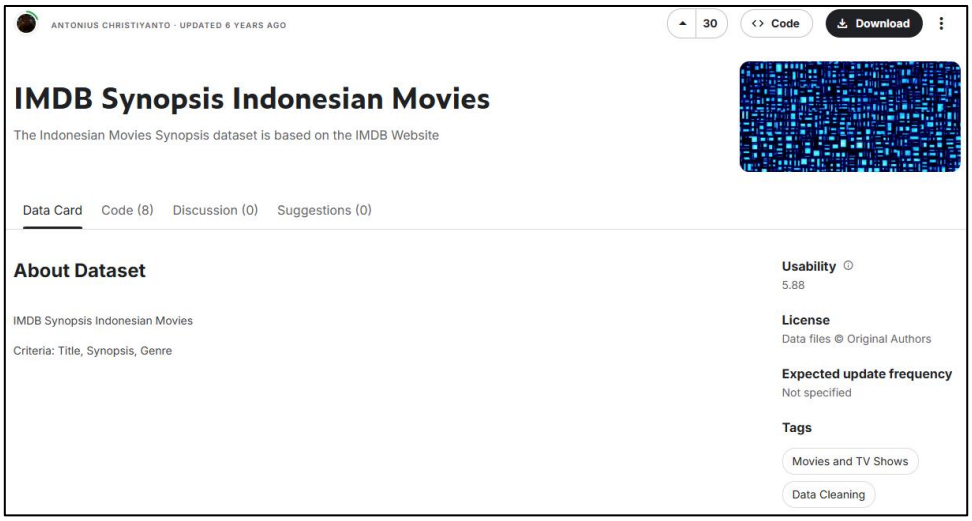
Nama Dosen : Teguh Iman Hermanto, M.Kom
Mata Kuliah : Machine Learning 2
Pembahasan : Basic NLP
Pokok Pemb : Membuat model Natural Language Processing Sederhana

DOWNLOAD DATASET

Import dataset ke dalam google colab menggunakan Alamat dataset berikut (import dataset menggunakn kaggle.json):
https://www.kaggle.com/datasets/antoniuscs/imdb-synopsis-indonesian-movies



IMPORT LIBARRY DAN LOAD DATASET

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from tensorflow.keras.preprocessing.text import Tokenizer
5 from tensorflow.keras.preprocessing.sequence import pad_sequences
6 import tensorflow as tf
7 from tensorflow.keras.utils import plot_model
```

```
1 df = pd.read_csv('imdb_indonesian_movies_2.csv')
```

PREPROCESSING DATA

```
df = df.drop(columns=['judul_film'])
```

```
df.head()
```

	ringkasan_sinopsis	genre
0	Raden Mas Said putra sulung Tumenggung Wilarik...	Drama
1	Soe Hok Gie adalah seorang aktivis yang hidup ...	Drama
2	Guru Bangsa Tjokroaminoto menceritakan tentang...	Drama
3	POL menceritakan kisah hidup yang luar biasa d...	Drama
4	Perjalanan pahlawan Indonesia KH Ahmad Dahlan ...	Drama

```
category = pd.get_dummies(df.genre)
df_baru = pd.concat([df, category], axis=1)
df_baru = df_baru.drop(columns='genre')
df_baru
```

	ringkasan_sinopsis	Drama	Horor	Komedi	Laga	Romantis
0	Raden Mas Said putra sulung Tumenggung Wilarik...	True	False	False	False	False
1	Soe Hok Gie adalah seorang aktivis yang hidup ...	True	False	False	False	False
2	Guru Bangsa Tjokroaminoto menceritakan tentang...	True	False	False	False	False
3	POL menceritakan kisah hidup yang luar biasa d...	True	False	False	False	False
4	Perjalanan pahlawan Indonesia KH Ahmad Dahlan ...	True	False	False	False	False
...	...	...	...	...	...	...
1000	Winter in Tokyo berpusat pada kehidupan Ishida...	False	False	False	False	True
1001	Markonah melarikan diri ke Jakarta karena akan...	False	False	False	False	True
1002	Tempat aking lebih dari 36 jam, Last Night ada...	False	False	False	False	True
1003	Proyek baru ini adalah tentang seorang lelaki ...	False	False	False	False	True
1004	Atika (Meriam Bellina) mantan penyanyi tenar, ...	False	False	False	False	True

1005 rows × 6 columns

```
sinopsis = df_baru['ringkasan_sinopsis'].values
label = df_baru[['Drama', 'Horor', 'Komedi', 'Laga', 'Romantis']].values
```

```
sinopsis_latih, sinopsis_test, label_latih, label_test = train_test_split(sinopsis, label, test_size=0.2)
```

```
tokenizer = Tokenizer(num_words=5000, oov_token='x')
tokenizer.fit_on_texts(sinopsis_latih)
tokenizer.fit_on_texts(sinopsis_test)

sekuens_latih = tokenizer.texts_to_sequences(sinopsis_latih)
sekuens_test = tokenizer.texts_to_sequences(sinopsis_test)

padded_latih = pad_sequences(sekuens_latih)
padded_test = pad_sequences(sekuens_test)
```

MODELING

```
input_length = padded_latih.shape[1]

model = tf.keras.Sequential([
    tf.keras.layers.Embedding(input_dim=5000, output_dim=16),
    tf.keras.layers.LSTM(64),
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(5, activation='softmax')
])
```

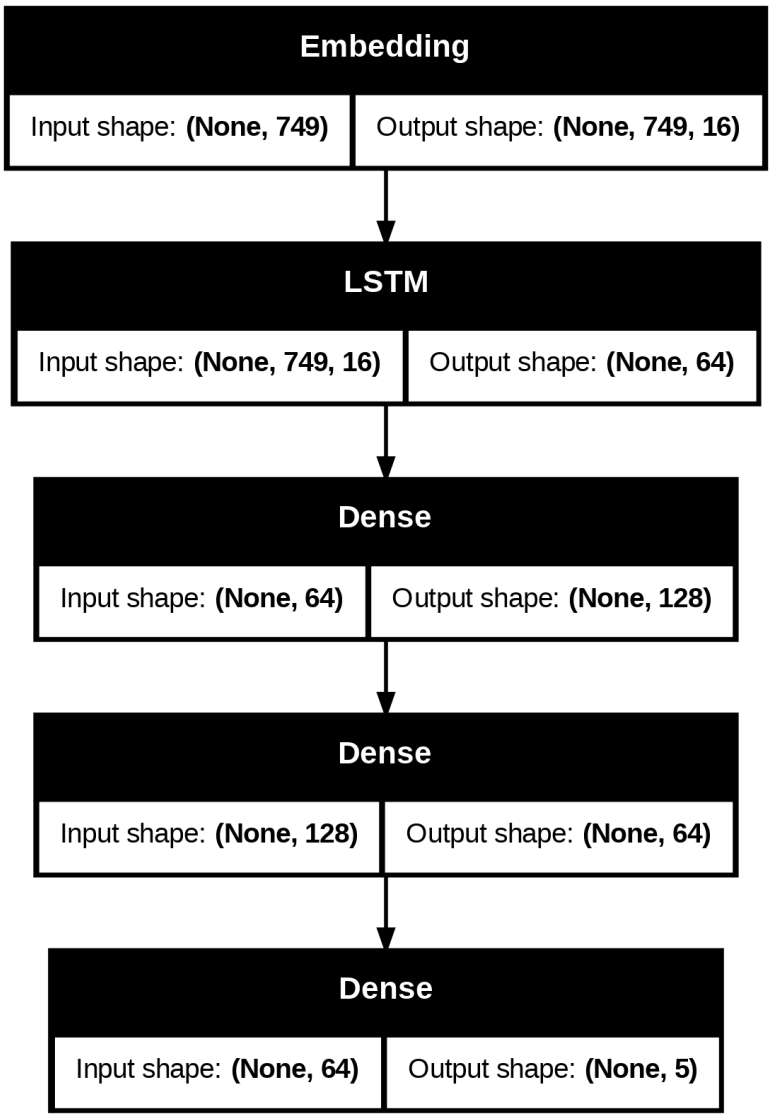
```
model.build(input_shape=(None, input_length))
```

```
model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
```

```
model.summary()
```

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, 749, 16)	80,000
lstm_6 (LSTM)	(None, 64)	20,736
dense_17 (Dense)	(None, 128)	8,320
dense_18 (Dense)	(None, 64)	8,256
dense_19 (Dense)	(None, 5)	325

```
plot_model(model, show_shapes = True)
```




```
history = model.fit(padded_latih, label_latih, epochs=15,
                    validation_data=(padded_test, label_test))
```

SIMULASI MODEL

```
hasil = ["Kisah cinta antara dua yuppies urban yang sinis yang harus saling berhadapan sebagai pengacara di ruang sidang"]
hasil = tokenizer.texts_to_sequences(hasil)
hasil = pad_sequences(hasil)
#print(model.predict(hasil))
hasil = np.argmax(model.predict(hasil), axis=1)
if hasil == [0]:
  print('Drama')
if hasil == [1]:
  print('Horor')
if hasil == [2]:
  print('Komedi')
if hasil == [3]:
  print('Laga')
if hasil == [4]:
  print('Romantis')
```