

## ✓ Lab 5 Transformers

```
#!pip install transformers
```

```
#!pip install tf-keras
```

```
!pip install opendatasets
```

```
🔗 Collecting opendatasets
  Downloading opendatasets-0.1.22-py3-none-any.whl.metadata (9.2 kB)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from opendatasets) (4.67.1)
Requirement already satisfied: kaggle in /usr/local/lib/python3.11/dist-packages (from opendatasets) (1.6.17)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from opendatasets) (8.1.8)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (1.17.0)
Requirement already satisfied: certifi>=2023.7.22 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2025.1.31)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.32.3)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (8.0.4)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (2.3.0)
Requirement already satisfied: bleach in /usr/local/lib/python3.11/dist-packages (from kaggle->opendatasets) (6.2.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.11/dist-packages (from bleach->kaggle->opendatasets) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.11/dist-packages (from python-slugify->kaggle->opendatasets) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->kaggle->opendatasets) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->kaggle->opendatasets) (3.10)
Downloading opendatasets-0.1.22-py3-none-any.whl (15 kB)
Installing collected packages: opendatasets
Successfully installed opendatasets-0.1.22
```

## ✓ Sentiment Analysis Using BERT model

```
import opendatasets as od
```

```
od.download("https://www.kaggle.com/datasets/ahmedabdulhamid/reviews-dataset/data?select=TestReviews.csv")
```

```
🔗 Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username: alenscaria
Your Kaggle Key: .....
Dataset URL: https://www.kaggle.com/datasets/ahmedabdulhamid/reviews-dataset
Downloading reviews-dataset.zip to ./reviews-dataset
100%|██████████| 3.90M/3.90M [00:00<00:00, 111MB/s]
```

```
from transformers import pipeline
import pandas as pd
```

```
review_dataset.head()
```

◀ 1 2 3 4 5 6 7 8 9 10 ▶

⏏ No model was supplied, defaulted to distilbert/distilbert-base-uncased-finetuned-sst-2-english and revision 714eb0f (<https://huggingface.co/distilbert/distilbert-base-uncased>). Using a pipeline without specifying a model name and revision in production is not recommended.  
/usr/local/lib/python3.11/dist-packages/huggingface\_hub/utils/\_auth.py:94: UserWarning:  
The secret `HF\_TOKEN` does not exist in your Colab secrets.  
To authenticate with the Hugging Face Hub, create a token in your settings tab (<https://huggingface.co/settings/tokens>), set it as secret in your Google Colab and restart your  
You will be able to reuse this secret in all of your notebooks.  
Please note that authentication is recommended but still optional to access public models or datasets.

```
predictions = sentiment_pipeline(text, truncation=True)
```

```
The predictions for 100 entries are
```


```
[{'label': 'POSITIVE', 'score': 0.9997124075889587}, {'label': 'POSITIVE', 'score': 0.9998700618743896}, {'label': 'POSITIVE', 'score': 0.9996854066848755}, {'label': 'POSITI
```

```
pred = [1 if d['label'] == 'POSITIVE' else 0 for d in predictions]
```

```
pred = np.array(pred)
```

```
from sklearn.metrics import accuracy_score  
y = np.array(review_dataset['class'])
```


```
print(f'The accuracy score for predicting the sentiment is : ', accuracy_score(y, pred))
```

 The accuracy score for predicting the sentiment is : 0.95


## ▼ Summarize a Para

```
import opendatasets as od
```

```
od.download("https://www.kaggle.com/datasets/gpreda/bbc-news?select=bbc_news.csv")
```

 Please provide your Kaggle credentials to download this dataset. Learn more: <http://bit.ly/kaggle-creds>  
Your Kaggle username: alenscaria  
Your Kaggle Key: .....  
Dataset URL: <https://www.kaggle.com/datasets/gpreda/bbc-news>  
Downloading bbc-news.zip to ./bbc-news  
100%|██████████| 3.64M/3.64M [00:00<00:00, 5.58MB/s]

```
!pip install torch transformers
```

 Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (2.5.1+cu124)  
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.48.3)  
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch) (3.17.0)  
Requirement already satisfied: typing-extensions>=4.8.0 in /usr/local/lib/python3.11/dist-packages (from torch) (4.12.2)  
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch) (3.4.2)  
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.6)  
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch) (2024.10.0)  
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch)  
 Downloading nvidia\_cuda\_nvrtc\_cu12-12.4.127-py3-none-manylinux2014\_x86\_64.whl.metadata (1.5 kB)  
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch)  
 Downloading nvidia\_cuda\_runtime\_cu12-12.4.127-py3-none-manylinux2014\_x86\_64.whl.metadata (1.5 kB)  
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch)  
 Downloading nvidia\_cuda\_cupti\_cu12-12.4.127-py3-none-manylinux2014\_x86\_64.whl.metadata (1.6 kB)  
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch)  
 Downloading nvidia\_cudnn\_cu12-9.1.0.70-py3-none-manylinux2014\_x86\_64.whl.metadata (1.6 kB)  
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch)  
 Downloading nvidia\_cublas\_cu12-12.4.5.8-py3-none-manylinux2014\_x86\_64.whl.metadata (1.5 kB)  
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch)  
 Downloading nvidia\_cufft\_cu12-11.2.1.3-py3-none-manylinux2014\_x86\_64.whl.metadata (1.5 kB)  
Collecting nvidia-curand-cu12==10.3.5.147 (from torch)  
 Downloading nvidia\_curand\_cu12-10.3.5.147-py3-none-manylinux2014\_x86\_64.whl.metadata (1.5 kB)  
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch)  
 Downloading nvidia\_cusolver\_cu12-11.6.1.9-py3-none-manylinux2014\_x86\_64.whl.metadata (1.6 kB)

```

Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch) (12.4.127)
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.1.0 in /usr/local/lib/python3.11/dist-packages (from torch) (3.1.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch) (1.3.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.24.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.28.1)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (1.26.4)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11.6)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.21.0)
Requirement already satisfied: safetensors>=0.4.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.31)
Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4 MB)
  363.4/363.4 MB 3.4 MB/s eta 0:00:00
Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (13.8 MB)
  13.8/13.8 MB 24.6 MB/s eta 0:00:00
Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
  24.6/24.6 MB 24.4 MB/s eta 0:00:00
Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
  883.7/883.7 kB 15.6 MB/s eta 0:00:00
Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
  664.8/664.8 MB 2.6 MB/s eta 0:00:00
Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)

```

```

import torch
from transformers import AutoTokenizer, AutoModelWithLMHead

```

```

# T5: Text To Text Transfer Transformer
tokenizer=AutoTokenizer.from_pretrained('T5-base')
model=AutoModelWithLMHead.from_pretrained('T5-base', return_dict=True)

```

```
⚙ /usr/local/lib/python3.11/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(

config.json: 100% ██████████ 1.21k/1.21k [00:00<00:00, 27.8kB/s]

spiece.model: 100% ██████████ 792k/792k [00:00<00:00, 2.74MB/s]

tokenizer.json: 100% ██████████ 1.39M/1.39M [00:00<00:00, 4.73MB/s]

/usr/local/lib/python3.11/dist-packages/transformers/models/auto/modeling_auto.py:1881: FutureWarning: The class `AutoModelWithLMHead` is deprecated and will be removed in a f
warnings.warn(

model.safetensors: 100% ██████████ 892M/892M [00:10<00:00, 98.6MB/s]

generation_config.json: 100% ██████████ 147/147 [00:00<00:00, 7.62kB/s]
```

```
sequence = ("Data science is an interdisciplinary field[10] focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to

inputs=tokenizer.encode("sumarize: " +sequence,return_tensors='pt', max_length=512, truncation=True)

output = model.generate(inputs, min_length=80, max_length=100)

summary=tokenizer.decode(output[0])
print(summary)
```

```
⚙ <pad> data science is an interdisciplinary field focused on extracting knowledge from typically large data sets . it incorporates skills from computer science, statistics, inf
```

```
import pandas as pd

news_dataset = pd.read_csv('/content/bbc-news/bbc_news.csv', nrows = 4)

news_dataset = news_dataset[['title', 'description']]
news_dataset.head()
```

```
⚙
```

	title	description
0	Ukraine: Angry Zelensky vows to punish Russian...	The Ukrainian president says the country will ...
1	War in Ukraine: Taking cover in a town under a...	Jeremy Bowen was on the frontline in Irpin, as...
2	Ukraine war 'catastrophic for global food'	One of the world's biggest fertiliser firms sa...
3	Manchester Arena bombing: Saffie Roussos's par...	The parents of the Manchester Arena bombing's ...

```
summaries = []
```

```

for description in news_dataset['description']:
    sequence = description

    inputs = tokenizer.encode("summarize: " + sequence, return_tensors='pt', max_length=512, truncation=True)

    output = model.generate(inputs, min_length=80, max_length=100)

    summary = tokenizer.decode(output[0], skip_special_tokens=True)

    summaries.append(summary)

news_dataset['summary'] = summaries

news_dataset[['description', 'summary']].head()

```



	description	summary
0	The Ukrainian president says the country will ...	president says country will not forgive or for...
1	Jeremy Bowen was on the frontline in Irpin, as...	Jeremy Bowen was on the frontline in Irpin, as...
2	One of the world's biggest fertiliser firms sa...	one of the world's biggest fertiliser firms sa...
3	The parents of the Manchester Arena bombing's ...	the parents of the youngest victim of the Manc...

## Text Continuation using GPT 2

```

import torch
from transformers import GPT2LMHeadModel, GPT2Tokenizer

model_name = "gpt2"
tokenizer = GPT2Tokenizer.from_pretrained(model_name)
model = GPT2LMHeadModel.from_pretrained(model_name)

model.eval()

```



```
merges.txt: 100% 456k/456k [00:00<00:00, 2.29MB/s]
```

```
config.json: 100% 665/665 [00:00<00:00, 15.8kB/s]
```

```
generation_config.json: 100% ██████████ 124/124 [00:00<00:00, 10.0kB/s]
```

◀ ▶

```
def generate_text(prompt, max_length=100, temperature=0.8, top_k=50):
    input_ids = tokenizer.encode(prompt, return_tensors="pt")
    output = model.generate(
        input_ids,
        max_length=max_length,
        temperature=temperature,
        top_k=top_k,
        pad_token_id=tokenizer.eos_token_id,
        do_sample=True
    )
    generated_text = tokenizer.decode(output[0], skip_special_tokens=True)
    return generated_text
```

```
prompt = "Following is a picture of a dog.This is a "  
generate_text(prompt)
```

⚡ 'Following is a picture of a dog.This is a \xa0photof an \xa0old \xa0horse that is in storage for future storage in a\xa0retail store. This old \xa0horse is\xa0in the\xa0reta  
il store\xa0and has not been removed by the seller since June.This may also be what a seller was selling in a box (i.e., in the same\xa0store).This is a \xa0photof a horse, o

## ▼ Translation : Eng to Fre

```
od.download("https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset?select=en-fr.csv")
```

⚡ Please provide your Kaggle credentials to download this dataset. Learn more: <http://bit.ly/kaggle-creds>  
Your Kaggle username: alenscaria  
Your Kaggle Key: .....  
Dataset URL: <https://www.kaggle.com/datasets/dhruvildave/en-fr-translation-dataset>  
Downloading en-fr-translation-dataset.zip to ./en-fr-translation-dataset  
100%|██████████| 2.54G/2.54G [01:09<00:00, 39.6MB/s]

```
en_fre_dataset = pd.read_csv("/content/en-fr-translation-dataset/en-fr.csv", nrows = 10)
```

```
en_fre_dataset.head(10)
```

⚡

	en	fr
0	Changing Lives   Changing Society   How It Wor...	Il a transformé notre vie   Il a transformé la...
1	Site map	Plan du site
2	Feedback	Rétroaction
3	Credits	Crédits
4	Français	English
5	What is light ?	Qu'est-ce que la lumière?
6	The white light spectrum Codes in the light Th...	La découverte du spectre de la lumière blanche...
7	The sky of the first inhabitants A contemporar...	Le ciel des premiers habitants La vision conte...
8	Cartoon	Bande dessinée
9	Links	Liens

```
from transformers import pipeline
```

```
translator = pipeline("translation_en_to_fr")
```


```
translations = []
```



```
for english in en_fre_dataset['en']:
    translation = translator(english, max_length= 40)
    translations.append(translation[0]['translation_text'])

en_fre_dataset['translation'] = translations

en_fre_dataset.head()
```

 No model was supplied, defaulted to google-t5/t5-base and revision a9723ea (<https://huggingface.co/google-t5/t5-base>).  
Using a pipeline without specifying a model name and revision in production is not recommended.  
Device set to use cpu  
Your input\_length: 51 is bigger than 0.9 \* max\_length: 40. You might consider increasing your max\_length manually, e.g. translator('...', max\_length=400)

	en	fr	translation
0	Changing Lives   Changing Society   How It Wor...	Il a transformé notre vie   Il a transformé la...	Évolution des vies   Évolution de la société  ...
1	Site map	Plan du site	Plan du site
2	Feedback	Rétroaction	Rétroaction