Laporan Teks Analitik

Nama : Risfi Ayu Sandika
NIM : 24917035
Mata Kuliah : Teks Analitik

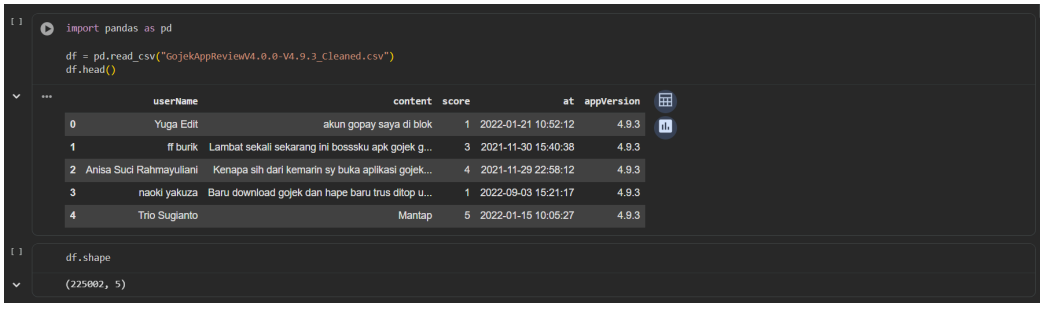
Berkembangnya teknologi secara cepat memberikan kesempatan untuk pengguna melakukan ulasan terhadap suatu teknologi tertentu. Ulasan yang dilakukan oleh pengguna yaitu opini dan pengalaman pengguna itu sangat berharga bagi perusahaan dalam meningkatkan kualitas layanan. Tetapi dengan jumlah data yang sangat besar membuat analisis manual menjadi tidak efesien dan memakan banyak waktu. Text analytic merupakan solusi untuk mengolah informasi penting pada teks secara otomatis dan efesien. Salah satunya adalah text classification terkhusus sentiment analysis yaitu mengklasifikasikan teks ke dalam sentimen yaitu posisi, negatif dan netral. Sehingga tugas akhir ini melakukan text classification berupa sentiment analysis dengan menggunakan dataset ulasan gojek yang didapatkan pada google play store (kaggle) berjumlah 225002 terdiri dari 5 kolom yaitu :

1. Username yaitu nama pengguna yang melakukan ulasan pada gojek
2. Content yaitu isi ulasan yang diberikan pengguna/username terkait gojek
3. Score yaitu nilai/score yang diberikan pengguna/username terkait penilaian gojek yang mana terdiri dari 1-5 dan 5 merupakan nilai tertinggi
4. At yaitu tanggal atau waktu pengguna mengunggah ulasan gojek
5. AppVersion yaitu versi aplikasi yang diberikan oleh pengguna yang memberikan ulasan

Jumlah kelas terdiri dari 3 yaitu :

Kelas	Score	Deskripsi
Negative	1-2	Ulasan keluhan / tidak puas
Neutral	3	Ulasan netral
Positive	4-5	Ulasan puas

Sample Dataset :

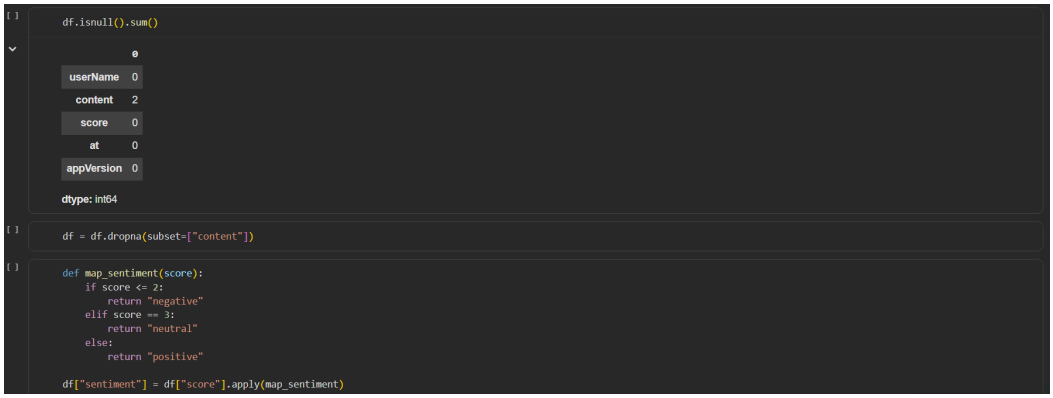


```
import pandas as pd

df = pd.read_csv("GojekAppReviewV4.0.0-V4.9.3_Cleaned.csv")
df.head()
```

	userName	content	score	at	appVersion
0	Yuga Edit	akun gopay saya di blok	1	2022-01-21 10:52:12	4.9.3
1	ff burik	Lambat sekali sekarang ini bosssku apk gojek g...	3	2021-11-30 15:40:38	4.9.3
2	Anisa Suci Rahmayuliani	Kenapa sih dari kemarin sy buka aplikasi gojek...	4	2021-11-29 22:58:12	4.9.3
3	naoki yakuza	Baru download gojek dan hape baru trus ditop u...	1	2022-09-03 15:21:17	4.9.3
4	Trio Sugianto	Mantap	5	2022-01-15 10:05:27	4.9.3

```
df.shape
```

```
(225002, 5)
```

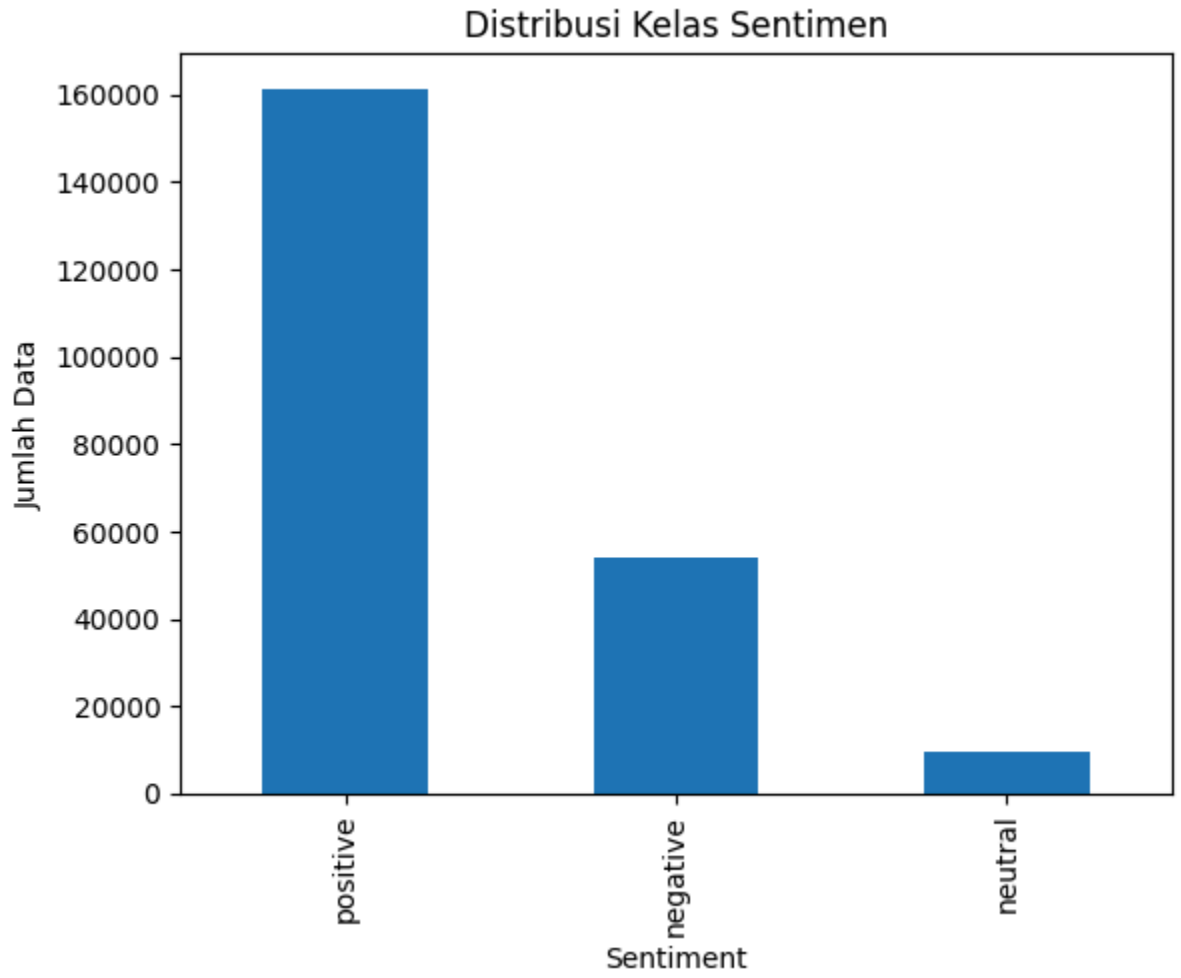


```
[ ] df.isnull().sum()

    0
userName 0
content 2
score 0
at 0
appVersion 0
dtype: int64

[ ] df = df.dropna(subset=["content"])

[ ] def map_sentiment(score):
        if score <= 2:
            return "negative"
        elif score == 3:
            return "neutral"
        else:
            return "positive"

    df["sentiment"] = df["score"].apply(map_sentiment)
```

userName	content	score	at	appVersion
Yuga Edit	akun gopay saya di blok	1	2022-01-21 10:52:12	4.9.3
ff burik	Lambat sekali sekarang ini bosssku apk gojek gk kaya dulu	3	2021-11-30 15:40:38	4.9.3
Anisa Suci Rahmayuliani	Kenapa sih dari kemarin sy buka aplikasi gojek malah keluar sendiri terus Saya kasih bintang 2 dulu kalo sudah normal sy kasih bintang 7	4	2021-11-29 22:58:12	4.9.3

naoki yakuza	Baru download gojek dan hape baru trus ditop u gopay transaksi dialfamart transaksi bloked transaksilapora n di pusat bantuan gak jelas yang ditanyakan apa jawaban lainlama lama gojek dikelola Tokopedia udah nyimpangapa gojek anak bangsa seperti dulu apa punya Tokopedia	1	2022-09-03 15:21:17	4.9.3
Trio Sugianto	Mantap	5	2022-01-15 10:05:27	4.9.3
Arlan Ramlan	Bagus	4	2022-02-01 5:50:40	4.9.3
Slamet Hariyanto	Coba dulu	2	2021-12-10 22:40:45	4.9.3
Hasan Thio	Ok	5	2022-02-01 3:07:45	4.9.3
RAFI BADZLIN	Gimana ini kak pin saya salah terus padahal udah di ubah masih salah	1	2022-12-17 8:56:52	4.9.3

Distribusi kelas sentimen :




```
[ ]
def map_sentiment(score):
    if score <= 2:
        return "negative"
    elif score == 3:
        return "neutral"
    else:
        return "positive"

df["sentiment"] = df["score"].apply(map_sentiment)

[ ]
df["sentiment"].value_counts()
```

sentiment	count
positive	161369
negative	54171
neutral	9460

dtype: int64

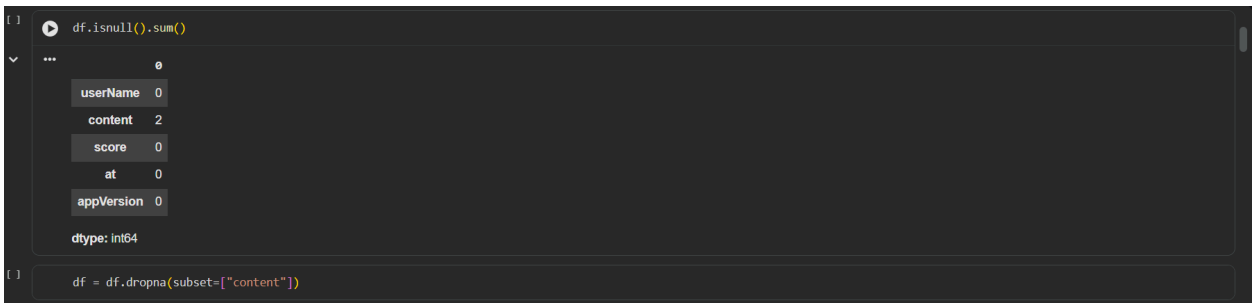
Ini merupakan gambar visualisasi dari distribusi kelas pada dataset ulasan gojek yang mana jumlah data sebagai berikut :

1. Positif : 161369
2. Negatif : 54171

3. Neutral : 9460

Selanjutnya dilakukan tahapan preprocessing sebagai berikut :

1. Dilakukan handling missing value pada setiap kolom untuk mengecek nilai kosong



```
df.isnull().sum()
```

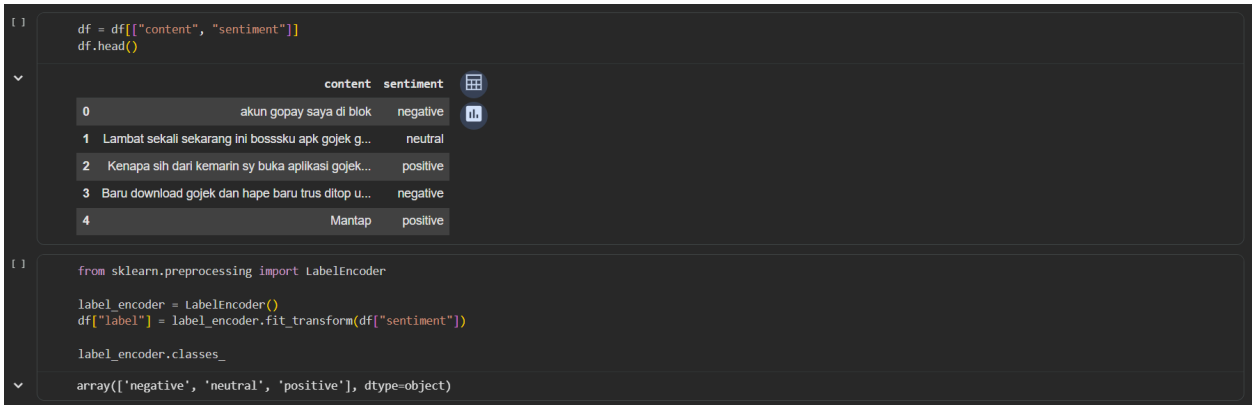
	0
userName	0
content	2
score	0
at	0
appVersion	0

dtype: int64

```
df = df.dropna(subset=["content"])
```

Kemudian setelah melakukan pengecekan ditemukan 2 data kosong pada kolom content dan data tersebut dihapus karena content atau ulasan merupakan fitur utama pada ulasan

2. Dilakukan label encoding karena sentiment teks berupa (negatif,positif dan neutral) maka dikonversi menjadi label numerik menggunakan labelencoder



```
df = df[["content", "sentiment"]]
df.head()
```

	content	sentiment
0	akun gopay saya di blok	negative
1	Lambat sekali sekarang ini bosssku apk gojek g...	neutral
2	Kenapa sih dari kemarin sy buka aplikasi gojek...	positive
3	Baru download gojek dan hape baru trus ditop u...	negative
4	Mantap	positive

```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
df["label"] = label_encoder.fit_transform(df["sentiment"])

label_encoder.classes_
```

```
array(['negative', 'neutral', 'positive'], dtype=object)
```

   a. Negatif = 0
   b. Neutral = 1
   c. Positif = 2

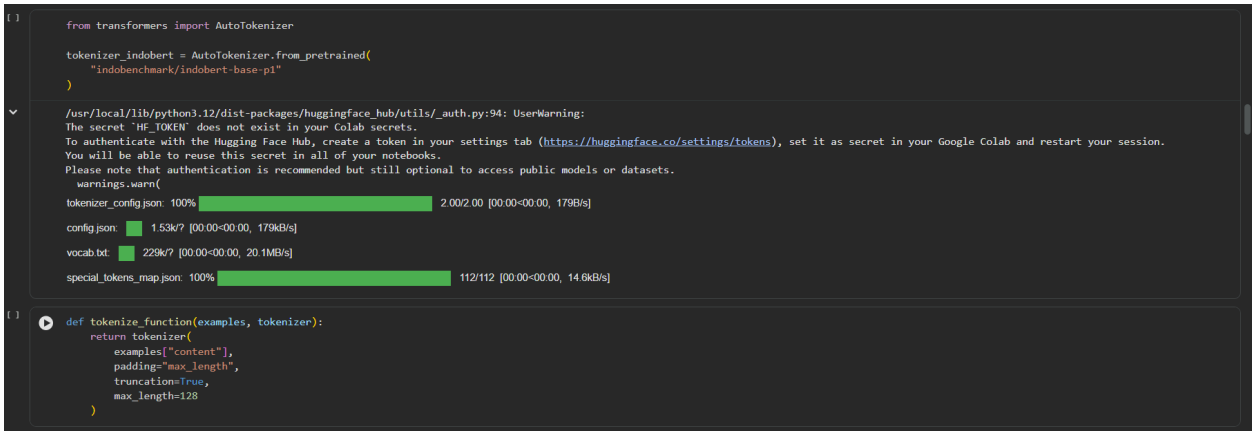
3. Dilakukan pembagian dataset, dataset dibagi menjadi :

```
from sklearn.model_selection import train_test_split

train_df, temp_df = train_test_split(
    df,
    test_size=0.30,
    random_state=42,
    stratify=df["label"]
)

val_df, test_df = train_test_split(
    temp_df,
    test_size=0.50,
    random_state=42,
    stratify=temp_df["label"]
)

len(train_df), len(val_df), len(test_df)
```

```
(157500, 33750, 33750)
```

   a. Training : 70% = 157500
   b. Validation : 15% = 33750
   c. Testing : 15% = 33750

4. Dilakukan tokenisasi menggunakan tokenizer dari indoBert `indobenchmark/indobert-base-p1` dengan parameter :

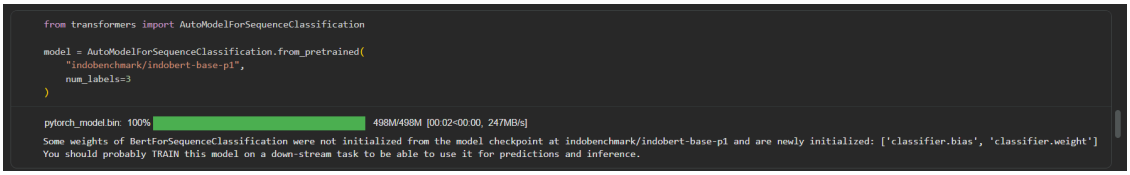


   a. Padding : max_lenght
   b. Truncation : true
   c. Maximum sequence lenght : 128 token

5. Dilakukan pemilihan model yang digunakan yaitu IndoBERT Base (indobenchmark/indobert-base-p1) alasan atas pemilihan model tersebut ialah :
   a. Dilatih untuk bahasa indonesia (sesuai dengan dataset)
   b. Tersedia secara publik di hugging face model hub
   c. Performa baik padan NLP berbahasa indonesia

   Model ini digunakan full fine-tuning dimana seluruh parameternya diperbarui selama proses training.
6. Dilakukan full fine-tuning yaitu seluruh parameter pre-trained language model terperbarui selama proses training, pemilihan full fine tuning karena :
   a. Dataset yang memiliki ukuran yang sangat besar sehingga resiko overfitting relatif lebih kecil
   b. Model menyesuaikan representasi bahasa secara optimal terhadap ulasan
   c. Memberikan performa yang baik

   Full fine tuning dilakukan menggunakan library hugging face transformers sebagai berikut :
   a. Model menggunakan AutoModelForSequenceClassification dengan jumlah label sesuai kelas sentiment



b. Berikut untuk dilakukannya eksploriasi secara otomatis menggunakan random search, hyperparameter dieksplorasi yaitu :
    1. Learning rate : 1e-5 hingga 5e-5
    2. Batch size : 8,16
    3. Number of epoch : 2,3

    Model terbaik berdasarkan hasil random search yaitu :
    1. Learning rate : 8.82
    2. Batch size : 16
    3. Number of epoch : 3

```python
def hp_space(trial):
    return {
        "learning_rate": trial.suggest_float(
            "learning_rate", 1e-5, 5e-5, log=True
        ),
        "per_device_train_batch_size": trial.suggest_categorical(
            "per_device_train_batch_size", [8, 16]
        ),
        "num_train_epochs": trial.suggest_int(
            "num_train_epochs", 2, 3
        ),
    }
```

```python
training_args = TrainingArguments(
    output_dir="./results_random_search",
    eval_strategy="epoch",
    save_strategy="epoch",
    per_device_eval_batch_size=16,
    weight_decay=0.01,
    logging_dir="./logs",
    report_to="none",
    load_best_model_at_end=True,
    metric_for_best_model="f1",
)
```

```python
trainer_rs = Trainer(
    model_init=load_indobert_model,
    args=training_args,
    train_dataset=train_indobert,
    eval_dataset=val_indobert,
    compute_metrics=compute_metrics,
    callbacks=[EarlyStoppingCallback(early_stopping_patience=1)],
)
```

```
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at indobenchmark/indobert-base-p1 and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

```python
best_run = trainer_rs.hyperparameter_search(
    direction="maximize",
    n_trials=1
)
```

```
[I 2026-01-09 15:48:44,982] A new study created in memory with name: no-name-74808e75-7b79-4726-b5dd-c87c722f0ae9
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at indobenchmark/indobert-base-p1 and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
[2953/29532 42:47, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.300600	0.302763	0.905274	0.904342	0.905274	0.889572
2	0.263300	0.305316	0.905156	0.890694	0.905156	0.891035
3	0.233200	0.325171	0.903348	0.888373	0.903348	0.892633

```
[I 2026-01-09 16:31:35,300] Trial 0 finished with value: 3.587702263746124 and parameters: {'learning_rate': 8.824713984787079e-06, 'num_train_epochs': 3, 'seed': 19, 'per_device_train_batch_size': 16}. Best is trial 0 with value: 3.587702263746124.
```

c. Fine tuning dilakukan secara otomatis dengan alur yaitu :
    1. Sistem akan memilih kombinasi hyperparameter
    2. Model IndoBert akan direset
    3. Dilakukan training ulang (fine-tuning)
    4. Evaluasi pada validation set
    5. Menghitung F1
    6. Memilih konfigurasi terbaik

7. Hasil dari evaluasi model yaitu sebagai berikut :

```
from transformers import Trainer

final_trainer = Trainer(
    model=load_indobert_model(),
    args=final_training_args,
    train_dataset=train_indobert,
    eval_dataset=val_indobert,
    compute_metrics=compute_metrics
)

final_trainer.train()
```

```
Some weights of BertForSequenceClassification were not initialized from the model checkpoint at indobenchmark/indobert-base-p1 and are newly initialized: ['classifier.bias', 'classifier.weight']
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
[29532/29532 42:54, Epoch 3/3]
```

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.288200	0.295053	0.904711	0.890911	0.904711	0.890529
2	0.256800	0.306657	0.905422	0.894049	0.905422	0.890100
3	0.234200	0.320729	0.902400	0.886797	0.902400	0.891995

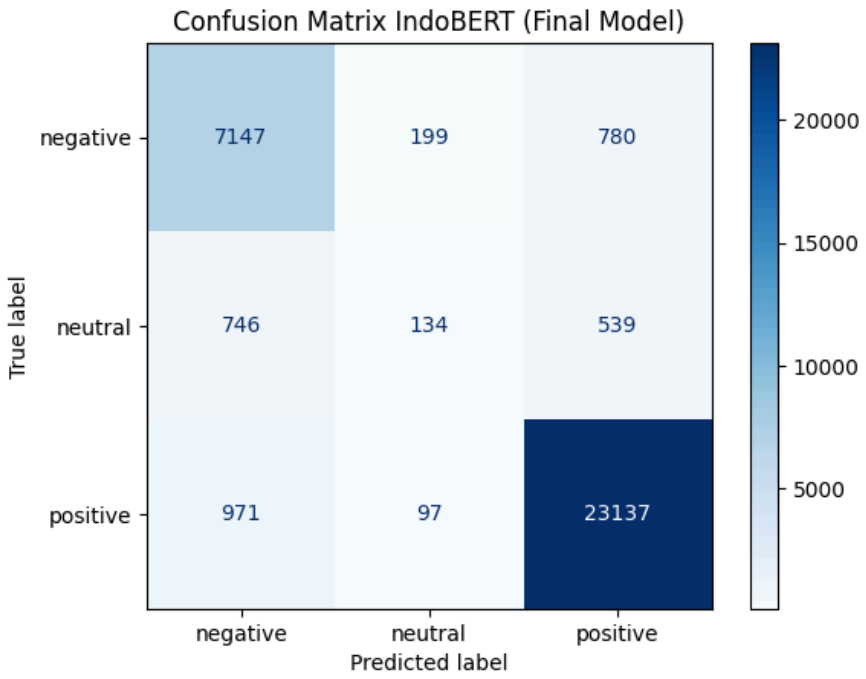
```
TrainOutput(global_step=29532, training_loss=0.2729260066200462, metrics={'train_runtime': 2574.4277, 'train_samples_per_second': 183.536, 'train_steps_per_second': 11.471, 'total_flos': 3.108027246912e+16, 'train_loss': 0.2729260066200462, 'epoch': 3.0})
```

```
final_test_results = final_trainer.evaluate(test_indobert)
final_test_results
```

```
[2110/2110 00:57]
{'eval_loss': 0.3241415023803711,
 'eval_accuracy': 0.9012740740740741,
 'eval_precision': 0.885739008269715,
 'eval_recall': 0.9012740740740741,
 'eval_f1': 0.8906632342106588,
 'eval_runtime': 57.1454,
 'eval_samples_per_second': 590.599,
 'eval_steps_per_second': 36.923,
 'epoch': 3.0}
```

Dari eksperimen ini dapat dilihat bahwa dataset bersifat imbalanced dan yang mana metrik evaluasi utama yang dipilih yaitu F1-score karena F1 mempertimbangkan precision dan recall secara bersamaan serta memberikan bobot sesuai proporsi kelas. Model menunjukan performa yang stabil sampai epoch 3 dengan nilai F1 mendekati 0.89

8. Ini merupakan hasil evaluasi dari model indoBERT yang telah melakukan full fine tuning menggunakan dataset yang tersedia dapat dijelaskan sebagai berikut :
    1. Kelas negatif : 7147 ulasan negatif dapat diprediksi benar sebagai negatif, 199 salah prediksi sebagai netral, 780 salah prediksi sebagai positif
    2. Kelas neutral : 134 benar diprediksi, 746 salah prediksi menjadi negatif dan 539 slaah prediksi menjadi positif (penyebab dikarenakan jumlah data neutral jauh lebih sedikit dibanding yang lain)
    3. Kelas positif : 23137 diprediksi benar, 971 diprediksi salah menjadi negatif dan 97 salah prediksi sebagai netral



9. Analisis yang dapat di tarik adalah :
    a. Indobert dapat menangkap dataset bahasa indonesia dengan baik
    b. F1 score baik dan meningkat pada epoch terkahir menandakan model tidka mengalami overfitting signifikan
    c. F1 score sebagai evaluasi model yang tidak hanya berfokus pada kelas mayoritas tetapi menjaga performa kelas minoritas

10. Kesimpulan yang dapat di tarik adalah :
    a. IndoBERT dengan full fine tuning mempu dan efektif digunakan untuk sentimen teks bahasa indonesia
    b. Fine tuning dengan hyperparameter yang tepat menghasilkan performa yang baik walaupun data tidak seimbang
    c. F1 score digunakan sebagai metrik evaluasi yang paling sesuai untuk kondisi dataset
    d. Pengembangan selanjutnya dapat menambahkan model pembanding dan implementasi LORA