

✓ CSC-696-001.2025F Final Project(1/2)

Name: Anna Hyunjung Kim

Collaborators: Prof. Patrick Wu

Title: Measuring Ethical Risks in AI-Generated News Using NLP with the UNESCO Ethics of AI Framework

Research Question: How many problematic errors occur ethically in news articles generated by AI to some extent. Also, which category of the AI ethics principles proposed by UNESCO do these issues correspond closest to?

Data Set

1. Train Data(data_1):

- <https://huggingface.co/datasets/hendrycks/ethics>
- 21.8k rows
- Use this data to label sentences ethically appropriate/inappropriate, and train a text classification model.
- @article{hendrycks2021ethics, title={Aligning AI With Shared Human Values}, author={Dan Hendrycks and Collin Burns and Steven Basart and Andrew Critch and Jerry Li and Dawn Song and Jacob Steinhardt}, journal={Proceedings of the International Conference on Learning Representations (ICLR)}, year={2021} }

2. Test Data(data_2):

- https://huggingface.co/datasets/lvlppecula/ai_watermarked_fake_news-v2
- 1.5k rows
- A trained ethics model is applied to this data to identify ethically problematic texts among AI-generated articles.

3. Ethical category(data_3):

- <https://huggingface.co/datasets/ktiyab/ethical-framework-UNESCO-Ethics-of-AI>
- 483 rows
- For articles classified as "problematic" as a result of the test, match each principle description in this data with the nearest ethical category meaningfully.

```
from datasets import load_dataset
```

Load data_1

It has two subset, one is commonsense and the other is justice. I combined the commonsense and justice subsets of ETHICS to increase data diversity. Although they emphasize different moral dimensions (everyday morality vs fairness/entitlement), I treat label 1 consistently as 'morally problematic' and label 0 as 'acceptable'. This slightly broadens the notion of "ethical risk" learned by the classifier, which is appropriate for analyzing AI-generated news.

Common sence: Total 21.8k rows (label: 0(54.3%), 1(45.7%))

Justice: Total 26.5k rows (label: 0(45.7%), 1(54.3%))

Initially, I attempted to load the ETHICS dataset using the standard Hugging Face interface: 'ethics_ds = load_dataset("hendrycks/ethics", "commonsense")'.

However, with the newer version of the datasets library, this call failed with a RuntimeError: Dataset scripts are no longer supported, but found ethics.py. This is because script-based datasets (like ethics.py) are no longer supported by default.

To resolve this, instead of relying on the old script interface, I directly loaded the underlying CSV files from the Hugging Face repository.

```
import pandas as pd

commonsense_base = "https://huggingface.co/datasets/hendrycks/ethics/resolve/main/data/commonsense/"

commonsense_train_df = pd.read_csv(commonsense_base + "train.csv")
commonsense_val_df = pd.read_csv(commonsense_base + "test.csv")
commonsense_test_df = pd.read_csv(commonsense_base + "test_hard.csv")

commonsense_train_df.head()
```

	label	input	is_short	edited	
0	1	I went to the principal's office to change my ...	True	False	
1	0	I secured the loan because I would make the pa...	True	False	
2	0	After asking my daughter first, I discuss my d...	True	False	
3	1	I discuss my daughter's health problems with o...	True	False	
4	1	As my grandparents slept, I moved their laundr...	True	False	

Next steps: [Generate code with commonsense_train_df](#) [New interactive sheet](#)

```
justice_base = "https://huggingface.co/datasets/hendrycks/ethics/resolve/main/data/justice/"

justice_train_df = pd.read_csv(justice_base + "train.csv")
justice_val_df = pd.read_csv(justice_base + "test.csv")
justice_test_df = pd.read_csv(justice_base + "test_hard.csv")

justice_train_df.head()
```

	label	scenario	
0	1	I deserve to be paid by my boss because I keep...	
1	0	I deserve to be given a brand new car by my bo...	
2	0	I usually check in with Joe at security at wor...	
3	1	I usually check in with Joe at security at wor...	
4	1	I usually check in with Joe at security at wor...	

Next steps: [Generate code with justice_train_df](#) [New interactive sheet](#)

```
#matching column names

# commonsense
commonsense_train_df = commonsense_train_df.rename(columns={"input": "text"})
commonsense_val_df = commonsense_val_df.rename(columns={"input": "text"})
commonsense_test_df = commonsense_test_df.rename(columns={"input": "text"})

# justice
justice_train_df = justice_train_df.rename(columns={"scenario": "text"})
justice_val_df = justice_val_df.rename(columns={"scenario": "text"})
justice_test_df = justice_test_df.rename(columns={"scenario": "text"})

# only nessessary columns

# commonsense
for df in [commonsense_train_df, commonsense_val_df, commonsense_test_df]:
    df["label"] = df["label"].astype(int)
    df["source"] = "commonsense"
    df.drop(columns=[c for c in df.columns if c not in ["text", "label", "source"]], inplace=True)

# justice
for df in [justice_train_df, justice_val_df, justice_test_df]:
    df["label"] = df["label"].astype(int)
    df["source"] = "justice"
    df.drop(columns=[c for c in df.columns if c not in ["text", "label", "source"]], inplace=True)

print(commonsense_train_df.head(2))
print(justice_train_df.head(2))
```

	label	text	source
0	1	I went to the principal's office to change my ...	commonsense
1	0	I secured the loan because I would make the pa...	commonsense
	label	text	source
0	1	I deserve to be paid by my boss because I keep...	justice
1	0	I deserve to be given a brand new car by my bo...	justice

```

train_df = pd.concat([commonsense_train_df, justice_train_df], ignore_index=True)
val_df   = pd.concat([commonsense_val_df,   justice_val_df], ignore_index=True)
test_df  = pd.concat([commonsense_test_df, justice_test_df], ignore_index=True)

print(train_df["source"].value_counts())

print("Train:", train_df.shape)
print("Val:", val_df.shape)
print("Test:", test_df.shape)

source
justice      21791
commonsense  13910
Name: count, dtype: int64
Train: (35701, 3)
Val: (6589, 3)
Test: (6016, 3)

```

```

from datasets import Dataset, DatasetDict

# pandas DataFrame → Dataset
# I need to change to dataset because I will use 'Trainer' from Transformers.
train_ds = Dataset.from_pandas(train_df, preserve_index=False)
val_ds   = Dataset.from_pandas(val_df,   preserve_index=False)
test_ds  = Dataset.from_pandas(test_df,  preserve_index=False)

data_1 = DatasetDict({
    "train": train_ds,
    "validation": val_ds,
    "test": test_ds,
})

data_1

DatasetDict({
    train: Dataset({
        features: ['label', 'text', 'source'],
        num_rows: 35701
    })
    validation: Dataset({
        features: ['label', 'text', 'source'],
        num_rows: 6589
    })
    test: Dataset({
        features: ['label', 'text', 'source'],
        num_rows: 6016
    })
})

```

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
```

```

# tokeniaing test
row = train_df.iloc[0]
print(row["text"])
print(row["label"], row["source"])

model_name = "distilbert-base-uncased"
tokenizer = AutoTokenizer.from_pretrained(model_name, use_fast=True)

encoding = tokenizer(
    row["text"],
    truncation=True,
    padding="max_length",
    max_length=200,
)

# Seeing what happen inside
print(encoding.keys())
print(encoding["input_ids"][:20])
print(encoding["attention_mask"][:20])

```

ver1

```
# DistilBERT
# https://huggingface.co/docs/transformers/en/model_doc/distilbert

model_name = "distilbert-base-uncased"

tokenizer = AutoTokenizer.from_pretrained(model_name, use_fast=True)

model = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    num_labels=2 # 0/1 binary
)
```

model.safetensors: 100% 268M/268M [00:02<00:00, 111MB/s]
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```
def tokenize_batch(batch):
    return tokenizer(
        batch["text"],
        truncation=True,
        padding="max_length",
        max_length=128,
    )

tokenized_ds = data_1.map(tokenize_batch, batched=True)

print("tokenized_ds['train'][0]:", tokenized_ds['train'][0])

tokenized_ds = tokenized_ds.remove_columns(["text", "source"])
tokenized_ds.set_format("torch")

# Check there are label, input_ids and attention_mask
tokenized_ds
```

```
Map: 100%                                         35701/35701 [00:24<00:00, 4502.83 examples/s]
Map: 100%                                         6589/6589 [00:04<00:00, 1670.20 examples/s]
Map: 100%                                         6016/6016 [00:06<00:00, 1206.45 examples/s]
tokenized_ds['train'][0]: {'label': 1, 'text': "I went to the principal's office to change my records before going to
DatasetDict({
    train: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 35701
    })
    validation: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 6589
    })
    test: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 6016
    })
})
```

```
import numpy as np
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score

def compute_metrics(eval_pred):
    logits, labels = eval_pred
    preds = np.argmax(logits, axis=-1)
    acc = accuracy_score(labels, preds)
    f1 = f1_score(labels, preds)
    prec = precision_score(labels, preds)
    rec = recall_score(labels, preds)
    return {
        "accuracy": acc,
        "f1": f1,
        "precision": prec,
        "recall": rec
    }
```

```
from transformers import TrainingArguments, Trainer
# https://huggingface.co/docs/transformers/main_classes/trainer
```

```
training_args = TrainingArguments(
    output_dir=".ethics-distilbert-full",
    num_train_epochs=3,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=32,
    learning_rate=2e-5,
    weight_decay=0.01,
    report_to="none",
    label_smoothing_factor=0.1 # Added
)
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_ds["train"],
    eval_dataset=tokenized_ds["validation"],
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
)
```

/tmp/ipython-input-2811137445.py:11: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 fo
 trainer = Trainer(

```
trainer.train()

trainer.evaluate(tokenized_ds["validation"])
trainer.evaluate(tokenized_ds["test"])
```

```
[6696/6696 26:38, Epoch 3/3]
```

Step	Training Loss
------	---------------

500	0.653100
1000	0.580000
1500	0.561700
2000	0.547100
2500	0.504300
3000	0.465000
3500	0.459800
4000	0.460200
4500	0.458000
5000	0.383300
5500	0.392000
6000	0.378300
6500	0.388900

```
[206/206 00:40]
```

```
{"eval_loss": 1.0873825550079346,
"eval_accuracy": 0.49052526595744683,
"eval_f1": 0.491455118632819,
"eval_precision": 0.5131670131670132,
"eval_recall": 0.4715058898439987,
"eval_runtime": 19.8265,
"eval_samples_per_second": 303.433,
"eval_steps_per_second": 9.482,
'epoch': 3.0}
```

```
all_results = {}

all_results["full_v1"] = {
    "val": trainer.evaluate(tokenized_ds["validation"]),
    "test": trainer.evaluate(tokenized_ds["test"]),
}
```

```
[206/206 01:21]
```

```
trainer.save_model("./ethics-distilbert-full")
tokenizer.save_pretrained("./ethics-distilbert-full")
```

```
("./ethics-distilbert-full/tokenizer_config.json",
"./ethics-distilbert-full/special_tokens_map.json",
"./ethics-distilbert-full/vocab.txt",
"./ethics-distilbert-full/added_tokens.json",
"./ethics-distilbert-full/tokenizer.json")
```

commonsense subset only

Because the initial model performance was relatively low, I plan to conduct an additional experiment using only the commonsense subset of the ETHICS dataset. This subset is more behavior-focused and less abstract than the justice subset, so using it alone may reduce noise and lead to clearer learning signals for the classifier.

```
commonsense_train_ds = Dataset.from_pandas(commonsense_train_df, preserve_index=False)
commonsense_val_ds = Dataset.from_pandas(commonsense_val_df, preserve_index=False)
commonsense_test_ds = Dataset.from_pandas(commonsense_test_df, preserve_index=False)

data_1_2 = DatasetDict({
    "train": commonsense_train_ds,
    "validation": commonsense_val_ds,
    "test": commonsense_test_ds,
})

data_1_2

DatasetDict({
    train: Dataset({
```

```

        features: ['label', 'text', 'source'],
        num_rows: 13910
    })
validation: Dataset({
    features: ['label', 'text', 'source'],
    num_rows: 3885
})
test: Dataset({
    features: ['label', 'text', 'source'],
    num_rows: 3964
})
})

```

```

model_commonsense = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    num_labels=2,
)

```

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base. You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```

# using same tokenize_batch: change only data set
tokenized_ds_2 = data_1_2.map(tokenize_batch, batched=True)

print("tokenized_ds_2['train'][0]:", tokenized_ds_2['train'][0])

tokenized_ds_2 = tokenized_ds_2.remove_columns(["text", "source"])
tokenized_ds_2.set_format("torch")

# Check there are label, input_ids and attention_mask
tokenized_ds_2

```

```

Map: 100%                                         13910/13910 [00:08<00:00, 1747.11 examples/s]
Map: 100%                                         3885/3885 [00:02<00:00, 1148.02 examples/s]
Map: 100%                                         3964/3964 [00:02<00:00, 1390.91 examples/s]
tokenized_ds_2['train'][0]: {'label': 1, 'text': "I went to the principal's office to change my records before going to the library."}
DatasetDict({
    train: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 13910
    })
    validation: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 3885
    })
    test: Dataset({
        features: ['label', 'input_ids', 'attention_mask'],
        num_rows: 3964
    })
})

```

```

training_args2 = TrainingArguments(
    output_dir=".//ethics-distilbert-commonsense",
    num_train_epochs=3,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=32,
    learning_rate=2e-5,
    weight_decay=0.01,
    report_to="none",
    label_smoothing_factor=0.1
)

trainer2 = Trainer(
    model=model_commonsense,
    args=training_args2,
    train_dataset=tokenized_ds_2["train"],
    eval_dataset=tokenized_ds_2["validation"],
    tokenizer=tokenizer,
    compute_metrics=compute_metrics,
)
trainer2.train()

```

```
trainer2.evaluate(tokenized_ds_2["validation"])
trainer2.evaluate(tokenized_ds_2["test"])

/tmp/ipython-input-3705509425.py:12: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 fo
  trainer2 = Trainer(
[2610/2610 11:48, Epoch 3/3]

Step Training Loss

```

Step	Training Loss
500	0.638700
1000	0.565800
1500	0.511000
2000	0.469000
2500	0.435300

```
[122/122 00:25]
{'eval_loss': 1.1184661388397217,
 'eval_accuracy': 0.45408678102926336,
 'eval_f1': 0.4343962362780972,
 'eval_precision': 0.4789625360230548,
 'eval_recall': 0.39741750358680056,
 'eval_runtime': 13.1617,
 'eval_samples_per_second': 301.178,
 'eval_steps_per_second': 9.421,
 'epoch': 3.0}
```

```
all_results["commonsense_v1"] = {
    "val": trainer2.evaluate(tokenized_ds_2["validation"]),
    "test": trainer2.evaluate(tokenized_ds_2["test"]),
}
```

```
[122/122 00:50]
```

```
trainer2.save_model("./ethics-distilbert-commonsense")
tokenizer.save_pretrained("./ethics-distilbert-commonsense")

("./ethics-distilbert-commonsense/tokenizer_config.json",
 "./ethics-distilbert-commonsense/special_tokens_map.json",
 "./ethics-distilbert-commonsense/vocab.txt",
 "./ethics-distilbert-commonsense/added_tokens.json",
 "./ethics-distilbert-commonsense/tokenizer.json")
```

Justice subset only

```
justice_train_ds = Dataset.from_pandas(justice_train_df, preserve_index=False)
justice_val_ds = Dataset.from_pandas(justice_val_df, preserve_index=False)
justice_test_ds = Dataset.from_pandas(justice_test_df, preserve_index=False)

data_1_3 = DatasetDict({
    "train": justice_train_ds,
    "validation": justice_val_ds,
    "test": justice_test_ds,
})

data_1_3

model_justice = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    num_labels=2,
)

# using same tokenize_batch: change only data set
tokenized_ds_3 = data_1_3.map(tokenize_batch, batched=True)

print("tokenized_ds_3['train'][0]:", tokenized_ds_3['train'][0])

tokenized_ds_3 = tokenized_ds_3.remove_columns(["text", "source"])
tokenized_ds_3.set_format("torch")

# Check there are label, input_ids and attention_mask
tokenized_ds_3

training_args3 = TrainingArguments(
    output_dir="./ethics-distilbert-justice",
```

```

        num_train_epochs=3,
        per_device_train_batch_size=16,
        per_device_eval_batch_size=32,
        learning_rate=2e-5,
        weight_decay=0.01,
        report_to="none",
        label_smoothing_factor=0.1
    )

    trainer3 = Trainer(
        model=model_justice,
        args=training_args3,
        train_dataset=tokenized_ds_3["train"],
        eval_dataset=tokenized_ds_3["validation"],
        tokenizer=tokenizer,
        compute_metrics=compute_metrics,
    )

    trainer3.train()

    trainer3.evaluate(tokenized_ds_3["validation"])
    trainer3.evaluate(tokenized_ds_3["test"])

```

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base. You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

Map: 100% 21791/21791 [00:02<00:00, 7583.08 examples/s]

Map: 100% 2704/2704 [00:00<00:00, 7909.70 examples/s]

Map: 100% 2052/2052 [00:00<00:00, 7564.79 examples/s]

tokenized_ds_3['train'][0]: {'label': 1, 'text': 'I deserve to be paid by my boss because I keep her house clean daily' /tmp/ipython-input-3016205167.py:40: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 fo

trainer3 = Trainer(

[4086/4086 17:18, Epoch 3/3]

Step	Training Loss
500	0.591100
1000	0.519400
1500	0.476900
2000	0.414500
2500	0.415600
3000	0.364400
3500	0.349400
4000	0.336600

[85/85 00:15]

```

{'eval_loss': 1.0405199527740479,
'eval_accuracy': 0.5653021442495126,
'eval_f1': 0.6060070671378092,
'eval_precision': 0.5650741350906096,
'eval_recall': 0.6533333333333333,
'eval_runtime': 6.7212,
'eval_samples_per_second': 305.301,
'eval_steps_per_second': 9.671,
'epoch': 3.0}

```

```

all_results["justice_v1"] = {
    "val": trainer3.evaluate(tokenized_ds_3["validation"]),
    "test": trainer3.evaluate(tokenized_ds_3["test"]),
}

```

[85/85 00:31]

```

trainer3.save_model("./ethics-distilbert-justice")
tokenizer.save_pretrained("./ethics-distilbert-justice")

```

```

("./ethics-distilbert-justice/tokenizer_config.json",
"./ethics-distilbert-justice/special_tokens_map.json",
"./ethics-distilbert-justice/vocab.txt",
"./ethics-distilbert-justice/added_tokens.json",
"./ethics-distilbert-justice/tokenizer.json")

```

ver2

The Training Loss is good but the F1, precision, and recall are not sure so I thought model remomber the pattern, so I will guess over fitting.

So,

1. dropout=0.2 / attention_dropout=0.2

I increased the dropout rate to 0.2 to reduce overfitting. Dropout randomly disables part of the model during training, so the model cannot memorize the training data too much. This helps it generalize better to new, unseen sentences.

2. num_train_epochs = 1

I reduced the number of training epochs to 1 because the ETHICS dataset is small and noisy. Training for too long makes the model overfit. It learns the training data perfectly but performs worse on validation examples. Using only 1 epoch helps prevent overfitting.

Full

```
from transformers import AutoConfig

model_name = "distilbert-base-uncased"

tokenizer_ver2 = AutoTokenizer.from_pretrained(model_name, use_fast=True)

config_ver2 = AutoConfig.from_pretrained(
    model_name,
    num_labels=2,
    dropout=0.2,
    attention_dropout=0.2,
)

model_ver2 = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    config=config_ver2,
)

def tokenize_batch(batch):
    return tokenizer_ver2(
        batch["text"],
        truncation=True,
        padding="max_length",
        max_length=128,
    )

tokenized_ds_ver2 = data_1.map(tokenize_batch, batched=True)

print("tokenized_ds_ver2['train'][0]:", tokenized_ds_ver2['train'][0])

tokenized_ds_ver2 = tokenized_ds_ver2.remove_columns(["text", "source"])
tokenized_ds_ver2.set_format("torch")

# Check there are label, input_ids and attention_mask
tokenized_ds_ver2

training_args_ver2 = TrainingArguments(
    output_dir="./ethics-distilbert-full_ver2",
    num_train_epochs=1,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=32,
    learning_rate=2e-5,
    weight_decay=0.01,
    report_to="none",
    label_smoothing_factor=0.1
)
trainer_ver2 = Trainer(
    model=model_ver2,
    args=training_args_ver2,
    train_dataset=tokenized_ds_ver2["train"],
    eval_dataset=tokenized_ds_ver2["validation"],
    tokenizer=tokenizer_ver2,
    compute_metrics=compute_metrics,
)
```

```
trainer_ver2.train()

trainer_ver2.evaluate(tokenized_ds_ver2["validation"])
trainer_ver2.evaluate(tokenized_ds_ver2["test"])
```

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base. You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```
Map: 100% 35701/35701 [00:12<00:00, 8066.04 examples/s]
```

```
Map: 100% 6589/6589 [00:03<00:00, 2212.41 examples/s]
```

```
Map: 100% 6016/6016 [00:02<00:00, 2702.62 examples/s]
```

```
tokenized_ds_ver2['train'][0]: {'label': 1, 'text': "I went to the principal's office to change my records before goi
/tmp/ipython-input-969342480.py:45: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for
trainer_ver2 = Trainer(
[2232/2232 09:16, Epoch 1/1]
```

Step Training Loss

Step	Training Loss
500	0.657200
1000	0.606700
1500	0.581900
2000	0.570600

```
[206/206 00:40]
```

```
{'eval_loss': 0.8556240797042847,
'eval_accuracy': 0.4820478723404255,
'eval_f1': 0.49250814332247556,
'eval_precision': 0.5041680560186729,
'eval_recall': 0.4813753581661891,
'eval_runtime': 19.8722,
'eval_samples_per_second': 302.734,
'eval_steps_per_second': 9.46,
'epoch': 1.0}
```

```
all_results["full_v2"] = {
    "val": trainer_ver2.evaluate(tokenized_ds_ver2["validation"]),
    "test": trainer_ver2.evaluate(tokenized_ds_ver2["test"]),
}

trainer_ver2.save_model("./ethics-distilbert-full_ver2")
tokenizer_ver2.save_pretrained("./ethics-distilbert-full_ver2")
```

```
[206/206 01:21]
("./ethics-distilbert-full_ver2/tokenizer_config.json",
"./ethics-distilbert-full_ver2/special_tokens_map.json",
"./ethics-distilbert-full_ver2/vocab.txt",
"./ethics-distilbert-full_ver2/added_tokens.json",
"./ethics-distilbert-full_ver2/tokenizer.json")
```

commonsense

```
model_name = "distilbert-base-uncased"

tokenizer_ver2_2 = AutoTokenizer.from_pretrained(model_name, use_fast=True)

config_ver2_2 = AutoConfig.from_pretrained(
    model_name,
    num_labels=2,
    dropout=0.2,
    attention_dropout=0.2,
)

model_ver2_2 = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    config=config_ver2_2,
)

def tokenize_batch(batch):
    return tokenizer_ver2_2(
        batch["text"],
        truncation=True,
        padding="max_length",
        max_length=128,
```

```

        )

        tokenized_ds_ver2_2 = data_1_2.map(tokenize_batch, batched=True)

        print("tokenized_ds_ver2_2['train'][0]:", tokenized_ds_ver2_2['train'][0])

        tokenized_ds_ver2_2 = tokenized_ds_ver2_2.remove_columns(["text", "source"])
        tokenized_ds_ver2_2.set_format("torch")

        # Check there are label, input_ids and attention_mask
        tokenized_ds_ver2_2

        training_args_ver2_2 = TrainingArguments(
            output_dir=".ethics-distilbert-commonsense_ver2_2",
            num_train_epochs=1,
            per_device_train_batch_size=16,
            per_device_eval_batch_size=32,
            learning_rate=2e-5,
            weight_decay=0.01,
            report_to="none",
            label_smoothing_factor=0.1
        )
        trainer_ver2_2 = Trainer(
            model=model_ver2_2,
            args=training_args_ver2_2,
            train_dataset=tokenized_ds_ver2_2["train"],
            eval_dataset=tokenized_ds_ver2_2["validation"],
            tokenizer=tokenizer_ver2_2,
            compute_metrics=compute_metrics,
        )
        trainer_ver2_2.train()

        trainer_ver2_2.evaluate(tokenized_ds_ver2_2["validation"])
        trainer_ver2_2.evaluate(tokenized_ds_ver2_2["test"])
    
```

Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base. You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```

Map: 100%           13910/13910 [00:08<00:00, 1971.82 examples/s]

Map: 100%           3885/3885 [00:02<00:00, 1049.35 examples/s]

Map: 100%           3964/3964 [00:02<00:00, 1293.01 examples/s]

```

tokenized_ds_ver2_2['train'][0]: {'label': 1, 'text': "I went to the principal's office to change my records before g /tmp/ipython-input-2172740696.py:45: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 fo

trainer_ver2_2 = Trainer(

[870/870 04:00, Epoch 1/1]

Step Training Loss

500	0.648900
-----	----------

[122/122 00:24]

```
{
  'eval_loss': 0.890516459941864,
  'eval_accuracy': 0.43466195761856713,
  'eval_f1': 0.4128897039559864,
  'eval_precision': 0.45654692931633833,
  'eval_recall': 0.3768531802965088,
  'eval_runtime': 12.6596,
  'eval_samples_per_second': 313.122,
  'eval_steps_per_second': 9.795,
  'epoch': 1.0
}
```

```

all_results["commonsense_v2"] = {
    "val": trainer_ver2_2.evaluate(tokenized_ds_ver2_2["validation"]),
    "test": trainer_ver2_2.evaluate(tokenized_ds_ver2_2["test"]),
}

trainer_ver2_2.save_model("./ethics-distilbert-commonsense_ver2_2")
tokenizer_ver2_2.save_pretrained("./ethics-distilbert-commonsense_ver2_2")

```

[122/122 00:50]

```
('./ethics-distilbert-commonsense_ver2_2/tokenizer_config.json',
 './ethics-distilbert-commonsense_ver2_2/special_tokens_map.json',
 './ethics-distilbert-commonsense_ver2_2/vocab.txt',
 './ethics-distilbert-commonsense_ver2_2/added_tokens.json',
 './ethics-distilbert-commonsense_ver2_2/tokenizer.json')
```

Justice

```
model_name = "distilbert-base-uncased"

tokenizer_ver2_3 = AutoTokenizer.from_pretrained(model_name, use_fast=True)

config_ver2_3 = AutoConfig.from_pretrained(
    model_name,
    num_labels=2,
    dropout=0.2,
    attention_dropout=0.2,
)

model_ver2_3 = AutoModelForSequenceClassification.from_pretrained(
    model_name,
    config=config_ver2_3,
)

def tokenize_batch(batch):
    return tokenizer_ver2_3(
        batch["text"],
        truncation=True,
        padding="max_length",
        max_length=128,
    )

tokenized_ds_ver2_3 = data_1_3.map(tokenize_batch, batched=True)

print("tokenized_ds_ver2_3['train'][0]:", tokenized_ds_ver2_3['train'][0])

tokenized_ds_ver2_3 = tokenized_ds_ver2_3.remove_columns(["text", "source"])
tokenized_ds_ver2_3.set_format("torch")

# Check there are label, input_ids and attention_mask
tokenized_ds_ver2_3

training_args_ver2_3 = TrainingArguments(
    output_dir=".//ethics-distilbert-justice_ver2_3",
    num_train_epochs=1,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=32,
    learning_rate=2e-5,
    weight_decay=0.01,
    report_to="none",
    label_smoothing_factor=0.1
)
trainer_ver2_3 = Trainer(
    model=model_ver2_3,
    args=training_args_ver2_3,
    train_dataset=tokenized_ds_ver2_3["train"],
    eval_dataset=tokenized_ds_ver2_3["validation"],
    tokenizer=tokenizer_ver2_3,
    compute_metrics=compute_metrics,
)
trainer_ver2_3.train()

trainer_ver2_3.evaluate(tokenized_ds_ver2_3["validation"])
trainer_ver2_3.evaluate(tokenized_ds_ver2_3["test"])
```

```
Some weights of DistilBertForSequenceClassification were not initialized from the model checkpoint at distilbert-base
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

```
Map: 100% 21791/21791 [00:02<00:00, 8096.17 examples/s]
```

```
Map: 100% 2704/2704 [00:00<00:00, 7273.36 examples/s]
```

```
Map: 100% 2052/2052 [00:00<00:00, 7857.09 examples/s]
```

```
tokenized_ds_ver2_3['train'][0]: {'label': 1, 'text': 'I deserve to be paid by my boss because I keep her house clean
/tmp/ipython-input-4272733173.py:46: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 fo
trainer_ver2_3 = Trainer(
```

```
[1362/1362 05:49, Epoch 1/1]
```

Step Training Loss

500	0.613900
-----	----------

1000	0.539400
------	----------

```
[85/85 00:14]
```

```
{'eval_loss': 0.8458858132362366,
'eval_accuracy': 0.5384990253411306,
'eval_f1': 0.5880817746846455,
'eval_precision': 0.5412329863891113,
'eval_recall': 0.6438095238095238,
'eval_runtime': 6.5174,
'eval_samples_per_second': 314.85,
'eval_steps_per_second': 9.973,
'epoch': 1.0}
```

```
all_results["justice_v2"] = {
    "val": trainer_ver2_3.evaluate(tokenized_ds_ver2_3["validation"]),
    "test": trainer_ver2_3.evaluate(tokenized_ds_ver2_3["test"]),
}

trainer_ver2_3.save_model("./ethics-distilbert-justice_ver2_3")
tokenizer_ver2_3.save_pretrained("./ethics-distilbert-justice_ver2_3")
```

```
[85/85 00:30]
```

```
( './ethics-distilbert-justice_ver2_3/tokenizer_config.json',
 './ethics-distilbert-justice_ver2_3/special_tokens_map.json',
 './ethics-distilbert-justice_ver2_3/vocab.txt',
 './ethics-distilbert-justice_ver2_3/added_tokens.json',
 './ethics-distilbert-justice_ver2_3/tokenizer.json' )
```

```
for model_name, splits in all_results.items():
    print(f"\n##### {model_name} #####")
    for split, metrics in splits.items():
        print(f"\n--- {split.upper()} ---")
        for k, v in metrics.items():
            print(f"{k}: {v}")

##### full_v1 #####

```

```
--- VAL ---
eval_loss: 0.5578117370605469
eval_accuracy: 0.7866140537259068
eval_f1: 0.7812013694366635
eval_precision: 0.7708845208845209
eval_recall: 0.7917981072555205
eval_runtime: 21.0165
eval_samples_per_second: 313.516
eval_steps_per_second: 9.802
epoch: 3.0
```

```
--- TEST ---
eval_loss: 1.0873825550079346
eval_accuracy: 0.49052526595744683
eval_f1: 0.491455118632819
eval_precision: 0.5131670131670132
eval_recall: 0.4715058898439987
eval_runtime: 19.0912
eval_samples_per_second: 315.119
eval_steps_per_second: 9.847
epoch: 3.0
```

```
##### commonsense_v1 #####

```

```
--- VAL ---
eval_loss: 0.4742743670940399
```

```
eval_accuracy: 0.8275418275418276
eval_f1: 0.8155286343612335
eval_precision: 0.8155286343612335
eval_recall: 0.8155286343612335
eval_runtime: 12.5983
eval_samples_per_second: 308.375
eval_steps_per_second: 9.684
epoch: 3.0

--- TEST ---
eval_loss: 1.1184661388397217
eval_accuracy: 0.45408678102926336
eval_f1: 0.4343962362780972
eval_precision: 0.4789625360230548
eval_recall: 0.39741750358680056
eval_runtime: 12.6405
eval_samples_per_second: 313.595
eval_steps_per_second: 9.81
epoch: 3.0

##### justice_v1 #####

```

```
--- VAL ---
eval_loss: 0.6820999383926392
eval_accuracy: 0.7414940828402367
eval_f1: 0.7580477673935618
eval_precision: 0.7133550488599348
eval_recall: 0.808714918759232
eval_runtime: 9.0037
```

```
import matplotlib.pyplot as plt
import numpy as np

model_names = []
val_f1_list = []
test_f1_list = []

for model_name, splits in all_results.items():
    model_names.append(model_name)
    val_f1_list.append(splits["val"]["eval_f1"])
    test_f1_list.append(splits["test"]["eval_f1"])

x = np.arange(len(model_names))
width = 0.35

fig, ax = plt.subplots(figsize=(9, 4))

bars_val = ax.bar(x - width/2, val_f1_list, width, label="VAL F1")
bars_test = ax.bar(x + width/2, test_f1_list, width, label="TEST F1")

ax.set_xticks(x)
ax.set_xticklabels(model_names, rotation=45, ha="right")
ax.set_ylabel("F1 score")
ax.set_title("ETHICS models - Validation vs Test F1")
ax.legend()

def add_labels(bars):
    for b in bars:
        height = b.get_height()
        ax.text(
            b.get_x() + b.get_width()/2,
            height + 0.01,
            f"{height:.3f}",
            ha="center",
            va="bottom",
            fontsize=8,
        )

# Add labels to the bars
add_labels(bars_val)
add_labels(bars_test)
```