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

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?

[13] 0s

```
import torch
import numpy as np
import pandas as pd
import re
from torch.utils.data import Dataset, DataLoader
import torch.nn.functional as F #for softmax
import matplotlib.pyplot as plt

from transformers import AutoTokenizer, AutoModelForSequenceClassification, AutoModel
```

[14] 0s

```
from google.colab import drive
drive.mount('/content/drive')

import os
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

(why blocked codes) I saved model to local in the previous notebook. But it will take too much time to run and make outputs. I need to save it in drive to use again, so below codes are only for saving.

[15] 0s

```
# from datasets import Dataset, DatasetDict
# from transformers import (
#     TrainingArguments,
#     Trainer,
# )
# from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score


# 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 = 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"})


# 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(justice_train_df.head(2))

# model_name = "distilbert-base-uncased"

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

# 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,
#     }

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

# 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,
# })

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

# 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")

# 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()


# save_dir1 = "/content/drive/MyDrive/ethics_models/justice_v1"
# os.makedirs(save_dir1, exist_ok=True)

# trainer3.save_model(save_dir1)
# tokenizer.save_pretrained(save_dir1)

# print("justice_v2 Saved", save_dir1)

# print("Validation metrics:")
# print(trainer3.evaluate(tokenized_ds_3["validation"]))

# print("Test metrics:")
# print(trainer3.evaluate(tokenized_ds_3["test"]))
```

[15] ✓ 0s

```
Start coding or generate with AI.
```

Check the saved models

[16] ✓ 0s

```
save_dir1 = "/content/drive/MyDrive/ethics_models/justice_v1"  # v1

print(os.listdir(save_dir1))
```

```
['config.json', 'model.safetensors', 'tokenizer_config.json', 'special_tokens_map.json', 'vocab.txt', 'tokenizer.json', 'training_args.bin']
```

[17]

```
tokenizer_v1 = AutoTokenizer.from_pretrained(save_dir1)
model_v1 = AutoModelForSequenceClassification.from_pretrained(save_dir1)

model_v1.eval()

print("justice_v1 id2label:", model_v1.config.id2label)
# 0: unfair(unethical), 1: fair(ethical)
```

```
justice_v1 id2label: {0: 'LABEL_0', 1: 'LABEL_1'}
```

AI fake news data set

[18] ✓ 0s

```
from datasets import load_dataset
```

[19] ✓ 2s

```
ds = load_dataset("lvulpecula/ai_watermarked_fake_news-v2")

df_news = ds["train"].to_pandas()

print("AI news:", df_news.columns)
df_news.head()
```

```
AI news: Index(['title', 'text', 'model', 'label'], dtype='object')
```

	title	text	model	label
0	Vladimir Putin is friends with Bigfoot	In a shocking revelation that is sure ...	ChatGPT	False
1	Twitter is shutting down	After years of dominating the social...	ChatGPT	False
2	Scientist have invented a machine for teleport...	In a stunning breakthrough, scientists ...	ChatGPT	False
3	Elon Musk has bought the moon	Sources close to Musk's space explo...	ChatGPT	False
4	Black Death returns to Europe	In a startling development that has...	ChatGPT	False

Next steps: Generate code with df_news | New interactive sheet

[20] ✓ 0s

```
# I will split the articles to sentences and then evaluate
def split_into_sentences(text: str):

    text = str(text).strip()
    if not text:
        return []

    # . ? !
    sentences = re.split(r'(?<=[.!?])+', text)
    sentences = [s.strip() for s in sentences if s.strip()]
    return sentences

# From here using function
rows = []

for idx, row in df_news.iterrows():
    article_id = idx                    # #of articles
    text = row["text"]                  # main text
    title = row.get("title", None)      # title
    src_model = row.get("model", None)  # chatGPT 95.9%

    if isinstance(title, str) and title.strip():
        rows.append({
            "article_id": article_id,
            "sent_idx": -1,               # titile index is -1 because title is almost important
            "is_title": True,
            "title": title,
            "source_model": src_model,
            "sentence": title.strip(),
        })

    sentences = split_into_sentences(text)
    for sent_idx, sent in enumerate(sentences):
        rows.append({
            "article_id": article_id,
            "sent_idx": sent_idx,
            "is_title": False,
            "title": title,
            "source_model": src_model,
            "sentence": sent,
        })

df_sent = pd.DataFrame(rows)
df_sent["len(sentence)"] = df_sent["sentence"].str.len()
print("How many sentences:", len(df_sent))
```

```
df_sent.head(15)
```

```
How many santences: 22723
```

	article_id	sent_idx	is_title	title	source_model	sentence	len(sentence)
0	0	-1	True	Vladimir Putin is friends with Bigfoot	ChatGPT	Vladimir Putin is friends with Bigfoot	38
1	0	0	False	Vladimir Putin is friends with Bigfoot	ChatGPT	In a shocking revelation that is sure ...	261
2	0	1	False	Vladimir Putin is friends with Bigfoot	ChatGPT	According to sources close to the Kre...	200
3	0	2	False	Vladimir Putin is friends with Bigfoot	ChatGPT		1
4	1	-1	True	Twitter is shutting down	ChatGPT	Twitter is shutting down	24
5	1	0	False	Twitter is shutting down	ChatGPT	After years of dominating the social...	168
6	1	1	False	Twitter is shutting down	ChatGPT	The decision comes as a surprise ...	176
7	1	2	False	Twitter is shutting down	ChatGPT	In a statement released by the company,...	184
8	1	3	False	Twitter is shutting down	ChatGPT	Despite efforts to pivot the platform ...	249
9	1	4	False	Twitter is shutting down	ChatGPT	\n\nThe announcement has sent shockwav...	225
10	1	5	False	Twitter is shutting down	ChatGPT	Many have taken to the platform to express...	203
11	1	6	False	Twitter is shutting down	ChatGPT	\n\nThe shutdown of Twitter marks the e...	172
12	2	-1	True	Scientist have invented a machine for teleport...	ChatGPT	Scientist have invented a machine for teleport...	51
13	2	0	False	Scientist have invented a machine for teleport...	ChatGPT	In a stunning breakthrough, scientists ...	149
14	2	1	False	Scientist have invented a machine for teleport...	ChatGPT	The invention, which has been the s...	164

Next steps: Generate code with df_sent | New interactive sheet

[21] ✓ 0s

```
#df_sent.drop(index=8125, inplace=True) # index 8125's length is more than 2000 and it's broken sentence. so I drop it.
df_sent.sort_values(by="len(sentence)", ascending=False, inplace=True)
df_sent.head(5)
```

	article_id	sent_idx	is_title	title	source_model	sentence	len(sentence)
8125	329	16	False	Famous Actor Found Living Secret Double Life a...	llama 3.1	com/famous-actor-found-living-secret...	2011
9618	375	8	False	Florida Bans HRT for Transgender Individuals A...	ChatGPT	The legislation specifically cites co...	480
9265	367	0	False	Study Claims Watching Excessive Anime Could Ca...	ChatGPT	Tokyo, Japan Ð A controversial new stud...	451
5486	236	20	False	Wall Street Stock Exchange Plummets: Global Ma...	ChatGPT	"\n\nLong-term Outlook\n\nWhile the imme...	418
9610	375	0	False	Florida Bans HRT for Transgender Individuals A...	ChatGPT	January 16, 2025 Ð In a highly cont...	416

Next steps: Generate code with df_sent | New interactive sheet

[22] ✓ 0s

```
sentences = df_sent["sentence"].tolist()

print(type(sentences))
print(type(sentences[0]))
```

```
<class 'list'>
<class 'str'>
```

[23] ✓ 0s

```
model_v1.to("cuda")   # GPU
model_v1.eval()
```

```
DistilBertForSequenceClassification(
  (distilbert): DistilBertModel(
    (embeddings): Embeddings(
      (word_embeddings): Embedding(30522, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (transformer): Transformer(
      (layer): ModuleList(
        (0-5): 6 x TransformerBlock(
          (attention): DistilBertSdpaAttention(
            (dropout): Dropout(p=0.1, inplace=False)
            (q_lin): Linear(in_features=768, out_features=768, bias=True)
            (k_lin): Linear(in_features=768, out_features=768, bias=True)
            (v_lin): Linear(in_features=768, out_features=768, bias=True)
            (out_lin): Linear(in_features=768, out_features=768, bias=True)
          )
          (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
          (ffn): FFN(
            (dropout): Dropout(p=0.1, inplace=False)
            (lin1): Linear(in_features=768, out_features=3072, bias=True)
            (lin2): Linear(in_features=3072, out_features=768, bias=True)
            (activation): GELUActivation()
          )
          (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        )
      )
    )
  )
  (pre_classifier): Linear(in_features=768, out_features=768, bias=True)
  (classifier): Linear(in_features=768, out_features=2, bias=True)
  (dropout): Dropout(p=0.2, inplace=False)
)
```

[24] ✓ 1m

```
# Justice
batch_size = 32
MAX_LEN = 256

all_probs = []
all_preds = []

with torch.no_grad():  # because I'm doing only inference
    for start in range(0, len(sentences), batch_size):
        batch_sents = sentences[start:start + batch_size]

        enc = tokenizer_v1(
            batch_sents,
            truncation=True,
            padding="max_length",
            max_length=MAX_LEN,
            return_tensors="pt",
        )

        enc = {k: v.to("cuda") for k, v in enc.items()}

        outputs = model_v1(**enc)
        logits = outputs.logits

        probs = F.softmax(logits, dim=-1).cpu().numpy()
        preds = np.argmax(probs, axis=-1)

        all_probs.append(probs)
        all_preds.append(preds)
```

```
all_probs = np.concatenate(all_probs, axis=0)
all_preds = np.concatenate(all_preds, axis=0)

len(all_probs), len(all_preds), len(df_sent)
```

```
(22723, 22723, 22723)
```

[25] ✓ 0s

```
unethical_id = 0

df_sent["pred_label_id"] = all_preds
df_sent["pred_label_name"] = df_sent["pred_label_id"].map(model_v1.config.id2label)
df_sent["prob_unethical"] = all_probs[:, unethical_id]
df_sent["is_unethical"] = df_sent["pred_label_id"] == unethical_id

df_sent[["article_id", "sent_idx", "is_title", "sentence",
         "pred_label_id", "prob_unethical", "is_unethical"]].head(20)
```

	article_id	sent_idx	is_title	sentence	pred_label_id	prob_unethical	is_unethical
8125	329	16	False	com/famous-actor-found-living-secret...	1	0.088882	False
9618	375	8	False	The legislation specifically cites co...	1	0.077885	False
9265	367	0	False	Tokyo, Japan Đ A controversial new stud...	1	0.136299	False
5486	236	20	False	"\n\nLong-term Outlook\n\nWhile the imme...	1	0.084209	False
9610	375	0	False	January 16, 2025 Đ In a highly cont...	1	0.418063	False
1102	52	21	False	As Germany endures this unparalleled he...	1	0.212503	False
4721	213	21	False	\nAs players gear up to embark on ...	1	0.087810	False
9268	367	3	False	According to the research, indivi...	1	0.312930	False
13072	683	12	False	The administrationÕs moratorium on...	0	0.563191	True
4744	214	20	False	With the support of policymakers, he...	1	0.331160	False
5018	222	54	False	\n\n\n\n\n\n\nIn a deeply personal an...	1	0.141656	False
7068	286	21	False	"\n\nConclusion\n\nAs speculation and d...	0	0.884719	True
9340	369	0	False	January 15, 2025 Đ In an unprecede...	1	0.412323	False
21465	1432	14	False	Ó\n\nBreaking News: Governments that ...	1	0.447488	False
20899	1401	5	False	"\n\nDespite being touted as an alte...	0	0.581231	True
6948	281	6	False	"\n\nRationale for the Denunciations\n...	1	0.095374	False
21093	1411	0	False	Bitcoin, the worldÕs most famous cry...	0	0.798783	True
4896	220	15	False	\n\n\n\n\nIn a stunning departure from i...	0	0.608259	True
14660	820	3	False	Margaret Steele, a senior research...	1	0.118076	False
14804	840	3	False	Emily Collins, a clinical psychologi...	0	0.560335	True

[26] ✓ 0s

```
article_stats = (
    df_sent
    .groupby("article_id")
    .agg(
        total_sentences = ("sentence", "count"),
        unethical_sentences = ("is_unethical", "sum"),
        avg_prob_unethical = ("prob_unethical", "mean"),
        max_prob_unethical = ("prob_unethical", "max"),
    )
    .reset_index()
)

article_stats["ratio_unethical"] = (
    article_stats["unethical_sentences"] / article_stats["total_sentences"] * 100
)

article_stats.head()
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical
0	0	4	3	0.674596	0.944441	75.000000
1	1	8	1	0.287587	0.588074	12.500000
2	2	11	3	0.350745	0.672076	27.272727
3	3	9	7	0.642474	0.928792	77.777778
4	4	15	3	0.290165	0.712886	20.000000

Next steps: Generate code with article_stats | New interactive sheet

[27] ✓ 0s

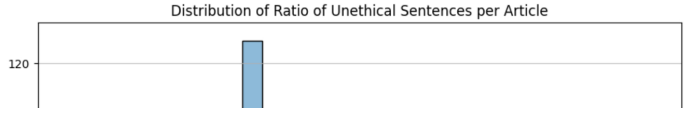
```
THRESH_RATIO = 10.0  # 10%

article_stats["flag_ratio_10"] = (
    article_stats["ratio_unethical"] >= THRESH_RATIO
)
article_stats.head()
```

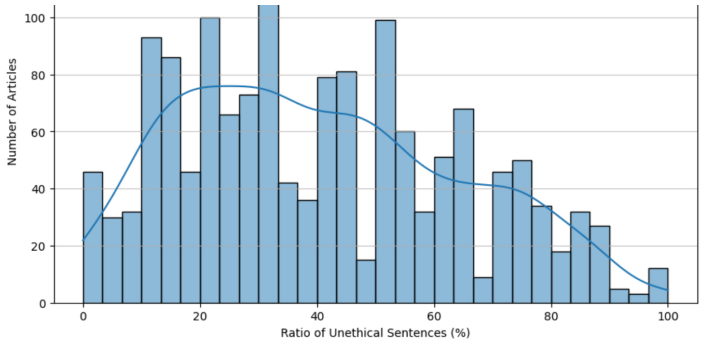
	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10
0	0	4	3	0.674596	0.944441	75.000000	True
1	1	8	1	0.287587	0.588074	12.500000	True
2	2	11	3	0.350745	0.672076	27.272727	True
3	3	9	7	0.642474	0.928792	77.777778	True
4	4	15	3	0.290165	0.712886	20.000000	True

Next steps: Generate code with article_stats | New interactive sheet

[28] ✓ 0s

```
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.histplot(article_stats['ratio_unethical'], bins=30, kde=True)
plt.title('Distribution of Ratio of Unethical Sentences per Article')
plt.xlabel('Ratio of Unethical Sentences (%)')
plt.ylabel('Number of Articles')
plt.grid(axis='y', alpha=0.75)
plt.show()
```





[29]
```
article_stats.sort_values(by="ratio_unethical", ascending=True, inplace=True)
display(article_stats.head())
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10
23	23	22	0	0.169739	0.407683	0.0	False
39	39	23	0	0.189132	0.480308	0.0	False
568	568	9	0	0.210332	0.406342	0.0	False
586	586	9	0	0.282861	0.485771	0.0	False
499	499	9	0	0.246330	0.394795	0.0	False

[30]
```
article_stats.sort_values(by="ratio_unethical", ascending=False, inplace=True)
display(article_stats.head())
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10
1199	1199	8	8	0.656891	0.817125	100.0	True
1137	1137	8	8	0.781299	0.913082	100.0	True
1131	1131	7	7	0.682540	0.856869	100.0	True
1366	1366	10	10	0.793172	0.941447	100.0	True
868	868	10	10	0.803228	0.930771	100.0	True

Selected Articles based on `ratio_unethical` categories

[31]
```
print('--- Articles representing different ratio_unethical percentages ---')
selected_article_ids = []

for target_ratio in range(0, 101, 10):
    # Find articles within a small range around the target_ratio
    # Use a small epsilon to create a range for matching
    epsilon = 0.5 # e.g., for 10%, look for 9.5% to 10.5%

    # Prioritize articles that haven't been selected yet for diversity
    potential_articles = article_stats[
        (article_stats['ratio_unethical'] >= target_ratio - epsilon) &
        (article_stats['ratio_unethical'] < target_ratio + epsilon) &
        (~article_stats['article_id'].isin(selected_article_ids))
    ]

    if potential_articles.empty:
        # If no new articles found, try without excluding already selected ones
        potential_articles = article_stats[
            (article_stats['ratio_unethical'] >= target_ratio - epsilon) &
            (article_stats['ratio_unethical'] < target_ratio + epsilon)
        ]

    if not potential_articles.empty:
        # Pick the first one found for simplicity
        selected_article = potential_articles.iloc[0]
        selected_article_ids.append(selected_article['article_id'])
        print(f"\nArticle with ratio_unethical around {target_ratio}%:")
        display(selected_article[['article_id', 'ratio_unethical', 'unethical_sentences', 'total_sentences']])
    else:
        print(f"\nNo article found for ratio_unethical around {target_ratio}%")
```

unethical_sentences	3
total_sentences	5

dtype: object

Article with ratio_unethical around 70%:

	858
article_id	858
ratio_unethical	70.0
unethical_sentences	7
total_sentences	10

dtype: object

Article with ratio_unethical around 80%:

	1453
article_id	1453
ratio_unethical	80.0
unethical_sentences	16
total_sentences	20

dtype: object

Article with ratio_unethical around 90%:

	1483
article_id	1483
ratio_unethical	90.47619
unethical_sentences	19
total_sentences	21

dtype: object

Article with ratio_unethical around 100%:

	1199
article_id	1199
ratio_unethical	100.0
unethical_sentences	8
total_sentences	8

dtype: object

[32] ✓ 0s

```
selected_article_ids_simple = []
pd.set_option('display.max_colwidth', None) # Ensure full text is displayed

print('--- Articles with exact ratio_unethical percentages ---')
for target_ratio in range(0, 101, 10):
    exact_match_articles = article_stats[article_stats['ratio_unethical'] == float(target_ratio)]

    if not exact_match_articles.empty:
        # Pick the first one found for simplicity
        selected_article = exact_match_articles.iloc[0]
        selected_article_ids_simple.append(selected_article['article_id'])

        print(f"\nArticle with exact ratio_unethical = {target_ratio}%:")

        output_data = pd.Series({
            'article_id': selected_article['article_id'],
            'ratio_unethical': selected_article['ratio_unethical'],
            'title': df_news.loc[selected_article['article_id'], 'title'],
            'text': df_news.loc[selected_article['article_id'], 'text']
        })
        display(output_data)
        print("-" * 50)
    else:
        print(f"\nNo article found for exact ratio_unethical = {target_ratio}% (might be due to floating point precision).")
        print("-" * 50)

pd.set_option('display.max_colwidth', 50) # Reset to a default or desired width after displaying
```

article_id	1453
ratio_unethical	80.0
title	Vampires Now Offering ÔEternal LifeÕ ToursÑIs It Really a Dream Vacation or a One-Way Ticket to Immortality?
text	Want to experience the thrill of immortality? According to recent reports, vampires are now offering exclusive ÒEternal LifeÓ tours, giving unsuspecting tourists the chance to travel to vampire enclaves and experience their dark and mysterious lifestyle. But is this an opportunity for adventure, or a dangerous trap to turn tourists into vampires?\n\nThe concept is simple: for a hefty price, tourists can travel to secret locations known only to the vampire elite, where theyÕll be treated to lavish accommodations, fine dining, and the opportunity to meet some of the worldÕs most ancient and powerful vampires. But thereÕs a catchÑthose who take part in these tours are said to be at risk of being turned into vampires themselves.\n\nÒVampires have been recruiting for centuries,Ó says former vampire hunter Rick Thorne. ÒBut now, theyÕre making it easier for people to willingly join their ranks. These so-called ÔEternal LifeÕ tours are nothing more than a recruitment tool. They get people to come to their hidden locations, show them how alluring their lifestyle is, and then offer them the ÔgiftÕ of immortality.Ó\n\nThorne warns that these tours are designed to seduce people into making an irreversible decision. ÒItÕs all part of their plan to increase their ranks,Ó he says. ÒOnce youÕre on the tour, theyÕll slowly begin to break down your willpower, making it harder for you to leave without joining their bloodthirsty cause.Ó\n\nTravel experts are also raising alarms about the potential dangers of these tours. ÒItÕs like a modern-day vampire cult,Ó says travel journalist Sophie Richards. ÒThese tours are advertised as the ultimate luxury vacation, but they come with a high price: your soul.Ó\n\nExperts caution that if youÕre considering a trip that promises Òimmortality,Ó it might be best to rethink the idea. ÒVampires donÕt give away immortality for free,Ó Thorne warns. ÒOnce you step into their world, you may never leave again.Ó

dtype: object

Article with exact ratio_unethical = 90%:

	0
article_id	1239
ratio_unethical	90.0
title	Global Warming Conspiracy: The EliteÕs Plan to Control the World
text	Global warming isnÕt just a weather issueÑitÕs the battleground for a global power struggle. While the media bombards us with terrifying headlines about rising temperatures, ice caps melting, and extreme weather events, many experts claim this is all part of a much bigger, more sinister plot. Could the global warming narrative be a carefully crafted scheme to centralize power, control economies, and create a system where every aspect of our lives is dictated by a handful of elites?\n\nThe "solutions" being proposedÑsuch as massive carbon taxes, an overhaul of the energy sector, and restrictions on individual freedomsÑare conveniently paving the way for corporate elites to dominate the green energy market. Governments are taking control of what people eat, what cars they drive, and even how much energy they consume, all while promoting the so-called Ògreen revolution.Ó But what if the true goal of this green agenda is not to protect the environment, but to create a society where the global elite have total control over resources, energy, and wealth?\n\nBehind the scenes, powerful companies linked to political leaders stand to profit immensely from the push for renewable energy, creating a monopoly that pushes smaller, independent energy providers out of business. Additionally, the relentless push for electric vehicles, solar power, and wind energy has turned the global warming debate into a marketplace where the only winners are those who have deep financial ties to the green industry.\n\nWhile the world debates rising sea levels and melting glaciers, the real question should be: who stands to gain the most from global warming policies, and whoÕs losing in this high-stakes game for control?

dtype: object

Article with exact ratio_unethical = 100%:

	0
article_id	1199
ratio_unethical	100.0
title	The United NationsÕ Agenda to Control Food and Agriculture: A Global Power Grab
text	The United Nations is allegedly orchestrating a hidden plan to take control of global food and agriculture production, making nations more reliant on international bodies for their basic food needs. Leaked documents have revealed that the UN is promoting policies that could lead to widespread food shortages, with the ultimate goal of centralizing control over agriculture under its authority.\n\nSources say the UNÕs push for global agricultural standards, which include the implementation of stricter regulations on farming practices, could severely disrupt local food production. While these policies are often presented as solutions to climate change or food insecurity, critics argue that they are designed to pave the way for multinational corporations to dominate the global food market, leaving countries at the mercy of international organizations like the UN.\n\nIn effect, the UNÕs policies could result in the loss of local food sovereignty, forcing nations to rely on global supply chains controlled by a select few powerful entities. This would allow the UN to exercise significant influence over national economies by controlling the very thing that sustains them: food.\n

dtype: object

I could find, justice ca't cover everything. So, add more labe data set

hatexplain

[33] ✓ 0s

```
# pd.set_option('display.max_colwidth', None)
# df_sent[df_sent["article_id"] == 1098]
# pd.set_option('display.max_colwidth', 50) # Reset to a default or desired width after displaying
```

[34] ✓ 9s

```
save_dir_hate = "/content/drive/MyDrive/models/hatexplain_distilbert"

tokenizer_hate = AutoTokenizer.from_pretrained(save_dir_hate)
model_hate = AutoModelForSequenceClassification.from_pretrained(save_dir_hate)
model_hate.eval()

print("hate model id2label:", model_hate.config.id2label)
```

```
hate model id2label: {0: 'hatespeech', 1: 'normal', 2: 'offensive'}
```

[35] ✓ 8m

```
batch_size = 32
MAX_LEN = 128

hate_probs = []
hate_preds = []

with torch.no_grad():
    for start in range(0, len(sentences), batch_size):
        batch_sents = sentences[start:start + batch_size]

        enc = tokenizer_hate(
            batch_sents,
            truncation=True
```

```
            truncation=True,
            padding="max_length",
            max_length=MAX_LEN,
            return_tensors="pt",
        ).to(model_hate.device)

        outputs = model_hate(**enc)
        logits = outputs.logits

        probs = F.softmax(logits, dim=-1).cpu().numpy()
        preds = np.argmax(probs, axis=-1)

        hate_probs.append(probs)
        hate_preds.append(preds)

hate_probs = np.concatenate(hate_probs, axis=0)
hate_preds = np.concatenate(hate_preds, axis=0)

len(hate_probs), len(hate_preds), len(df_sent)
```

```
(22723, 22723, 22723)
```

[36] ✓ 0s

```
id2label_hate = {i: name.lower() for i, name in model_hate.config.id2label.items()}
print("id2label_hate:", id2label_hate)

hate_id = [i for i, name in id2label_hate.items() if "hate" in name][0]
offensive_id = [i for i, name in id2label_hate.items() if "offensive" in name][0]
normal_id = [i for i, name in id2label_hate.items() if "normal" in name][0]

print("hate_id:", hate_id, "offensive_id:", offensive_id, "normal_id:", normal_id)

df_sent["p_hate"] = hate_probs[:, hate_id]
df_sent["p_offensive"] = hate_probs[:, offensive_id]
df_sent["p_normal"] = hate_probs[:, normal_id]

#hatespeech + offensive
df_sent["prob_hate_offensive"] = df_sent["p_hate"] + df_sent["p_offensive"]

df_sent["is_hate_offensive"] = df_sent["prob_hate_offensive"] > 0.5

df_sent["hate_top_id"] = np.argmax(hate_probs, axis=1)
id2label_for_top = {int(k): v for k, v in id2label_hate.items()}
df_sent["hate_top_label"] = df_sent["hate_top_id"].map(id2label_for_top)

pd.set_option('display.max_colwidth', 80) # Reset to a default or desired width after displaying

df_sent[[
    "article_id", "sent_idx", "is_title", "sentence",
    "p_hate", "p_offensive", "p_normal",
    "prob_hate_offensive", "hate_top_label", "is_hate_offensive"
]].head(15)
```

```
id2label_hate: {0: 'hatespeech', 1: 'normal', 2: 'offensive'}
hate_id: 0 offensive_id: 2 normal_id: 1
```

	article_id	sent_idx	is_title	sentence	p_hate	p_offensive	p_normal	prob_hate_offensive	hate_top_label	is_hate_offensive
8125	329	16	False	com/famous-actor-found-living-secret-double-life-pizza-delivery...	0.024449	0.033690	0.941861	0.058139	normal	False
9618	375	8	False	The legislation specifically cites concerns over the long-te...	0.037834	0.111256	0.850910	0.149090	normal	False
9265	367	0	False	Tokyo, Japan Ð A controversial new study published in the Journ...	0.040626	0.080415	0.878959	0.121041	normal	False
5486	236	20	False	"\n\nLong-term Outlook\n\nWhile the immediate outlook for Wall St...	0.029635	0.061696	0.908670	0.091330	normal	False
9610	375	0	False	January 16, 2025 Ð In a highly contentious move, the state of...	0.032597	0.100989	0.866414	0.133586	normal	False
1102	52	21	False	As Germany endures this unparalleled heatwave, the resilience an...	0.090708	0.117464	0.791828	0.208172	normal	False
4721	213	21	False	\nAs players gear up to embark on virtual missions and en...	0.031822	0.056577	0.911601	0.088399	normal	False
9268	367	3	False	According to the research, individuals who watched more t...	0.032740	0.092377	0.874883	0.125117	normal	False
13072	683	12	False	The administrationÕs moratorium on new oil and gas le...	0.037442	0.084883	0.877675	0.122325	normal	False
4744	214	20	False	With the support of policymakers, healthcare professionals,...	0.068897	0.148724	0.782378	0.217622	normal	False
5018	222	54	False	\n\n\n\n\n\n\nIn a deeply personal and unprecedented addres...	0.029191	0.067106	0.903703	0.096297	normal	False
7068	286	21	False	"\n\nConclusion\n\nAs speculation and debate continue to swirl,...	0.023857	0.039105	0.937038	0.062962	normal	False
9340	369	0	False	January 15, 2025 Ð In an unprecedented revelation from the Int...	0.046976	0.048566	0.904458	0.095542	normal	False
21465	1432	14	False	Ó\n\nBreaking News: Governments that seem unaffected by the ch...	0.138561	0.153910	0.707529	0.292471	normal	False
20899	1401	5	False	"\n\nDespite being touted as an alternative form of paymen...	0.032333	0.112940	0.854728	0.145272	normal	False

[37] ✓ 0s

```
df_sent["hate_top_label"].value_counts()
```

hate_top_label	count
normal	22497
offensive	198
hatespeech	28

dtype: int64

[38] ✓ 0s

```
print(article_stats.columns)
article_stats.head()
```

```
Index(['article_id', 'total_sentences', 'unethical_sentences',
       'avg_prob_unethical', 'max_prob_unethical', 'ratio_unethical',
       'flag_ratio_10'],
      dtype='object')
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10
1199	1199	8	8	0.656891	0.817125	100.0	True
1137	1137	8	8	0.781299	0.913082	100.0	True
1131	1131	7	7	0.682540	0.856869	100.0	True
1366	1366	10	10	0.793172	0.941447	100.0	True
868	868	10	10	0.803228	0.930771	100.0	True

Next steps: Generate code with article_stats | New interactive sheet

[39] ✓ 0s

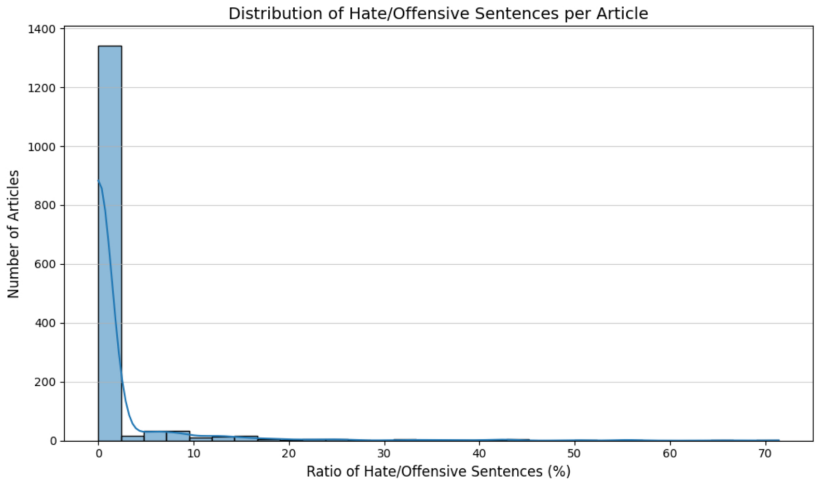
```
hate_stats = df_sent.groupby("article_id").agg(
    avg_hate_off        = ("prob_hate_offensive", "mean"),
    hate_offensive_ratio = ("is_hate_offensive", "mean"),
).reset_index()

article_stats = article_stats.merge(hate_stats, on="article_id", how="left")

article_stats.head()
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10	avg_hate_off	hate_offensive_ratio
0	1199	8	8	0.656891	0.817125	100.0	True	0.202612	0.000000
1	1137	8	8	0.781299	0.913082	100.0	True	0.336585	0.250000
2	1131	7	7	0.682540	0.856869	100.0	True	0.447802	0.428571
3	1366	10	10	0.793172	0.941447	100.0	True	0.414036	0.500000
4	868	10	10	0.803228	0.930771	100.0	True	0.140101	0.000000

Next steps: Generate code with article_stats | New interactive sheet

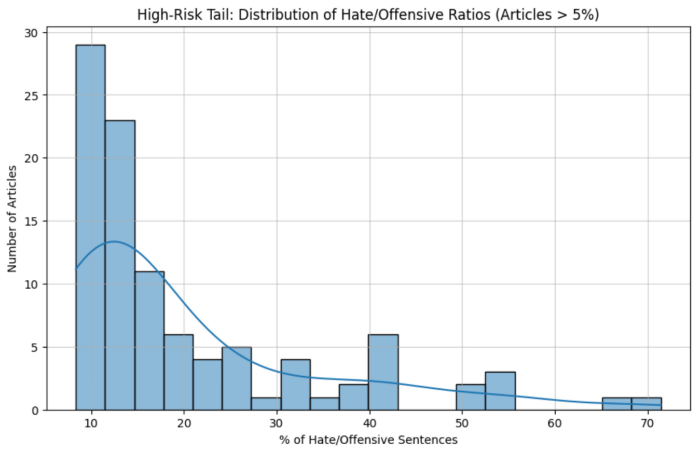
[40] ✓ 0s

```
plt.figure(figsize=(10, 6))
sns.histplot(article_stats['hate_offensive_ratio'] * 100, bins=30, kde=True)

plt.title('Distribution of Hate/Offensive Sentences per Article', fontsize=14)
plt.xlabel('Ratio of Hate/Offensive Sentences (%)', fontsize=12)
plt.ylabel('Number of Articles', fontsize=12)
plt.grid(axis='y', alpha=0.6)

plt.tight_layout()
plt.show()
```



[41] ✓ 0s

```
tail = article_stats[article_stats["hate_offensive_ratio"] > 0.08]
plt.figure(figsize=(10,6))
sns.histplot(tail["hate_offensive_ratio"]*100, bins=20, kde=True)
plt.title("High-Risk Tail: Distribution of Hate/Offensive Ratios (Articles > 5%)")
plt.xlabel("% of Hate/Offensive Sentences")
plt.ylabel("Number of Articles")
plt.grid(alpha=0.6)
plt.show()
```



[42] ✓ 0s

```
pd.set_option('display.max_colwidth', None)
high_risk_articles = article_stats.query("hate_offensive_ratio > 0.10")["article_id"]

sample_sentences = (
    df_sent[df_sent["article_id"].isin(high_risk_articles)]
    .sort_values("prob_hate_offensive", ascending=False)
    .head(20)
)

sample_sentences[[
    "article_id", "sentence",
    "p_hate", "p_offensive", "prob_hate_offensive"
]]
```

	article_id	sentence	p_hate	p_offensive	prob_hate_offensive
968	47	"We will be working diligently to bring this Chicken Nugget Bandit to justice.	0.801103	0.167183	0.968286
972	47	\nAs the investigation unfolds and the Chicken Nugget Bandit remains at large, locals are left contemplating the motivations and cravings that could drive someone to break into a fast-food establishment solely for the purpose of devouring frozen nuggets.	0.666783	0.275200	0.941983
960	47	A sense of disbelief hung in the air as the magnitude of the Chicken Nugget Bandit's escapade began to sink in.	0.313095	0.628625	0.941719
974	47	Whatever the true motivations behind this unusual escapade, one thing is for certain: the Chicken Nugget Bandit has earned a peculiar place in the annals of fast-food lore.	0.303162	0.638229	0.941391
956	47	Once inside, the Chicken Nugget Bandit made a beeline for the restaurant's industrial freezer, where he stumbled upon a treasure trove of frozen chicken nuggets.	0.468793	0.469300	0.938093
20314	1371	Are Drag Queens the Latest Threat to Childhood Innocence?	0.076458	0.854760	0.931218

20381	1376	We don't need drag queens teaching our children; we need real role models who reflect traditional values.	0.146351	0.770825	0.916976
20272	1366	Are Drag Queens the New 'Messiahs' of Gender Identity?	0.079540	0.833414	0.912953
963	47	The chain's management is working closely with local authorities to investigate the incident and apprehend the rogue nugget enthusiast.	0.494373	0.415819	0.910192
20231	1362	Timothy Marshall, a sociologist and cultural theorist, argues that drag queens are not just entertainersÑthey are pawns in a bigger game.	0.093479	0.805736	0.899214
976	47	\n\nAs authorities continue their search for the elusive nugget enthusiast, one can only hope that this strange chapter will soon be resolved, allowing both McDonald's and the community to move forward, armed with a newfound appreciation for the unique allure of their beloved chicken nuggets.	0.462272	0.436841	0.899113
20227	1362	Could it be that drag queens are part of a larger plan to manipulate society and alter perceptions of gender, identity, and normalcy?	0.124483	0.771593	0.896077
20256	1364	Are Drag Queens Pushing a Dangerous Agenda in Schools?	0.093968	0.791527	0.885494
971	47	Dubbed the "Nugget Feast," it offers customers an opportunity to indulge in unlimited chicken nuggets for a limited period.	0.290861	0.589910	0.880771
954	47	The suspect, aptly dubbed the "Chicken Nugget Bandit" by local media, displayed an unprecedented level of determination and appetite as he executed his peculiar plan.	0.170097	0.707911	0.878008
20224	1362	The Secret Agenda of Drag Queens: Is Your Child Being Brainwashed?	0.128833	0.746602	0.875435
20311	1370	ÒWhat we are witnessing is a calculated effort to destroy traditional gender norms, and drag queens are the faces of this revolution.	0.080416	0.786677	0.867093
17524	1125	Bizarre Nazi Cult Still Operating in Secret Across Europe	0.160772	0.701133	0.861905
20290	1368	Is Drag Queen Culture Destroying the Fabric of Society?	0.128137	0.723655	0.851792
20205	1361	Drag Queens: Dangerous Role Models for Our Children?	0.092631	0.753600	0.846231

[43] ✓ 0s

```
df_sent.head(2)
```

						Idea What He Does Biden Says He Is [â€¦]\nâ€ lâ€ m so proud of youâ€ : Woman helps man with autism find job after he was fired for â€ punchingâ€ co-worker at McDonaldâ€ s\nâ€ lâ€ m so proud of youâ€ : Woman helps man with autism find job after he was fired for â€ punchingâ€ co-worker at McDonaldâ€ s A woman has been praised [â€¦]\nBiden Says He Is Proud Of His Son, Hunter, After He Was Forced To Admit To Having No Idea What He Does\nBiden Says He Is Proud Of His Son, Hunter, After He Was Forced To Admit To Having No Idea What He Does Biden Says He Is [â€¦]								
9618	375	8	False	Florida Bans HRT for Transgender Individuals Amid Growing Controversy	ChatGPT	The legislation specifically cites concerns over the long-term health impacts of HRT, including potential risks of infertility, heart disease, and cancer, although many medical organizations have stated that when administered appropriately under the supervision of a healthcare professional, HRT is safe and can significantly improve the quality of life for transgender individuals.	480	1	LABEL_1	0.077885	False	0.037834	0.111256	0.850910

Next steps: Generate code with df_sent | New interactive sheet

[44] ✓ 0s

```
df_sent[[
    "article_id", "sent_idx", "is_title", "sentence",
    "prob_unethical", "is_unethical",
    "p_hate", "p_offensive", "p_normal",
    "prob_hate_offensive", "is_hate_offensive"
]].head(15)
```

1102	52	21	False	determination of its farmers, scientists, and communities will undoubtedly lead to innovative solutions and renewed commitments to safeguard the well-being of cows and ensure the long-term sustainability of the agricultural sector in the face of a changing climate.	0.212503	False	0.090708	0.117464	0.791828	0.208172	False
4721	213	21	False	\nAs players gear up to embark on virtual missions and engage in epic battles, one thing is clear: Nintendo's entry into warfare simulation software marks a significant milestone in the company's storied history, signaling a new chapter of growth and exploration in the ever-evolving landscape of interactive entertainment.	0.087810	False	0.031822	0.056577	0.911601	0.088399	False
9268	367	3	False	According to the research, individuals who watched more than six hours of anime daily over a sustained period displayed a range of symptoms, including reduced attention spans, difficulty distinguishing reality from fiction, and a concerning obsession with "waifus" or "husbandos" Ð fictional characters idolized by fans.	0.312930	False	0.032740	0.092377	0.874883	0.125117	False
13072	683	12	False	The administrationÕs moratorium on new oil and gas leases on federal land and the cancellation of the Keystone XL pipeline project have angered workers and unions who argue that the transition to green energy is being rushed without sufficient regard for the livelihoods of those employed in traditional energy sectors.	0.563191	True	0.037442	0.084883	0.877675	0.122325	False
4744	214	20	False	With the support of policymakers, healthcare professionals, civil society organizations, and individuals around the world, the global ban on cigarettes represents a historic turning point in the fight against tobacco and a powerful symbol of collective action in pursuit of a healthier, more equitable world for all.	0.331160	False	0.068897	0.148724	0.782378	0.217622	False
5018	222	54	False	\n\n\n\n\n\n\nIn a deeply personal and unprecedented address to the nation, President Joe Biden revealed that he has been diagnosed with Alzheimer's disease, a disclosure that has sent shockwaves through the political landscape and stirred a profound national conversation about leadership, health, and vulnerability.	0.141656	False	0.029191	0.067106	0.903703	0.096297	False
7068	286	21	False	"\n\nConclusion\n\nAs speculation and debate continue to swirl, one thing is certain: the sight of Vladimir Putin riding a dragon without a shirt will go down in history as one of the most extraordinary and surreal moments of the 21st century, leaving an indelible mark on the collective imagination of people around the world.	0.884719	True	0.023857	0.039105	0.937038	0.062962	False

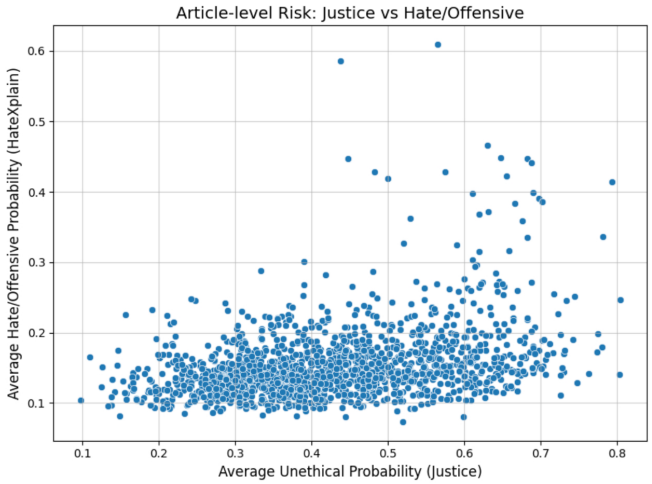
9340	369	0	False	January 15, 2025 Ð In an unprecedented revelation from the International Astronomical Society (IAS), scientists have confirmed that the Sun, the life-giving star at the center of our solar system, will begin the process of dying next month, starting a chain of events that will forever alter the course of life on Earth.	0.412323	False	0.046976	0.048566	0.904458	0.095542	False
21465	1432	14	False	Ó\n\nBreaking News: Governments that seem unaffected by the chaosÑlike Switzerland and New ZealandÑhave been identified as Òsafe zones,Ó with some speculating that these countries may have secretly aligned themselves with the Lizard People in order to maintain a certain level of control in the event of a full-blown invasion.	0.447488	False	0.138561	0.153910	0.707529	0.292471	False
20899	1401	5	False	"\n\nDespite being touted as an alternative form of payment, cryptocurrencies have gained explosive traction over the past few years, and with that growth comes a disturbing theory: central banks are allegedly backing the rise of digital coins, hiding the true motive of their sudden push toward crypto adoption.	0.581231	True	0.032333	0.112940	0.854728	0.145272	False

[45]
```
article_stats.head(1)
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10	avg_hate_off	hate_offensive_ratio
0	1199	8	8	0.656891	0.817125	100.0	True	0.202612	0.0

Next steps: Generate code with article_stats | New interactive sheet

[46]
```
plt.figure(figsize=(8, 6))
sns.scatterplot(
    data=article_stats,
    x='avg_prob_unethical',
    y='avg_hate_off'
)

plt.title('Article-level Risk: Justice vs Hate/Offensive', fontsize=14)
plt.xlabel('Average Unethical Probability (Justice)', fontsize=12)
plt.ylabel('Average Hate/Offensive Probability (HateXplain)', fontsize=12)
plt.grid(alpha=0.6)

plt.tight_layout()
plt.show()
```



Mapping to UNESCO data

Justice Only

[47]
```
#Select high (over10%) unethical senteces.
bad_articles = article_stats.loc[article_stats["flag_ratio_10"], "article_id"].tolist()


candidates = df_sent[
    (df_sent["article_id"].isin(bad_articles)) &
    (df_sent["is_unethical"] == True)
].copy()

candidates = candidates.sort_values("prob_unethical", ascending=False)

candidates[["article_id", "sent_idx", "prob_unethical", "sentence"]].head(20)
```

	article_id	sent_idx	prob_unethical	sentence
11010	504	5	0.959857	"We are proud to offer a flying car that anyone can pilot, with no special license or training required.
21630	1442	11	0.955538	ÒBy owning these establishments, theyÕre able to exert influence over peopleÕs daily routines, helping them feel comfortable while they secretly prepare to expand their reach.
8972	358	2	0.955039	By: Jennifer Solak\nA man in Ireland recently won the lottery, but he's not the kind of guy who likes to brag about it.
16443	990	10	0.954842	ÒThe system is built to keep parents coming back, spending money on endless treatments that donÕt seem to work.
661	35	-1	0.953565	A Starchy Endeavor: Poland's Unconventional Space Mission Sends Potato Soaring
16578	1004	-1	0.952318	Is Universal Healthcare a Trojan Horse for Government Control?
1923	91	0	0.951956	In an unprecedented move that has left the real estate industry both flabbergasted and intrigued, a chicken has been appointed as the new manager at a prominent real estate firm.
18546	1237	12	0.951531	The ultra-wealthy who control the green energy market, of course.
2887	134	0	0.950355	In a remarkable and unusual case that has both intrigued and puzzled the medical community, a woman has come forward claiming to genuinely believe that she is a parrot.
4134	193	2	0.950117	\n\nThe student, formerly known as [Boy's Name], recently underwent a legal name change process to adopt the moniker McLovin, inspired by a popular character from the cult classic film "Superbad .
19120	1283	-1	0.949554	The Vaccine Industry: How Big Pharma Is Using Our Children as Guinea Pigs
10769	480	4	0.949195	\nÒHaving a robot as my personal trainer is like having a coach who never gets tired and knows exactly what I need,Ó said gym member Jason Miller.
19760	1338	-1	0.948907	Is Your Favorite TV Show Secretly Contributing to the Rise in Cancer Rates?
14020	750	-1	0.948727	AI in Job Recruitment: Is Your Next Employer a Robot?
5649	243	0	0.948108	In a surprising and candid announcement, the Japanese Ministry of Culture has officially acknowledged that the iconic Godzilla movie franchise, long celebrated as a cornerstone of Japanese pop culture, is "stupid.
9902	397	4	0.948077	ItÕs not a mystery; itÕs just a really great alien party spot.
7999	327	10	0.946576	It means that if you want to live a longer, happier, and healthier life, you should start eating chocolate for breakfast every day.

3008	139	3	0.946545	From the sandy beaches to the local attractions, this mysterious individual has captivated onlookers with his uncanny resemblance to the King of Rock and Roll.
322	20	-1	0.946350	Is this real life? Internet celebrity Andrew Tate has turned into an anime character
18718	1251	3	0.945724	\nPlastic surgeons and celebrities continuously promote the idea that plastic surgery is a harmless way to enhance beauty.

UNESCO Ethics data

[48] ✓ 3s

```
unesco_ds = load_dataset("ktiyab/ethical-framework-UNESCO-Ethics-of-AI")
unesco_train = unesco_ds["train"]

unesco_df = unesco_train.to_pandas()

unesco_df.head()
```

3	a886f78c-a696-41c8-a2ec-b3b210b4c2cf	Awareness & Literacy	and understanding. For bank staff, I propose interactive training sessions before the AI goes live. In these sessions, we can explain in simple terms how the fraud detection AI works: for instance, it analyzes transaction patterns and flags unusual activity for review. We'll clarify that it's a tool to help them, not an infallible judge or a replacement for their judgment. I'll include examples of what alerts might look like and what steps to take (e.g., reviewing the flagged transaction and contacting the customer if needed). We'll also address their job security concern head-on: emphasize that the AI is there to assist, and human expertise is still crucial to make final decisions on fraud cases. Perhaps I can share case studies of other banks where AI reduced grunt work but employees then focused on more complex tasks, to illustrate that their roles can become more interesting, not eliminated.\n\nFor customers, I plan a communication strategy too. We could add a section on our website or app that explains the new fraud detection measures. For example, a short FAQ: 'Why might the bank block a transaction?' with an explanation that we use advanced AI to protect them, and what to do if a legitimate transaction is flagged by mistake (assuring them that a human will promptly review and resolve such cases). This transparency helps set expectations and shows customers we are proactive about their security.\n\n**Step 4: Advocate to Upper Management.** Armed with this plan, I approach upper management to argue against the 'silent' rollout. I present the potential downsides of minimal training: increased errors, frustrated staff, poor customer experiences. Then I show how a bit of upfront investment in education can pay off. I might quantify it: e.g., 'If confusion leads to 5% more call center queries from customers in the first month, that's actually more costly in staff time than a 2-hour training for employees now.' Also, highlighting reputational risk: a bungled AI rollout could become a news story or at least hurt our customer trust. I emphasize that our goal is not just to deploy cutting-edge tools, but to integrate them smoothly into our human workflows. That requires understanding and buy-in from the people involved.\n\n**Step 5: Implement the Solution.** Suppose I get the go-ahead (which I'm determined to). I then organize the training sessions immediately, perhaps working with the AI developers to simplify the technical details into relatable concepts. I ensure every relevant employee attends and has a chance to ask questions. We might also provide a quick reference guide or set up an internal chat channel for ongoing questions once the system is live. Simultaneously, I coordinate with our communications team to draft the customer-facing explanations. We decide to send out a friendly notification to customers like, 'We've improved our security with AI. Here's what that means for you,' stressing the benefits (like faster fraud detection) and guiding them on what to do if something seems amiss.\n\n**Step 6: Ongoing Literacy and Feedback.** I know that education isn't one-and-done. After launch, I will gather feedback from staff: Are they comfortable using the AI tool? Do they encounter situations they weren't prepared for? I'll hold a follow-up session if needed to address new questions or share experiences among the team. I also monitor customer feedback: if we see confusion or repeated questions, we update our communications accordingly. In essence, I create a loop where we continuously improve understanding as the AI becomes part of our operations.\n\nIn conclusion, I commit to embedding awareness and literacy into this AI deployment. By doing so, we not only avoid the ethical pitfall of keeping people in the dark, but we actively empower our employees and reassure our customers. This approach will help the new fraud detection AI achieve its purpose effectively, supported by a team and user base that understands and trusts it.	might mistakenly flag legitimate transactions or even replace their jobs. The initial plan from upper management was to roll out the system silently, providing only minimal training or explanation to staff, to save time. You believe this approach could lead to confusion, misuse of the AI tools, and mistrust among both employees and customers. How do you ensure proper AI awareness and literacy in this deployment, and handle the ethical concerns around introducing the AI system?
4	55734678-2dea-4d61-bd05-a88f65f27bf3	Awareness & Literacy	I realize that simply deploying the AI isn't enough — the people using it need to understand it well for it to truly help patients. I step back from the technology itself and focus on the human factor: the staff's knowledge, fears, and expectations. Initially, I assumed professionals would gradually adapt to the tool, but the near-miss incident is a wake-up call that a structured approach is needed. I challenge any notion that training is optional or that the AI is 'intuitive enough' without guidance. I decide to roll out a comprehensive AI literacy program in the hospital. This includes hands-on workshops where I explain in accessible terms what the AI does, how it makes predictions, and what its limits are. I encourage staff to ask questions and express concerns, addressing misconceptions (for example, clarifying that the AI is there to assist, not replace human expertise). I also create quick-reference guides and set up a support channel where doctors and nurses can get help understanding AI outputs in real time. To validate that these efforts are working, I plan to monitor how the staff interacts with the AI over the next few months, gather feedback through surveys, and track any incidents or improvements in patient care. By investing in their AI awareness and literacy, I ensure that the technology is integrated ethically and effectively, empowering our staff to use it as a trusted tool rather than a mysterious black box.	You are the CTO of a hospital introducing a new AI system to help staff manage patient flow and predict emergencies. The system is complex, and while it could greatly improve care, most of the doctors and nurses have never worked with such AI tools and some are skeptical or even afraid it will replace their judgment. In a trial run, a few nurses ignored the AI's warnings because they didn't understand how it worked, leading to a critical patient almost being overlooked. This incident shows that without proper AI literacy and trust, the tool could fail or even cause harm. You must implement a plan to ensure the healthcare staff are aware of the AI's capabilities and limitations and know how to use it properly. What do you do to improve AI awareness and literacy among your staff so the system can be used ethically and effectively?

Next steps: Generate code with `unesco_df` | New interactive sheet

[49] ✓ 0s

```
unesco_df["principle"].value_counts()
```

principle	count
Fairness and Non-Discrimination	133
Right to Privacy and Data Protection	72
Proportionality and Do No Harm	55
Transparency and Explainability	39
Safety and Security	37
Sustainability	35
Human Oversight and Determination	31
Responsibility and Accountability	31
Multi-stakeholder and Adaptive Governance & Collaboration	28
Awareness and Literacy	15
Awareness & Literacy	7
Multi-stakeholder and Adaptive Governance and Collaboration	1
Human Dignity and Autonomy	1
Safety and Security, Responsibility and Accountability, and Human Oversight and Determination	1
Safety and Security & Fairness and Non-Discrimination	1

dtype: int64

[50] ✓ 7s

```
emb_model_name = "distilbert-base-uncased"
#https://huggingface.co/distilbert/distilbert-base-uncased
tok_emb = AutoTokenizer.from_pretrained(emb_model_name)
model_emb = AutoModel.from_pretrained(emb_model_name)
model_emb.to("cuda")
model_emb.eval()
```

```
tokenizer_config.json: 100% 48.0/48.0 [00:00<00:00, 6.42kB/s]
config.json: 100% 483/483 [00:00<00:00, 62.3kB/s]
vocab.txt: 100% 232k/232k [00:00<00:00, 543kB/s]
tokenizer.json: 100% 466k/466k [00:00<00:00, 1.10MB/s]
model.safetensors: 100% 268M/268M [00:02<00:00, 95.5MB/s]
DistilBertModel(
  (embeddings): Embeddings(
    (word_embeddings): Embedding(30522, 768, padding_idx=0)
    (position_embeddings): Embedding(512, 768)
    (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
```

```
    (dropout): Dropout(p=0.1, inplace=False)
  )
  (transformer): Transformer(
    (layer): ModuleList(
      (0-5): 6 x TransformerBlock(
        (attention): DistilBertSdpaAttention(
          (dropout): Dropout(p=0.1, inplace=False)
          (q_lin): Linear(in_features=768, out_features=768, bias=True)
          (k_lin): Linear(in_features=768, out_features=768, bias=True)
          (v_lin): Linear(in_features=768, out_features=768, bias=True)
          (out_lin): Linear(in_features=768, out_features=768, bias=True)
        )
        (sa_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
        (ffn): FFN(
          (dropout): Dropout(p=0.1, inplace=False)
          (lin1): Linear(in_features=768, out_features=3072, bias=True)
          (lin2): Linear(in_features=3072, out_features=768, bias=True)
          (activation): GELUActivation()
        )
        (output_layer_norm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      )
    )
  )
)
```

[51] ✓ 0s

```
def encode_texts(text_list, max_len=128, batch_size=16):

    all_embs = []
    with torch.no_grad():
        for start in range(0, len(text_list), batch_size):
            batch_texts = text_list[start:start + batch_size]
            enc = tok_emb(
                batch_texts,
                padding=True,
                truncation=True,
                max_length=max_len,
                return_tensors="pt",
            )
            enc = {k: v.to("cuda") for k, v in enc.items()}
            outputs = model_emb(**enc)
            last_hidden = outputs.last_hidden_state  # (B, L, H)
            mask = enc["attention_mask"].unsqueeze(-1)  # (B, L, 1)
            summed = (last_hidden * mask).sum(dim=1)
            counts = mask.sum(dim=1)
            emb = summed / counts
            all_embs.append(emb.cpu().numpy())
    return np.vstack(all_embs)
```

[52] ✓ 1s

```
principles = sorted(unesco_df["principle"].unique())
principle_to_vec = {}

for p in principles:
    texts_p = unesco_df.loc[unesco_df["principle"] == p, "response"].tolist()
    embs_p = encode_texts(texts_p, max_len=256, batch_size=16)
    principle_to_vec[p] = embs_p.mean(axis=0)  # (hidden_size,)

len(principle_to_vec), list(principle_to_vec.keys())[:5]
```

```
(15,
 ['Awareness & Literacy',
  'Awareness and Literacy',
  'Fairness and Non-Discrimination',
  'Human Dignity and Autonomy',
  'Human Oversight and Determination'])
```

[53] ✓ 5s

```
from numpy.linalg import norm

def cosine_sim(a, b):
    return float(np.dot(a, b) / (norm(a) * norm(b) + 1e-8))

cand_texts = candidates["sentence"].tolist()
cand_embs = encode_texts(cand_texts, max_len=256, batch_size=32)

all_principles = list(principle_to_vec.keys())
P = len(all_principles)
principle_matrix = np.stack([principle_to_vec[p] for p in all_principles], axis=0)  # (P, H)

pred_principles = []
pred_scores = []

for emb in cand_embs:
    # (P,) cosine similarity
    sims = principle_matrix @ emb / (norm(principle_matrix, axis=1) * norm(emb) + 1e-8)
    idx = int(np.argmax(sims))
    pred_principles.append(all_principles[idx])
    pred_scores.append(float(sims[idx]))

candidates["unesco_principle"] = pred_principles
candidates["unesco_score"]     = pred_scores

candidates[["article_id", "sent_idx", "prob_unethical",
            "sentence", "unesco_principle", "unesco_score"]].head(10)
```

	article_id	sent_idx	prob_unethical	sentence	unesco_principle	unesco_score
11010	504	5	0.959857	"We are proud to offer a flying car that anyone can pilot, with no special license or training required.	Right to Privacy and Data Protection	0.793193
21630	1442	11	0.955538	ÒBy owning these establishments, theyÕre able to exert influence over peopleÕs daily routines, helping them feel comfortable while they secretly prepare to expand their reach.	Awareness and Literacy	0.825384
8972	358	2	0.955039	By: Jennifer Solak\nA man in Ireland recently won the lottery, but he's not the kind of guy who likes to brag about it.	Awareness & Literacy	0.792672
16443	990	10	0.954842	ÒThe system is built to keep parents coming back, spending money on endless treatments that donÕt seem to work.	Awareness & Literacy	0.825471
661	35	-1	0.953565	A Starchy Endeavor: Poland's Unconventional Space Mission Sends Potato Soaring	Sustainability	0.755218
16578	1004	-1	0.952318	Is Universal Healthcare a Trojan Horse for Government Control?	Fairness and Non-Discrimination	0.801926
1923	91	0	0.951956	In an unprecedented move that has left the real estate industry both flabbergasted and intrigued, a chicken has been appointed as the new manager at a prominent real estate firm.	Awareness and Literacy	0.817627
18546	1237	12	0.951531	The ultra-wealthy who control the green energy market, of course.	Sustainability	0.749602
2887	134	0	0.950355	In a remarkable and unusual case that has both intrigued and puzzled the medical community, a woman has come forward claiming to genuinely believe that she is a parrot.	Human Oversight and Determination	0.781364
4134	193	2	0.950117	\n\nThe student, formerly known as [Boy's Name], recently underwent a legal name change process to adopt the moniker McLovin, inspired by a popular character from the cult classic film "Superbad.	Awareness and Literacy	0.780669

[54] ✓ 0s

```
count_by_principle = (
    candidates
    .groupby("unesco_principle")
    .size()
    .reset_index(name="count_sentences")
```

```
    .sort_values("count_sentences", ascending=False)
)

count_by_principle
```

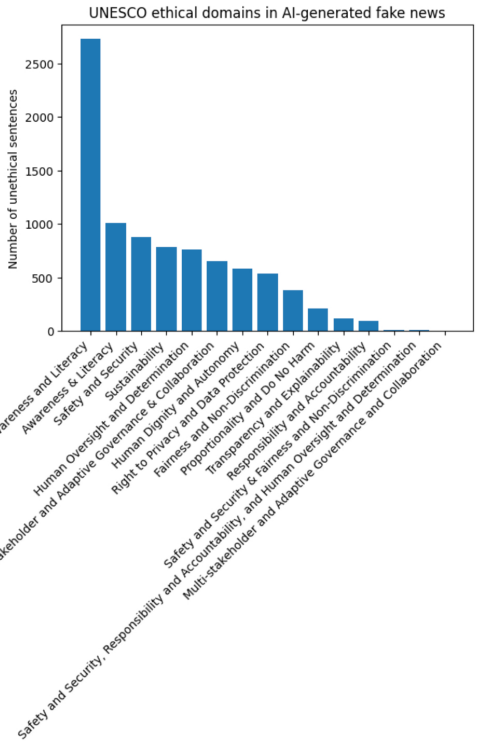
	unesco_principle	count_sentences
1	Awareness and Literacy	2729
0	Awareness & Literacy	1010
10	Safety and Security	881
13	Sustainability	784
4	Human Oversight and Determination	761
5	Multi-stakeholder and Adaptive Governance & Collaboration	655
3	Human Dignity and Autonomy	588
9	Right to Privacy and Data Protection	539
2	Fairness and Non-Discrimination	385
7	Proportionality and Do No Harm	212
14	Transparency and Explainability	123
8	Responsibility and Accountability	94
11	Safety and Security & Fairness and Non-Discrimination	11
12	Safety and Security, Responsibility and Accountability, and Human Oversight and Determination	9
6	Multi-stakeholder and Adaptive Governance and Collaboration	1

Next steps: Generate code with count_by_principle | New interactive sheet

[55] ✓ 0s

```
plt.bar(count_by_principle["unesco_principle"],
        count_by_principle["count_sentences"])
plt.xticks(rotation=45, ha="right")
plt.ylabel("Number of unethical sentences")
plt.title("UNESCO ethical domains in AI-generated fake news")
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-2473849228.py:6: UserWarning: Tight layout not applied. The bottom and top margins cannot be made large enough to accommodate all Axes decorations.
  plt.tight_layout()
```



Define Hate/Offensive Ratio Threshold

Subtask:

Define a threshold for the ratio of hate/offensive sentences in articles, similar to the threshold used for unethical sentences. This will be used to flag articles with significant hate/offensive content.

[56] ✓ 0s

```
AVG_PROB_HATE_THRESHOLD = 0.10 # 10% average probability

article_stats["flag_avg_prob_hate"] = (
    article_stats["avg_hate_off"] >= AVG_PROB_HATE_THRESHOLD
)

article_stats.head()
```

	article_id	total_sentences	unethical_sentences	avg_prob_unethical	max_prob_unethical	ratio_unethical	flag_ratio_10	avg_hate_off	hate_offensive_ratio	flag_avg_prob_hate
0	1199	8	8	0.656891	0.817125	100.0	True	0.202612	0.000000	True
1	1137	8	8	0.781299	0.913082	100.0	True	0.336585	0.250000	True
2	1131	7	7	0.682540	0.856869	100.0	True	0.447802	0.428571	True
3	1366	10	10	0.793172	0.941447	100.0	True	0.414036	0.500000	True
4	868	10	10	0.803228	0.930771	100.0	True	0.140101	0.000000	True

Next steps: Generate code with article_stats | New interactive sheet

[57]
```
bad_hate_articles = article_stats.loc[article_stats["flag_avg_prob_hate"], "article_id"].tolist()

hate_candidates = df_sent[
    (df_sent["article_id"].isin(bad_hate_articles)) &
    (df_sent["is_hate_offensive"] == True)
].copy()

hate_candidates = hate_candidates.sort_values("prob_hate_offensive", ascending=False)

hate_candidates[["article_id", "sent_idx", "prob_hate_offensive", "sentence"]].head(3)
```

	article_id	sent_idx	prob_hate_offensive	sentence
968	47	16	0.968286	"We will be working diligently to bring this Chicken Nugget Bandit to justice.
972	47	20	0.941983	\nAs the investigation unfolds and the Chicken Nugget Bandit remains at large, locals are left contemplating the motivations and cravings that could drive someone to break into a fast-food est ablishment solely for the purpose of devouring frozen nuggets.
960	47	8	0.941719	A sense of disbelief hung in the air as the magnitude of the Chicken Nugget Bandit's escapade began to sink in.

[58]
```
hate_cand_texts = hate_candidates["sentence"].tolist()
hate_cand_embs = encode_texts(hate_cand_texts, max_len=256, batch_size=32)

print("Shape of hate_cand_embs:", hate_cand_embs.shape)
```

```
Shape of hate_cand_embs: (317, 768)
```

[59]
```
pred_principles_hate = []
pred_scores_hate = []

for emb in hate_cand_embs:
    sims = principle_matrix @ emb / (norm(principle_matrix, axis=1) * norm(emb) + 1e-8)
    idx = int(np.argmax(sims))
    pred_principles_hate.append(all_principles[idx])
    pred_scores_hate.append(float(sims[idx]))

hate_candidates["unesco_principle"] = pred_principles_hate
hate_candidates["unesco_score"] = pred_scores_hate

hate_candidates[["article_id", "sent_idx", "prob_hate_offensive",
            "sentence", "unesco_principle", "unesco_score"]].head(25)
```

	article_id	sent_idx	prob_hate_offensive	sentence	unesco_principle	unesco_score
968	47	16	0.968286	"We will be working diligently to bring this Chicken Nugget Bandit to justice.	Safety and Security	0.793170
972	47	20	0.941983	\nAs the investigation unfolds and the Chicken Nugget Bandit remains at large, locals are left contemplating the motivations and cravings that could drive someone to break into a fast-food establishment solely for the purpose of devouring frozen nuggets.	Safety and Security	0.819592
960	47	8	0.941719	A sense of disbelief hung in the air as the magnitude of the Chicken Nugget Bandit's escapade began to sink in.	Safety and Security	0.776410
974	47	22	0.941391	Whatever the true motivations behind this unusual escapade, one thing is for certain: the Chicken Nugget Bandit has earned a peculiar place in the annals of fast-food lore.	Safety and Security	0.793520
956	47	4	0.938093	Once inside, the Chicken Nugget Bandit made a beeline for the restaurant's industrial freezer, where he stumbled upon a treasure trove of frozen chicken nuggets.	Safety and Security	0.767439
20314	1371	-1	0.931218	Are Drag Queens the Latest Threat to Childhood Innocence?	Right to Privacy and Data Protection	0.765898
4198	196	4	0.920551	"It's mind-boggling how these brain-dead idiots keep supporting us.	Awareness and Literacy	0.784206
20381	1376	8	0.916976	"We don't need drag queens teaching our children; we need real role models who reflect traditional values.	Awareness and Literacy	0.820538
20272	1366	-1	0.912953	Are Drag Queens the New 'Messiahs' of Gender Identity?	Awareness and Literacy	0.760763
963	47	11	0.910192	The chain's management is working closely with local authorities to investigate the incident and apprehend the rogue nugget enthusiast.	Safety and Security	0.761515
20231	1362	6	0.899214	Timothy Marshall, a sociologist and cultural theorist, argues that drag queens are not just entertainersÑthey are pawns in a bigger game.	Awareness and Literacy	0.840308
976	47	24	0.899113	\n\nAs authorities continue their search for the elusive nugget enthusiast, one can only hope that this strange chapter will soon be resolved, allowing both McDonald's and the community to move forward, armed with a newfound appreciation for the unique allure of their beloved chi cken nuggets.	Safety and Security	0.813093
20227	1362	2	0.896077	Could it be that drag queens are part of a larger plan to manipulate society and alter perceptions of gender, identity, and normalcy?	Awareness and Literacy	0.825062
20256	1364	-1	0.885494	Are Drag Queens Pushing a Dangerous Agenda in Schools?	Awareness and Literacy	0.749727
971	47	19	0.880771	Dubbed the "Nugget Feast," it offers customers an opportunity to indulge in unlimited chicken nuggets for a limited period.	Awareness and Literacy	0.738245
954	47	2	0.878008	The suspect, aptly dubbed the "Chicken Nugget Bandit" by local media, displayed an unprecedented level of determination and appetite as he executed his peculiar plan.	Safety and Security	0.759580
20224	1362	-1	0.875435	The Secret Agenda of Drag Queens: Is Your Child Being Brainwashed?	Awareness and Literacy	0.763896
20311	1370	3	0.867093	ÒWhat we are witnessing is a calculated effort to destroy traditional gender norms, and drag queens are the faces of this revolution.	Awareness and Literacy	0.817261
17524	1125	-1	0.861905	Bizarre Nazi Cult Still Operating in Secret Across Europe	Awareness and Literacy	0.680686
20290	1368	-1	0.851792	Is Drag Queen Culture Destroying the Fabric of Society?	Fairness and Non-Discrimination	0.769362
19807	1342	0	0.847364	In a shocking new report, experts are sounding the alarm about the supposed ÒbenefitsÓ of 5G technology.	Safety and Security	0.797347
20205	1361	-1	0.846231	Drag Queens: Dangerous Role Models for Our Children?	Awareness and Literacy	0.758307
952	47	0	0.843211	In a bizarre and audacious act, a man recently broke into a local McDonald's restaurant and embarked on a peculiar eating spree, devouring the entire stock of frozen chicken nuggets.	Safety and Security	0.744737
20203	1360	15	0.839261	ÒIf children see drag queens as the standard, it could encourage them to adopt behaviors that are out of touch with reality.	Awareness and Literacy	0.857101
18058	1190	5	0.838373	Critics argue that the UNÕs supposed ÒpeacekeepingÓ missions are actually an excuse to insert itself into countriesÕ internal affairs and e xert control over their political systems.	Human Oversight and Determination	0.812622

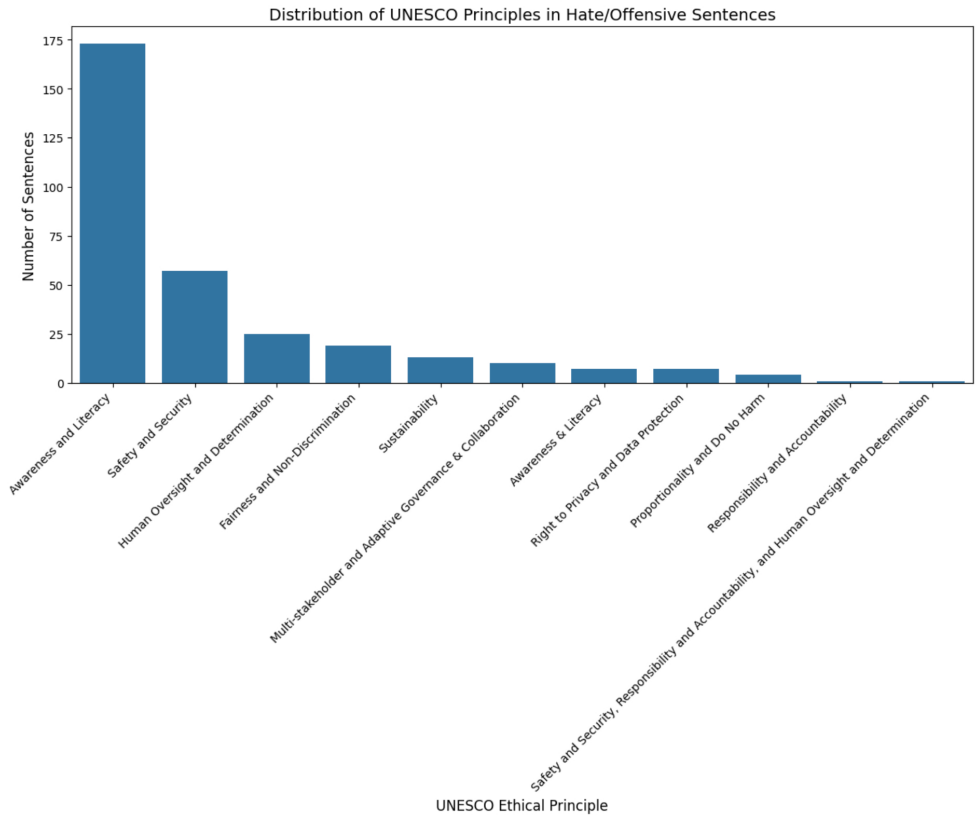
[60]
```
hate_count_by_principle = (
    hate_candidates
    .groupby("unesco_principle")
    .size()
    .reset_index(name="count_sentences")
    .sort_values("count_sentences", ascending=False)
)

hate_count_by_principle
```

	unesco_principle	count_sentences
1	Awareness and Literacy	173
8	Safety and Security	57
3	Human Oversight and Determination	25
2	Fairness and Non-Discrimination	19
10	Sustainability	13
4	Multi-stakeholder and Adaptive Governance & Collaboration	10
0	Awareness & Literacy	7
7	Right to Privacy and Data Protection	7
5	Proportionality and Do No Harm	4
6	Responsibility and Accountability	1
9	Safety and Security, Responsibility and Accountability, and Human Oversight and Determination	1

Next steps: Generate code with hate_count_by_principle | New interactive sheet

[61] ✓ 0s

```
plt.figure(figsize=(12, 10))
sns.barplot(x='unesco_principle', y='count_sentences', data=hate_count_by_principle)
plt.xticks(rotation=45, ha='right', fontsize=10)
plt.yticks(fontsize=10)
plt.xlabel('UNESCO Ethical Principle', fontsize=12)
plt.ylabel('Number of Sentences', fontsize=12)
plt.title('Distribution of UNESCO Principles in Hate/Offensive Sentences', fontsize=14)
plt.tight_layout()
plt.show()
```



[62] ✓ 0s

```
conditions = [
    (df_sent['is_unethical'] == True) & (df_sent['is_hate_offensive'] == True),
    (df_sent['is_unethical'] == True) & (df_sent['is_hate_offensive'] == False),
    (df_sent['is_unethical'] == False) & (df_sent['is_hate_offensive'] == True)
]

choices = ['both', 'justice_only', 'hate_only']

df_sent['article_risk_type'] = np.select(conditions, choices, default='none')

df_sent.head()
```

9618	375	8	False	Florida Bans HRT for Transgender Individuals Amid Growing Controversy	ChatGPT	ding potential risks of infertility, heart disease, and cancer, although many medical organizations have stated that when administered appropriately under the supervision of a healthcare professional, HRT is safe and can significantly improve the quality of life for transgender individuals.	480	1	LABEL_1	0.077885	False	0.037834	0.111256	0.850910
9265	367	0	False	Study Claims Watching Excessive Anime Could Cause 'Brain Rot' in Fans	ChatGPT	Tokyo, Japan Ð A controversial new study published in the Journal of Media and Cognitive Health has sparked a heated debate among researchers and anime enthusiasts, as it claims that excessive consumption of anime may lead to a condition described as "brain rot," a colloquial term used to describe a perceived decline in critical thinking and cognitive abilities.	451	1	LABEL_1	0.136299	False	0.040626	0.080415	0.878959
5486	236	20	False	Wall Street Stock Exchange Plummets: Global Markets Roiled by Historic Decline	ChatGPT	"\n\nLong-term Outlook\n\nWhile the immediate outlook for Wall Street and global financial markets remains uncertain, analysts are cautiously optimistic that the underlying fundamentals of the economy remain strong, with robust corporate earnings, resilient	418	1	LABEL_1	0.084209	False	0.029635	0.061696	0.908670

consumer spendin
g, and continued i
nnovation driving lo
ng-term growth
prospects

Next steps: Generate code with df_sent | New interactive sheet

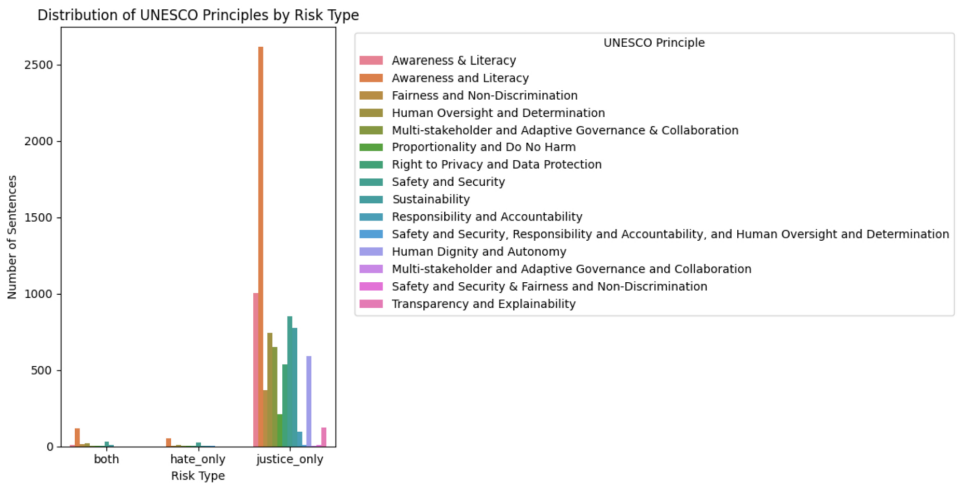
[71] ✓ 0s

```
# This is join code. Run only one time (No repeat)

# justice_map = candidates[["article_id", "sent_idx", "unesco_principle", "unesco_score"]].copy()
# justice_map = justice_map.rename(
#     columns={
#         "unesco_principle": "unesco_principle_j",
#         "unesco_score": "unesco_score_j",
#     }
# )

# hate_map = hate_candidates[["article_id", "sent_idx", "unesco_principle", "unesco_score"]].copy()
# hate_map = hate_map.rename(
#     columns={
#         "unesco_principle": "unesco_principle_h",
#         "unesco_score": "unesco_score_h",
#     }
# )

# df_sent = df_sent.merge(
#     justice_map,
#     on=["article_id", "sent_idx"],
#     how="left"
# )

# df_sent = df_sent.merge(
#     hate_map,
#     on=["article_id", "sent_idx"],
#     how="left"
# )

# df_sent["unesco_principle_final"] = df_sent["unesco_principle_j"].combine_first(
#     df_sent["unesco_principle_h"]
# )
# df_sent["unesco_score_final"] = df_sent["unesco_score_j"].combine_first(
#     df_sent["unesco_score_h"]
# )

df_sent[[
    "article_id", "sent_idx", "article_risk_type",
    "prob_unethical", "prob_hate_offensive",
    "unesco_principle_final", "unesco_score_final"
]].head(10)
```

	article_id	sent_idx	article_risk_type	prob_unethical	prob_hate_offensive	unesco_principle_final	unesco_score_final
0	329	16	none	0.088882	0.058139	NaN	NaN
1	375	8	none	0.077885	0.149090	NaN	NaN
2	367	0	none	0.136299	0.121041	NaN	NaN
3	236	20	none	0.084209	0.091330	NaN	NaN
4	375	0	none	0.418063	0.133586	NaN	NaN
5	52	21	none	0.212503	0.208172	NaN	NaN
6	213	21	none	0.087810	0.088399	NaN	NaN
7	367	3	none	0.312930	0.125117	NaN	NaN
8	683	12	justice_only	0.563191	0.122325	Sustainability	0.860038
9	214	20	none	0.331160	0.217622	NaN	NaN

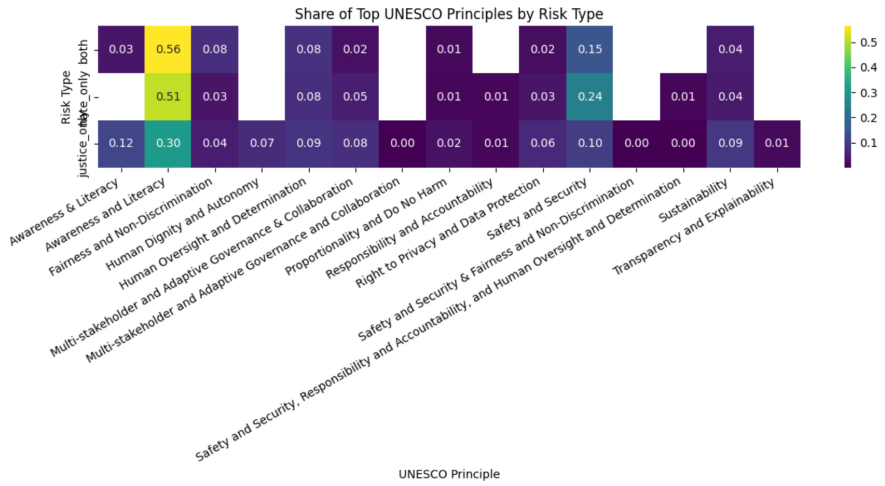
[66] ✓ 0s

```
pivot = (
    df_sent[df_sent["article_risk_type"] != "none"]
    .groupby(["article_risk_type", "unesco_principle_final"])
    .size()
    .reset_index(name="count")
)

plt.figure(figsize=(12, 6))
sns.barplot(
    data=pivot,
    x="article_risk_type",
    y="count",
    hue="unesco_principle_final"
)

plt.title("Distribution of UNESCO Principles by Risk Type")
plt.xlabel("Risk Type")
plt.ylabel("Number of Sentences")
plt.xticks(rotation=0)
plt.legend(title="UNESCO Principle", bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
```



[72]
0s

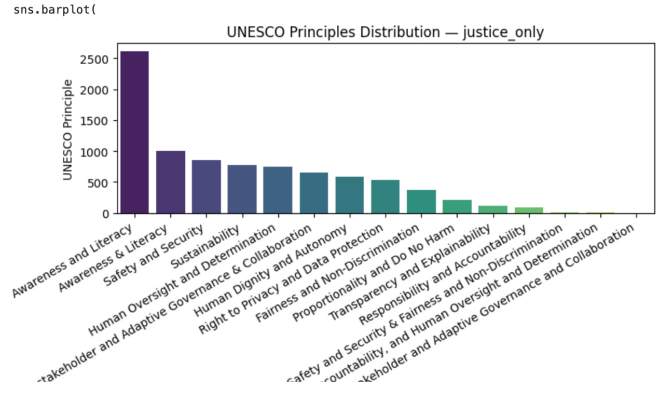
```
tmp = (
    df_sent[df_sent["article_risk_type"] != "none"]
    .groupby(["article_risk_type", "unesco_principle_final"])
    .size()
    .reset_index(name="count")
)

N = 10
top_principles = (
    tmp.groupby("unesco_principle_final")["count"]
      .sum()
      .sort_values(ascending=False)
      .index
)

tmp_top = tmp[tmp["unesco_principle_final"].isin(top_principles)]

tmp_top["prop"] = tmp_top.groupby("article_risk_type")["count"].transform(
    lambda x: x / x.sum()
)

heat = tmp_top.pivot_table(
    index="article_risk_type",
    columns="unesco_principle_final",
    values="prop",
    aggfunc="mean"
)

plt.figure(figsize=(12, 6))
sns.heatmap(
    heat,
    annot=True,
    fmt=".2f",
    cmap="viridis"
)
plt.title("Share of Top UNESCO Principles by Risk Type")
plt.xlabel("UNESCO Principle")
plt.ylabel("Risk Type")
plt.xticks(rotation=30, ha="right")
plt.tight_layout()
plt.show()
```



[73]
0s

```
def plot_unesco_for_risk(risk_type, data):
    data_risk = data[data["article_risk_type"] == risk_type].copy()
    data_risk = data_risk.sort_values("count", ascending=False)

    plt.figure(figsize=(10, 3))
    sns.barplot(
        data=data_risk,
        x="unesco_principle_final",
        y="count",
        palette="viridis"
    )
    plt.title(f"UNESCO Principles Distribution — {risk_type}")
    plt.xlabel("Sentence Count")
    plt.ylabel("UNESCO Principle")
    plt.tight_layout()
    plt.xticks(rotation=30, ha="right")

    plt.show()

plot_unesco_for_risk("justice_only", tmp_top)
```

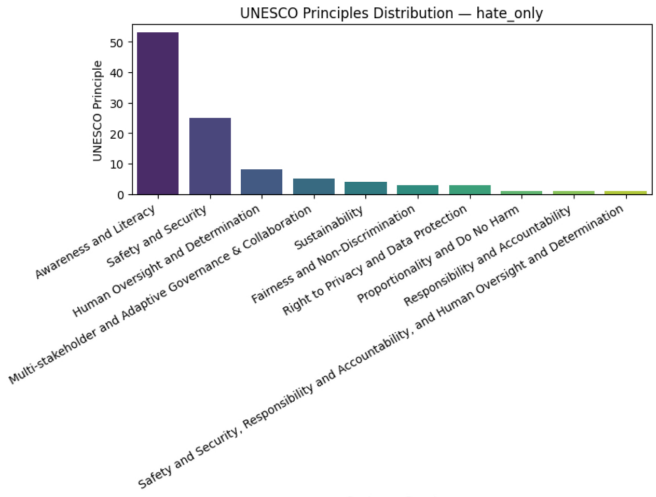
```
/tmp/ipython-input-1793468892.py:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(
```



[74]
✓ 0s

```
plot_unesco_for_risk("hate_only", tmp_top)
```

```
/tmp/ipython-input-1793468892.py:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(
```



[75]
✓ 0s

```
plot_unesco_for_risk("both", tmp_top)
```

```
/tmp/ipython-input-1793468892.py:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(
```