TODO

- Train a specific BERT before using it
- Add a step to use an LLM (probably LLnan3 locally) to add features, labels etc

In [1]:

```
import pandas as pd
import numpy as np
import os
import glob
from helper_functions.topic_modelling.flatten_articles import flatten_articl
from helper_functions.topic_modelling.sentiment_analysis import perform_sent
from helper_functions.topic_modelling.add_counts_columns_parallel import (
    add_synonym_frequency_columns,
    add_category_count_columns
)
from helper_functions.topic_modelling.aggregate_articles import aggregate_ar

# 1️⃣ Load enriched event dataset with:
#    - articles (list of cleaned + truncated text strings)
#    - NER_admin0_list / admin1 / admin2
#    - event metadata (ADMIN0/1/2, CS_score, period, etc.)
# =================================================================

exploded_df = pd.read_parquet("../data/gdelt/events/scraped_urls/cleaned_fil

# =================================================================
# 3️⃣ (OPTIONAL SAFETY) Ensure clean_text is string
# =================================================================
exploded_df["clean_text"] = exploded_df["clean_text"].astype(str)

# =================================================================
# 4️⃣ Sentiment Analysis
# =================================================================
exploded_df = perform_sentiment_analysis(exploded_df, text_col="clean_text")
print("Sentiment analysis done.")

# =================================================================
# 5️⃣ Keyword frequency & category count features
# =================================================================
exploded_df = add_synonym_frequency_columns(exploded_df, text_col="clean_tex
print("Frequency counts done")
exploded_df = add_category_count_columns(exploded_df, text_col="clean_text")
print("Keyword & category counts added.")

# =================================================================
# 6️⃣ Save in chunks
# =================================================================
out_dir = "../data/gdelt/events/5_modelled"
os.makedirs(out_dir, exist_ok=True)
exploded_df.to_parquet(out_dir + "/events_exploded_with_counts.parquet", inc
print("✅ Processing complete.")
exploded_df.head()
```

```
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
Processing chunks: 100%|██████████| 7/7 [01:29<00:00, 12.83s/it]
```

```
Sentiment analysis done.
Pandas Apply:   0%|          | 0/131727 [00:00<?, ?it/s]
Frequency counts done
  0%|          | 0/131727 [00:00<?, ?it/s]
```

```
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
/Users/marco.bertetti/Desktop/git_repos/phd_nlp/venv/lib/python3.9/site-pack
ages/urllib3/__init__.py:35: NotOpenSSLWarning: urllib3 v2 only supports Ope
nSSL 1.1.1+, currently the 'ssl' module is compiled with 'LibreSSL 2.8.3'. S
ee: https://github.com/urllib3/urllib3/issues/3020
  warnings.warn(
```

```
Keyword & category counts added.
✅ Processing complete.
```

Out[1]:

	url	header	body	
0	http://truefire.com/blog/guitar-lessons/8-free...	8 Free African Style Guitar Lessons - TrueFire...	Paul Simon, Ry Cooder, Taj Mahal, Corey Harris...	http://truefire
1	http://truepublica.org.uk/global/european-use-...	European Use of Military Drones Expanding - Tr...	ByChris Coleofdronewars.uk– Two weeks ago a ne...	http://truepub
2	http://truepublica.org.uk/united-kingdom/us-sp...	US Special Forces Will Wage War in Africa “For...	By Thomas Gaist and Eddie Haywoodwsws.org– The...	http
3	http://truepundit.com/boko-haram-expected-to-q...	Boko Haram Expected To Quadruple Use Of Girls ...	Boko Haram, the Islamic State’s affiliate in ...	http://tr
4	http://truepundit.com/fact-check-yes-terrorist...	FACT CHECK: Yes, Terrorists Have Come From The...	“The various people who have, in fact, commit...	http://true

5 rows × 34 columns