```
import pandas as pd

df = pd.read_csv("multilang_sarcasm_dataset.csv")
```

```
df
```

	article_url	article_title	is_sarcastic	lang	title_length
0	https://www.huffingtonpost.com/entry/versace-b...	former versace store clerk sues over secret 'b...	0	en	78
1	https://www.huffingtonpost.com/entry/roseanne-...	the 'roseanne' revival catches up to our thorn...	0	en	84
2	https://local.theonion.com/mom-starting-to-fea...	mom starting to fear son's web series closest ...	1	en	79
3	https://politics.theonion.com/boehner-just-wan...	boehner just wants wife to listen, not come up...	1	en	84
4	https://www.huffingtonpost.com/entry/jk-rowlin...	j.k. rowling wishes snape happy birthday in th...	0	en	64
...	...	...	...	...	...
67474	https://speld.nl/2022/08/27/dit-is-de-enige-we...	dit is de enige wettelijk toegestane manier om...	1	nl	73
67475	https://speld.nl/2022/03/27/nieuwe-fitnesspas-...	nieuwe fitnesspas 200 euro zodat je iedere dag...	1	nl	81
67476	https://speld.nl/2022/09/23/wilco-stond-5-minu...	wilco stond 5 minuten in de rij voor de kassa ...	1	nl	85
67477	https://speld.nl/2022/09/17/nemen-de-britten-w...	nemen de britten wel genoeg tijd om te rouwen?	1	nl	46
67478	https://speld.nl/2022/02/12/einde-pandemie-in-...	einde pandemie in zicht, vrouwen alsnog geadvi...	1	nl	67

67479 rows × 5 columns

SetFit MODEL - 1

Model: MPNET BASE v2

N = 64

```
from setfit import SetFitModel, SetFitTrainer
from sentence_transformers.losses import CosineSimilarityLoss
from datasets import Dataset
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, f1_score
import pandas as pd
import os

# === CONFIG ===
DATA_PATH = "multilang_sarcasm_dataset.csv"
MODEL_PATH = "model/setfit_multilang_sarcasm_en_32"
N_SHOT = 64
MAX_TEST_SAMPLES = 1000

# === LOAD & PREPROCESS ===
df = pd.read_csv(DATA_PATH)

# Filter to English headlines
df = df[df["lang"] == "en"]

# Rename columns to match SETFIT input format
df = df[["article_title", "is_sarcastic"]].rename(columns={"article_title": "text", "is_sarcastic": "label"})

# Drop any potential NaNs
df = df.dropna(subset=["text", "label"])

# === TRAIN/TEST SPLIT ===
train_df, test_df = train_test_split(df, test_size=0.2, stratify=df["label"], random_state=42)

# Few-shot sampling
def sample_few_shot(df, n=64):
    return df.groupby("label").apply(lambda x: x.sample(n=min(n, len(x)), random_state=42)).reset_index(drop=True)

fewshot_train_df = sample_few_shot(train_df, N_SHOT)
test_subset_df = test_df.sample(n=min(len(test_df), MAX_TEST_SAMPLES), random_state=42)

# Convert to HuggingFace datasets
train_dataset = Dataset.from_pandas(fewshot_train_df)
test_dataset = Dataset.from_pandas(test_subset_df)
```

```
# === LOAD BASE MODEL ===
model = SetFitModel.from_pretrained("sentence-transformers/all-mpnet-base-v2")

# === TRAIN SETUP ===
trainer = SetFitTrainer(
    model=model,
    train_dataset=train_dataset,
    eval_dataset=test_dataset,
    loss_class=CosineSimilarityLoss,
    batch_size=16,
    num_iterations=50,
    num_epochs=1,
    column_mapping={"text": "text", "label": "label"},
)

trainer.train()

# Save model
model.save_pretrained(MODEL_PATH)

# Evaluate
y_true = test_dataset["label"]
y_pred = model.predict(test_dataset["text"])
acc = accuracy_score(y_true, y_pred)
f1 = f1_score(y_true, y_pred)

print(f"Accuracy: {acc:.4f} | F1 Score: {f1:.4f}")
```

```
/var/folders/lv/xd91rcv91cq23cjjl0_c93nh0000gn/T/ipykernel_50657/1317443071.py:32: DeprecationWarning: DataFrameGroupBy.appl
  return df.groupby("label").apply(lambda x: x.sample(n=min(n, len(x)), random_state=42)).reset_index(drop=True)
`SentenceTransformer._target_device` has been deprecated, please use `SentenceTransformer.device` instead.
model_head.pkl not found on HuggingFace Hub, initialising classification head with random weights. You should TRAIN this mod
Applying column mapping to training dataset
Generating Training Pairs: 100%|██████████| 50/50 [00:00<00:00, 1044.50it/s]
***** Running training *****
  Num examples = 12800
  Num epochs = 1
  Total optimization steps = 800
  Total train batch size = 16
```

[800/800 02:57, Epoch 1/1]

Step	Training Loss
500	0.060100

```
Accuracy: 0.8220 | F1 Score: 0.8172
```

MODEL : MPNET BASE v2

N = 32

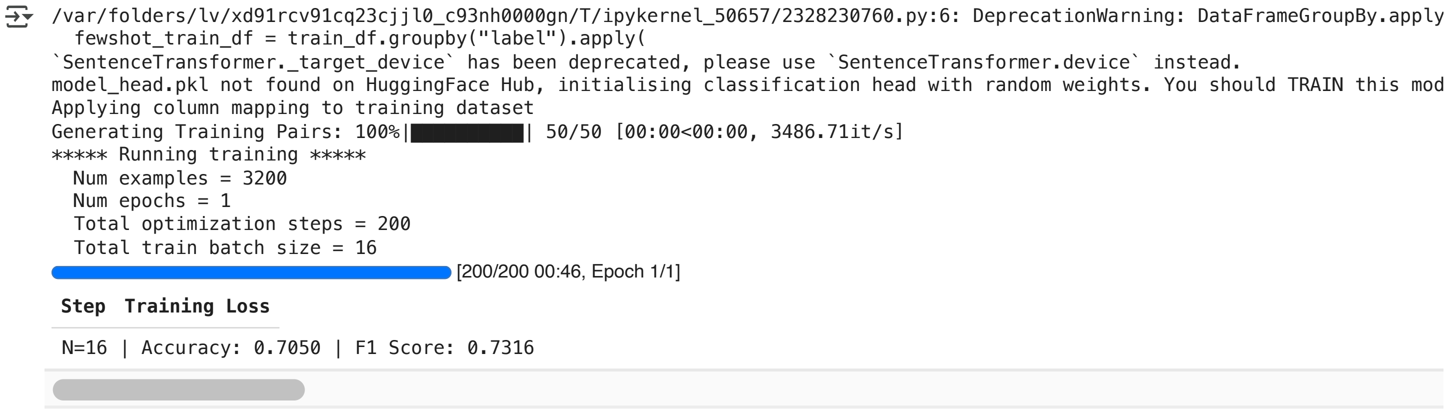
```
# === FEW-SHOT SETUP FOR N_SHOT = 32 ===
N_SHOT = 32
DATA_PATH = "multilang_sarcasm_dataset.csv"
MODEL_PATH = "model/setfit_multilang_sarcasm_en_32"

fewshot_train_df = train_df.groupby("label").apply(
    lambda x: x.sample(n=min(N_SHOT, len(x)), random_state=42)
).reset_index(drop=True)

train_dataset = Dataset.from_pandas(fewshot_train_df)
test_dataset = Dataset.from_pandas(test_subset_df)

model = SetFitModel.from_pretrained("sentence-transformers/all-mpnet-base-v2")
trainer = SetFitTrainer(
    model=model,
    train_dataset=train_dataset,
    eval_dataset=test_dataset,
    loss_class=CosineSimilarityLoss,
    batch_size=16,
    num_iterations=50,
    num_epochs=1,
    column_mapping={"text": "text", "label": "label"},
)
trainer.train()

model.save_pretrained(MODEL_PATH)

# Evaluate
y_pred = model.predict(test_dataset["text"])
acc = accuracy_score(test_dataset["label"], y_pred)
f1 = f1_score(test_dataset["label"], y_pred)
print(f" N=32 | Accuracy: {acc:.4f} | F1 Score: {f1:.4f}")
```

```
/var/folders/lv/xd91rcv91cq23cjjl0_c93nh0000gn/T/ipykernel_50657/3250611203.py:6: DeprecationWarning: DataFrameGroupBy.apply
  fewshot_train_df = train_df.groupby("label").apply(
`SentenceTransformer._target_device` has been deprecated, please use `SentenceTransformer.device` instead.
model_head.pkl not found on HuggingFace Hub, initialising classification head with random weights. You should TRAIN this mod
Applying column mapping to training dataset
Generating Training Pairs: 100%|██████████| 50/50 [00:00<00:00, 1655.25it/s]
***** Running training *****
  Num examples = 6400
  Num epochs = 1
  Total optimization steps = 400
  Total train batch size = 16
[400/400 01:30, Epoch 1/1]
 Step  Training Loss
 N=32 | Accuracy: 0.7570 | F1 Score: 0.7538
```

MODEL : MPNET BASE v2

N = 16

```
# === FEW-SHOT SETUP FOR N_SHOT = 16 ===
N_SHOT = 16
MODEL_PATH = f"model/setfit_multilang_sarcasm_en_N{N_SHOT}"
DATA_PATH = "multilang_sarcasm_dataset.csv"

fewshot_train_df = train_df.groupby("label").apply(
    lambda x: x.sample(n=min(N_SHOT, len(x)), random_state=42)
).reset_index(drop=True)

train_dataset = Dataset.from_pandas(fewshot_train_df)
test_dataset = Dataset.from_pandas(test_subset_df)

model = SetFitModel.from_pretrained("sentence-transformers/all-mpnet-base-v2")
trainer = SetFitTrainer(
    model=model,
    train_dataset=train_dataset,
    eval_dataset=test_dataset,
    loss_class=CosineSimilarityLoss,
    batch_size=16,
    num_iterations=50,
    num_epochs=1,
    column_mapping={"text": "text", "label": "label"},
)
trainer.train()

model.save_pretrained(MODEL_PATH)

# Evaluate
y_pred = model.predict(test_dataset["text"])
acc = accuracy_score(test_dataset["label"], y_pred)
f1 = f1_score(test_dataset["label"], y_pred)
print(f" N=16 | Accuracy: {acc:.4f} | F1 Score: {f1:.4f}")
```



```
/var/folders/lv/xd91rcv91cq23cjjl0_c93nh0000gn/T/ipykernel_50657/2328230760.py:6: DeprecationWarning: DataFrameGroupBy.apply
  fewshot_train_df = train_df.groupby("label").apply(
`SentenceTransformer._target_device` has been deprecated, please use `SentenceTransformer.device` instead.
model_head.pkl not found on HuggingFace Hub, initialising classification head with random weights. You should TRAIN this mod
Applying column mapping to training dataset
Generating Training Pairs: 100%|██████████| 50/50 [00:00<00:00, 3486.71it/s]
***** Running training *****
  Num examples = 3200
  Num epochs = 1
  Total optimization steps = 200
  Total train batch size = 16
[200/200 00:46, Epoch 1/1]
 Step  Training Loss
 N=16 | Accuracy: 0.7050 | F1 Score: 0.7316
```