Notebook

October 20, 2025

[3]:
```
!pip install datasets
```

```
Collecting datasets
  Downloading datasets-3.3.2-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.17.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-
packages (from datasets) (1.26.4)
Requirement already satisfied: pyarrow>=15.0.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Collecting dill<0.3.9,>=0.3.0 (from datasets)
  Downloading dill-0.3.8-py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages
(from datasets) (2.2.2)
Requirement already satisfied: requests>=2.32.2 in
/usr/local/lib/python3.11/dist-packages (from datasets) (2.32.3)
Requirement already satisfied: tqdm>=4.66.3 in /usr/local/lib/python3.11/dist-
packages (from datasets) (4.67.1)
Collecting xxhash (from datasets)
  Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(12 kB)
Collecting multiprocess<0.70.17 (from datasets)
  Downloading multiprocess-0.70.16-py311-none-any.whl.metadata (7.2 kB)
Requirement already satisfied: fsspec<=2024.12.0,>=2023.1.0 in
/usr/local/lib/python3.11/dist-packages (from
fsspec[http]<=2024.12.0,>=2023.1.0->datasets) (2024.10.0)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-
packages (from datasets) (3.11.13)
Requirement already satisfied: huggingface-hub>=0.24.0 in
/usr/local/lib/python3.11/dist-packages (from datasets) (0.28.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-
packages (from datasets) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-
packages (from datasets) (6.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (2.4.6)
Requirement already satisfied: aiosignal>=1.1.2 in
```

```
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-
packages (from aiohttp->datasets) (25.1.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.1.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.18.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.24.0->datasets)
(4.12.2)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(3.4.1)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-
packages (from requests>=2.32.2->datasets) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests>=2.32.2->datasets)
(2025.1.31)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas->datasets) (2025.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas->datasets) (1.17.0)
Downloading datasets-3.3.2-py3-none-any.whl (485 kB)
                        485.4/485.4 kB
19.4 MB/s eta 0:00:00
Downloading dill-0.3.8-py3-none-any.whl (116 kB)
                        116.3/116.3 kB
8.0 MB/s eta 0:00:00
Downloading multiprocess-0.70.16-py311-none-any.whl (143 kB)
                        143.5/143.5 kB
10.2 MB/s eta 0:00:00
Downloading
xxhash-3.5.0-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
                        194.8/194.8 kB
10.3 MB/s eta 0:00:00
Installing collected packages: xxhash, dill, multiprocess, datasets
Successfully installed datasets-3.3.2 dill-0.3.8 multiprocess-0.70.16
```

```
xxhash-3.5.0
```

[1]:
```
import pandas as pd
from huggingface_hub import list_datasets
```

[4]:
```
from datasets import load_dataset
emotions = load_dataset('emotion')
emotions.set_format(type='pandas')
```

```
/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 session.
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(
README.md:   0%|          | 0.00/9.05k [00:00<?, ?B/s]
train-00000-of-00001.parquet:   0%|          | 0.00/1.03M [00:00<?, ?B/s]
validation-00000-of-00001.parquet:   0%|          | 0.00/127k [00:00<?, ?B/s]
test-00000-of-00001.parquet:   0%|          | 0.00/129k [00:00<?, ?B/s]
Generating train split:   0%|          | 0/16000 [00:00<?, ? examples/s]
Generating validation split:   0%|          | 0/2000 [00:00<?, ? examples/s]
Generating test split:   0%|          | 0/2000 [00:00<?, ? examples/s]
```

[5]:
```
df = emotions['train'][:]
```

[6]:
```
df.head()
```

[6]:

	text	label
0	i didnt feel humiliated	0
1	i can go from feeling so hopeless to so damned…	0
2	im grabbing a minute to post i feel greedy wrong	3
3	i am ever feeling nostalgic about the fireplac…	2
4	i am feeling grouchy	3

[7]:
```
classes = emotions['train'].features['label'].names
classes
```

[7]:
```
['sadness', 'joy', 'love', 'anger', 'fear', 'surprise']
```

[8]:
```
df['label_name'] = df['label'].apply(lambda x: classes[x])
```

[9]:
```
df.head()
```

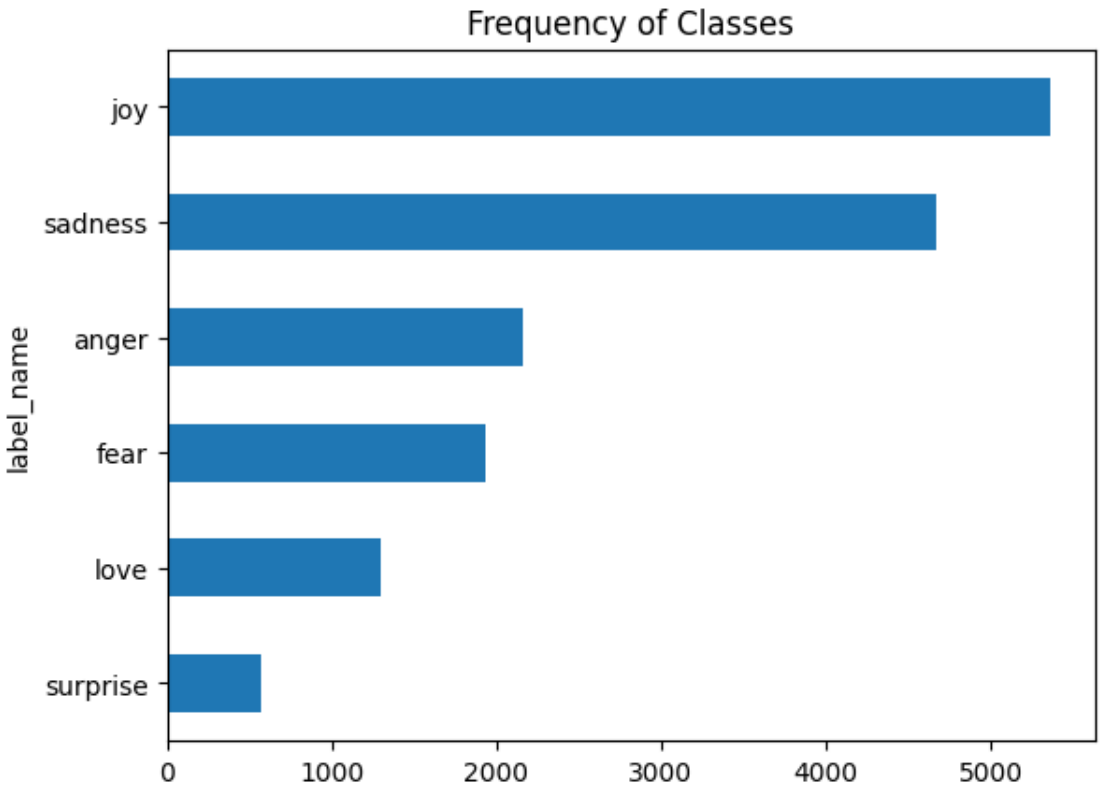
[9]:

	text	label	label_name
0	i didnt feel humiliated	0	sadness
1	i can go from feeling so hopeless to so damned…	0	sadness
2	im grabbing a minute to post i feel greedy wrong	3	anger
3	i am ever feeling nostalgic about the fireplac…	2	love
4	i am feeling grouchy	3	anger

[10]:
```
# Data Analysis
```

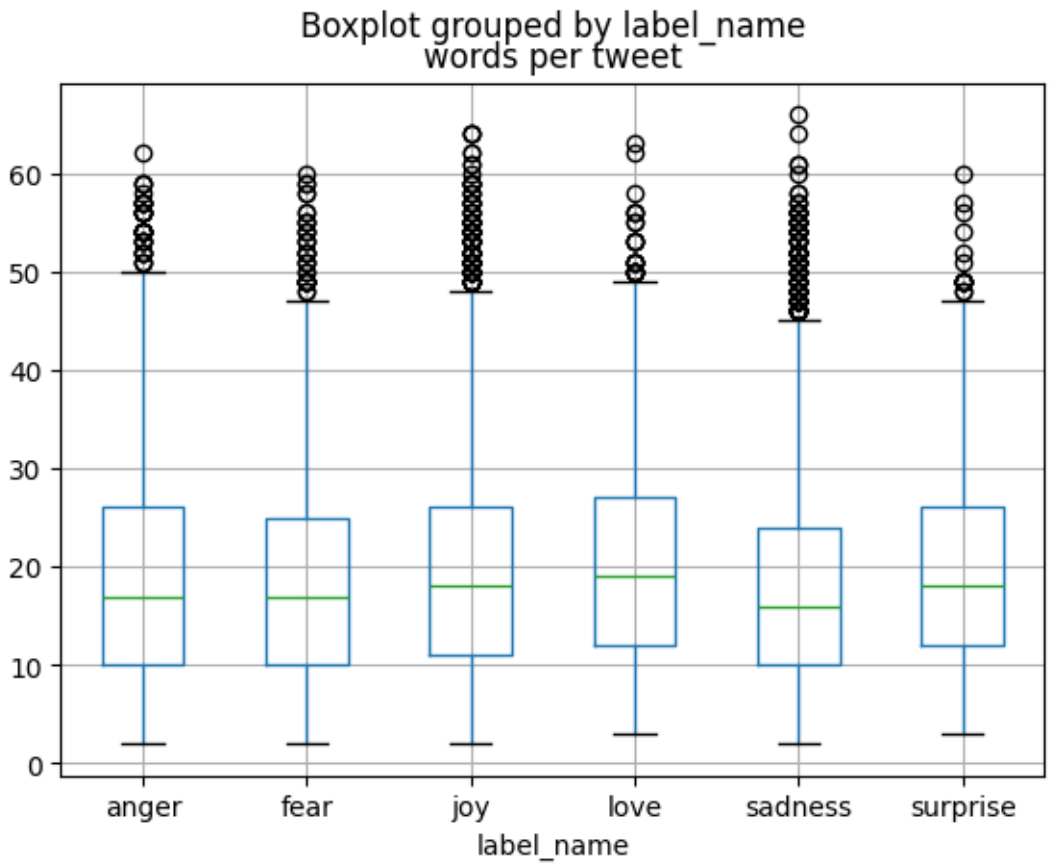
[11]:
```
import matplotlib.pyplot as plt
```

[12]:
```
df['label_name'].value_counts(ascending=True).plot.barh()
plt.title('Frequency of Classes')
plt.show()
```



[13]:
```
df['words per tweet'] = df['text'].str.split().apply(len)
df.boxplot('words per tweet', by='label_name')
```

[13]:
```
<Axes: title={'center': 'words per tweet'}, xlabel='label_name'>
```



[14]:
```
from transformers import AutoTokenizer
model_ckpt = 'distilbert-base-uncased'
tokenizer = AutoTokenizer.from_pretrained(model_ckpt)
```

```
The cache for model files in Transformers v4.22.0 has been updated. Migrating
your old cache. This is a one-time only operation. You can interrupt this and
resume the migration later on by calling `transformers.utils.move_cache()`.

0it [00:00, ?it/s]

tokenizer_config.json:   0%|          | 0.00/48.0 [00:00<?, ?B/s]

config.json:   0%|          | 0.00/483 [00:00<?, ?B/s]

vocab.txt:   0%|          | 0.00/232k [00:00<?, ?B/s]

tokenizer.json:   0%|          | 0.00/466k [00:00<?, ?B/s]
```

0.1 Tokenization

[15]:
```
emotions.reset_format()
```

[16]:
```
def tokenize(batch):
  temp = tokenizer(batch['text'], padding=True, truncation=True)
  return temp
```

[17]:
```
emotions_encoded = emotions.map(tokenize, batched=True, batch_size=None)
```

```
Map:   0%|          | 0/16000 [00:00<?, ? examples/s]
Map:   0%|          | 0/2000 [00:00<?, ? examples/s]
Map:   0%|          | 0/2000 [00:00<?, ? examples/s]
```

[18]:
```
emotions_encoded
```

[18]:
```
DatasetDict({
    train: Dataset({
        features: ['text', 'label', 'input_ids', 'attention_mask'],
        num_rows: 16000
    })
    validation: Dataset({
        features: ['text', 'label', 'input_ids', 'attention_mask'],
        num_rows: 2000
    })
    test: Dataset({
        features: ['text', 'label', 'input_ids', 'attention_mask'],
        num_rows: 2000
    })
})
```

[31]:
```
train_df = emotions_encoded['train'].to_tf_dataset(
    columns=['input_ids', 'attention_mask'],
    shuffle=True,
    batch_size=32,
    label_cols=['label']
)
val_df = emotions_encoded['validation'].to_tf_dataset(
    columns=['input_ids', 'attention_mask'],
    shuffle=False,
    batch_size=32,
    label_cols=['label']
)
test_df = emotions_encoded['test'].to_tf_dataset(
    columns=['input_ids', 'attention_mask'],
    shuffle=False,
    batch_size=32,
    label_cols=['label']
)
```

```
/usr/local/lib/python3.11/dist-packages/datasets/arrow_dataset.py:405:
```

```
FutureWarning: The output of `to_tf_dataset` will change when a passing single
element list for `labels` or `columns` in the next datasets version. To return a
tuple structure rather than dict, pass a single string.
Old behaviour: columns=['a'], labels=['labels'] -> (tf.Tensor, tf.Tensor)
             : columns='a', labels='labels' -> (tf.Tensor, tf.Tensor)
New behaviour: columns=['a'],labels=['labels'] -> ({'a': tf.Tensor}, {'labels':
tf.Tensor})
             : columns='a', labels='labels' -> (tf.Tensor, tf.Tensor)
  warnings.warn(
```

0.2 Model

[32]:
```
import tensorflow as tf
from transformers import TFAutoModelForSequenceClassification
num_labels = 6

model = TFAutoModelForSequenceClassification.from_pretrained(model_ckpt,␣
↪num_labels=num_labels)
```

```
Some weights of the PyTorch model were not used when initializing the TF 2.0
model TFDistilBertForSequenceClassification: ['vocab_transform.bias',
'vocab_transform.weight', 'vocab_layer_norm.weight', 'vocab_layer_norm.bias',
'vocab_projector.bias']
- This IS expected if you are initializing TFDistilBertForSequenceClassification
from a PyTorch model trained on another task or with another architecture (e.g.
initializing a TFBertForSequenceClassification model from a BertForPreTraining
model).
- This IS NOT expected if you are initializing
TFDistilBertForSequenceClassification from a PyTorch model that you expect to be
exactly identical (e.g. initializing a TFBertForSequenceClassification model
from a BertForSequenceClassification model).
Some weights or buffers of the TF 2.0 model
TFDistilBertForSequenceClassification were not initialized from the PyTorch
model and are newly initialized: ['pre_classifier.weight',
'pre_classifier.bias', 'classifier.weight', 'classifier.bias']
You should probably TRAIN this model on a down-stream task to be able to use it
for predictions and inference.
```

[33]:
```
from transformers import AdamWeightDecay
```

[34]:
```
model.compile(optimizer=AdamWeightDecay(learning_rate=5e-5, weight_decay_rate=0.
↪01),
              loss=tf.keras.losses.
↪SparseCategoricalCrossentropy(from_logits=True),
              metrics=['accuracy'])
```

[35]:
```
model.fit(train_df, validation_data=val_df, epochs=3)
```

```
Epoch 1/3
500/500 [==============================] - 175s 323ms/step - loss: 0.4553 -
accuracy: 0.8406 - val_loss: 0.1886 - val_accuracy: 0.9260
Epoch 2/3
500/500 [==============================] - 154s 307ms/step - loss: 0.1365 -
accuracy: 0.9419 - val_loss: 0.1314 - val_accuracy: 0.9375
Epoch 3/3
500/500 [==============================] - 158s 317ms/step - loss: 0.1047 -
accuracy: 0.9503 - val_loss: 0.1328 - val_accuracy: 0.9415
```

[35]:
```
<tf_keras.src.callbacks.History at 0x7a5cd045b650>
```

[37]:
```
model.evaluate(test_df)
```

```
63/63 [==============================] - 8s 85ms/step - loss: 0.1447 - accuracy:
0.9265
```

[37]:
```
[0.14469902217388153, 0.9265000224113464]
```

[42]:
```
import numpy as np
```

[45]:
```
text = 'i am sad'
inputs = tokenizer(text, return_tensors='tf', padding=True, truncation=True)
logits = model(**inputs)

logits = logits.logits
pred = np.argmax(logits, axis=1).item()
pred, classes[pred]
```

[45]:
```
(0, 'sadness')
```

[ ]:
```

```

This notebook was converted with convert.ploomber.io