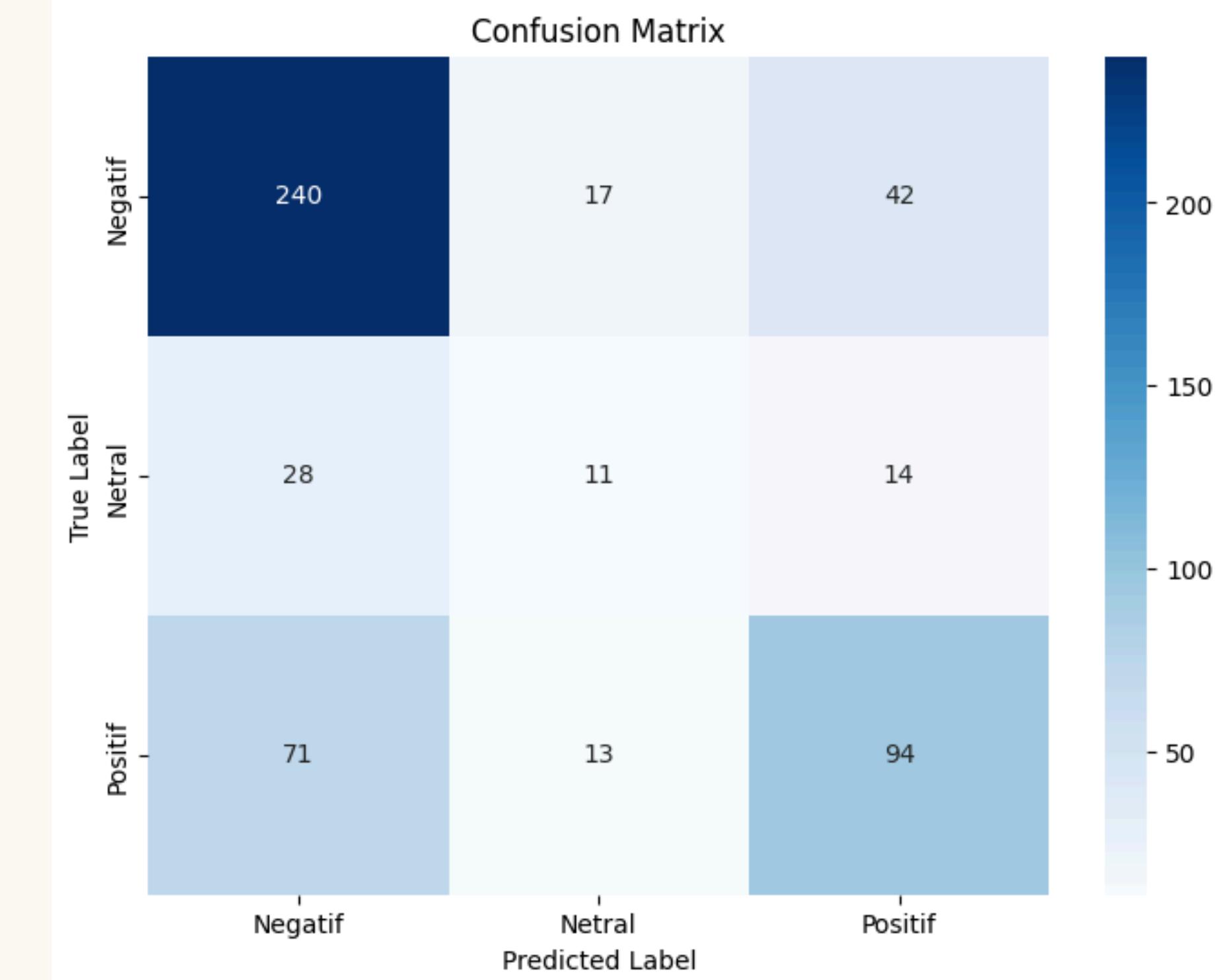
PREDICTION MODEL AND EVALUATION

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.6306	0.6135	0.6306	0.6197
Multinomial Naive Bayes	0.5989	0.6242	0.5989	0.6089
Support Vector Machine	0.6517	0.6248	0.6517	0.5982
Decision Tree	0.5488	0.5291	0.5488	0.5382
Random Forest	0.6359	0.6054	0.6359	0.6035
Extreme Gradient Boosting	0.6279	0.5805	0.6279	0.5976

MODELING & EVALUATION

- Setelah data seimbang, dilakukan modeling data dengan algoritma : Logistic Regression, Multinomial Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, Extreme Gradient Boosting
- Modeling dilakukan untuk menghitung nilai accuracy, precision, recall dan F1-score yang digunakan untuk mengukur performa model.
- Model terbaik ialah Logistic Regression berdasarkan F1-Score

CONFUSION MATRIX LOGISTIC REGRESSION



KESIMPULAN

- **Sentimen negatif** mendominasi
Disusul dengan sentimen positif kemudian netral.
Menunjukkan kekhawatiran atau ketidakpuasan
yang signifikan di kalangan masyarakat
- **Logistic Regression**
61.97%

Model dengan **F1-Score** tertinggi

SARAN

- Menggunakan algoritma
Catboost
CatBoost memiliki mekanisme *Ordered Boosting*
untuk mengurangi *overfitting*, dan memberikan
performa optimal tanpa banyak tuning, bahkan
pada **dataset** yang **tidak seimbang**.
- **Cloglog**
Threshold *Cloglog* berguna dalam
menangani dataset tidak seimbang
karena sifat asimetriknya yang lebih
sensitif terhadap kelas minoritas
- **Deep Neural Network**
Deep neural network memiliki kemampuan
dalam menangani data teks yang
kompleks dan **beragam**

THANK YOU!

