News Media Analytics for Risk Monitoring:

Chungmann Kim

17/09/2025

IPC global partners

Why Add News to Risk Monitoring?

- Traditional indicators (prices, conflict, weather, displacement) are robust and increasingly near real-time, but often lack narrative context
- News media can provide **rich, real-time narratives** on emerging risks (conflict, political instability, displacement, aid blockages, economic shocks).
- Using NLP and LLMs, news can **complement structured evidence** by adding depth and situational awareness to IPC's risk monitoring (Wanrooij et al, 2024; Balashankar et al, 2023).

Objective: From News to Humanitarian Signals

- Build an **NLP/LLM pipeline** to scrape and process news articles
- **Define potential use cases** of processed news data for humanitarian risk monitoring
- Demonstrate practical applications of media-based risk indicators, including counts, sentiment, and contextual summaries

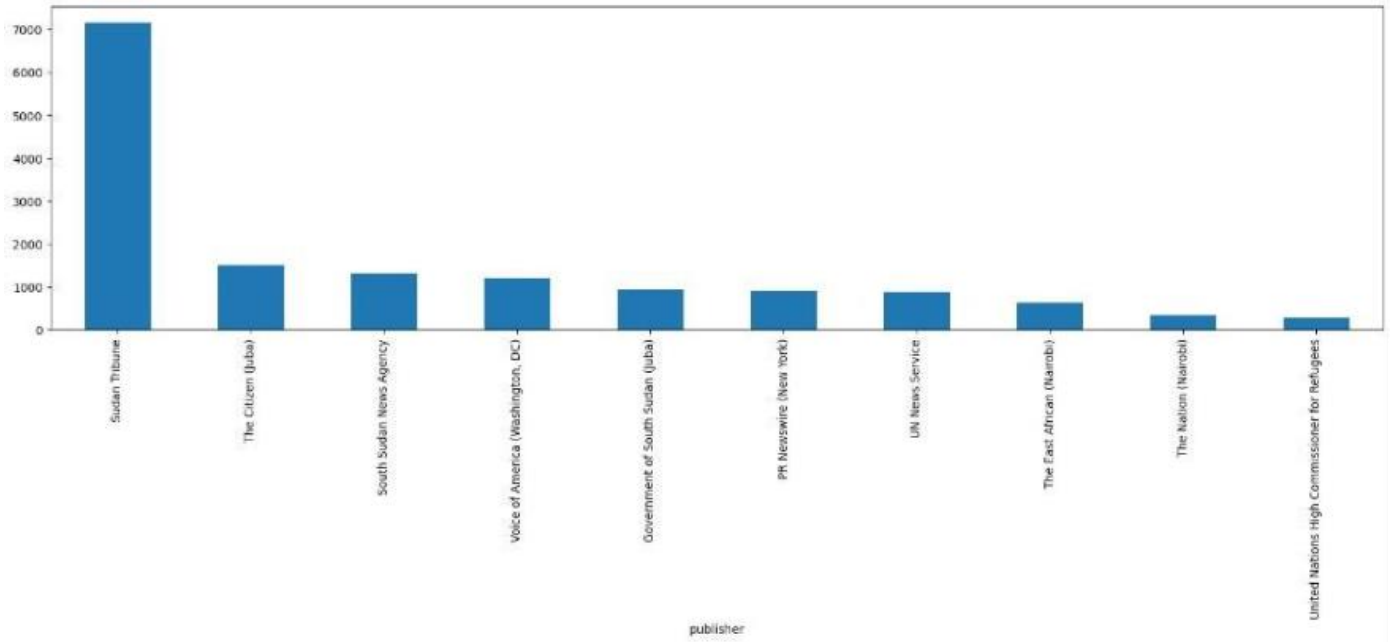
Quick Overview

- **Articles collected**: 18,000+ (2011–2023) from AllAfrica, focusing on South Sudan
- **Geocoded**: 11,000+ articles mapped to ADM1/ADM2 levels using Named Entity Recognition (NER) and geoparsing
- **Topic classification**: Applied topic modeling and LLM-based keyword reduction to classify articles into 11 humanitarian themes (e.g., Conflict, Humanitarian Aid, Food Crisis)
- **Sentiment analysis**: Performed sentiment scoring at regional and thematic levels
- **Validation**: Correlated news-derived indicators with IPC outcomes, food prices, displacement, and conflict data
- **Demonstrated use cases**:
  - Article counts by topic and region
  - Sentiment score trends over time
  - LLM-powered contextual summaries of key events

Raw Data Profile

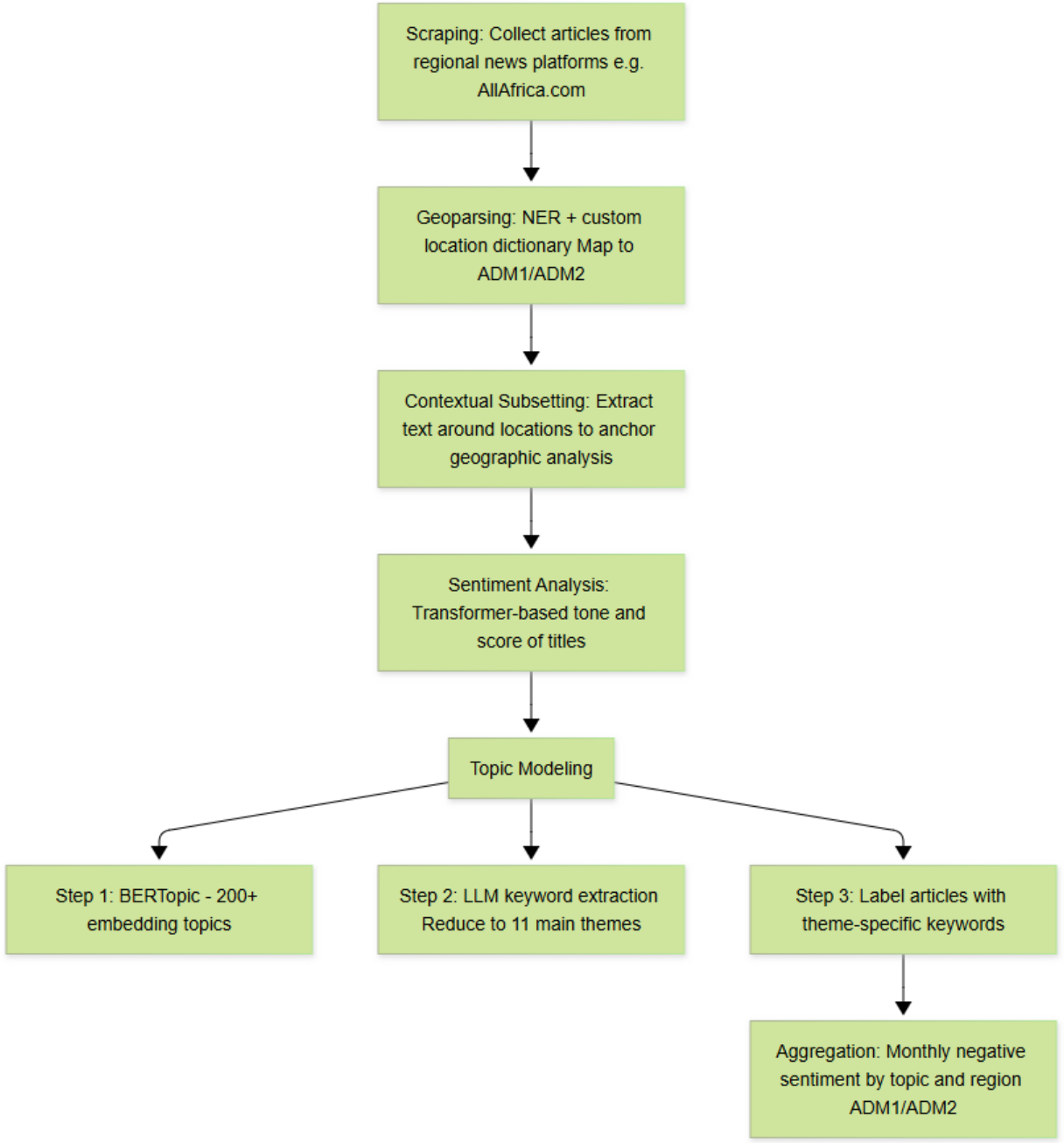
- **Source**: AllAfrica.com, with major contributions from outlets such as Sudan Tribune
- **Coverage**: 2011–2023.
- **Volume**: 18,000 unique articles collected
- Diverse reporting across South Sudan

Diverse reporting sources



Data Processing Pipeline

- The pipeline converts unstructured text into structured subnational indicators that can be trended and mapped.



Example: From News to Indicator

- **Raw Article (excerpt):**
  - *"Heavy fighting erupted near Bor in Jonglei State, displacing hundreds of families. Aid agencies reported disruptions to food distribution."*
- **Step 1 – Geoparsing:**
  - NER detects: *"Bor"* → matched to **Jonglei (ADM1)**.
  - Location anchored: ADM1 = Jonglei, ADM2 = Bor.
- **Step 2 – Contextual Subsetting:**
  - Keep nearby text: *"…fighting erupted near Bor…displacing hundreds…"*
- **Step 3 – Sentiment Analysis:**
  - DistilBERT classifies → **Negative or Positive (score: 0.91)**
- **Step 4 – Topic Modeling:**
  - Keywords matched to themes:
    - *"fighting, displacing"* → **Conflict & Displacement**
    - *"aid, food distribution"* → **Humanitarian Aid**
- **Step 5 – Aggregation:**
  - Record for **Bor, Jonglei, May 2017**:
    - Conflict articles +1
    - Displacement articles +1
    - Aid disruption articles +1
    - Avg. sentiment = –0.91

Method: Geoparsing Details

- Used BERT-based NER model (dslim/bert-base-NER) to extract **LOC** and **GPE** entities from cleaned text
- Matched entities against a curated gazetteer of South Sudan administrative units.
- Applied LLM-assisted disambiguation to standardize spelling variations and resolve ambiguities (e.g., “Wau” vs. “Wow”)

Method: Topic Modelling Details

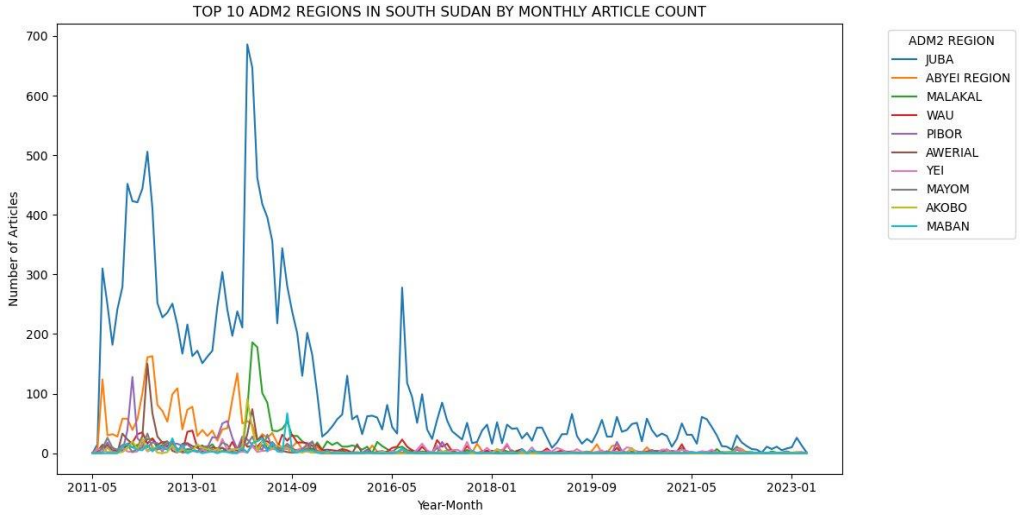
- **Embedding & Clustering:** Used HuggingFace BERTopic to create document embeddings and generate 200+ topic clusters.
- **Keyword Extraction:** Fed representative document from each cluster into an LLM to identify keywords tied to 11 predefined humanitarian themes (Balashankar et al., 2023), including *Conflict and Violence, Humanitarian Aid, Economic Issues, and Food Crisis*.
- **Classification:** Applied a dictionary-based multi-label classification, allowing articles to be assigned to multiple themes when overlaps occurred.

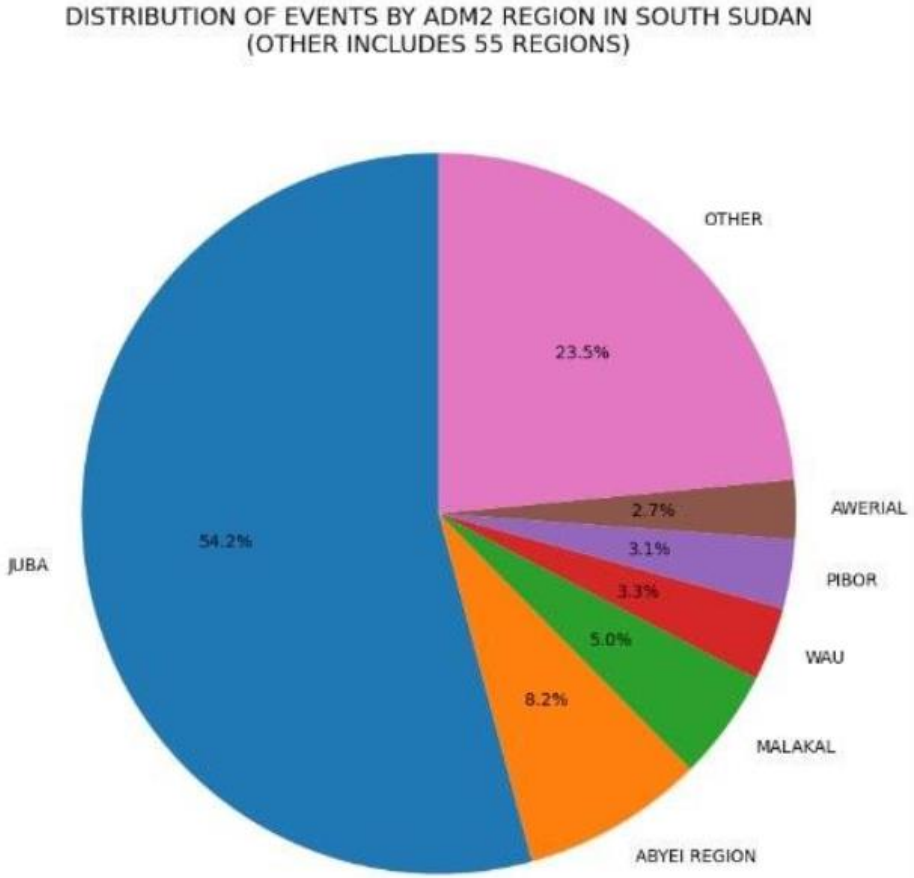
Method: Sentiment Analysis Details

- **Model**: DistilBERT fine-tuned on SST-2
- **Preprocessing**: Stopword removal, text cleaning, truncation at 512 tokens
- **Output**: Binary labels (Positive/Negative) with probability scores [0,1]
- **Aggregation**: Averaged sentiment scores per article, then aggregated by ADM1/ADM2 and theme over time

Results: Geocoded Coverage

- 11,000 articles successfully mapped to ADM1/ADM2 (~60%)
- About 50% referenced **Juba**
- Regional distribution shows coverage of conflict-heavy areas (Jonglei, Upper Nile) and aid operations



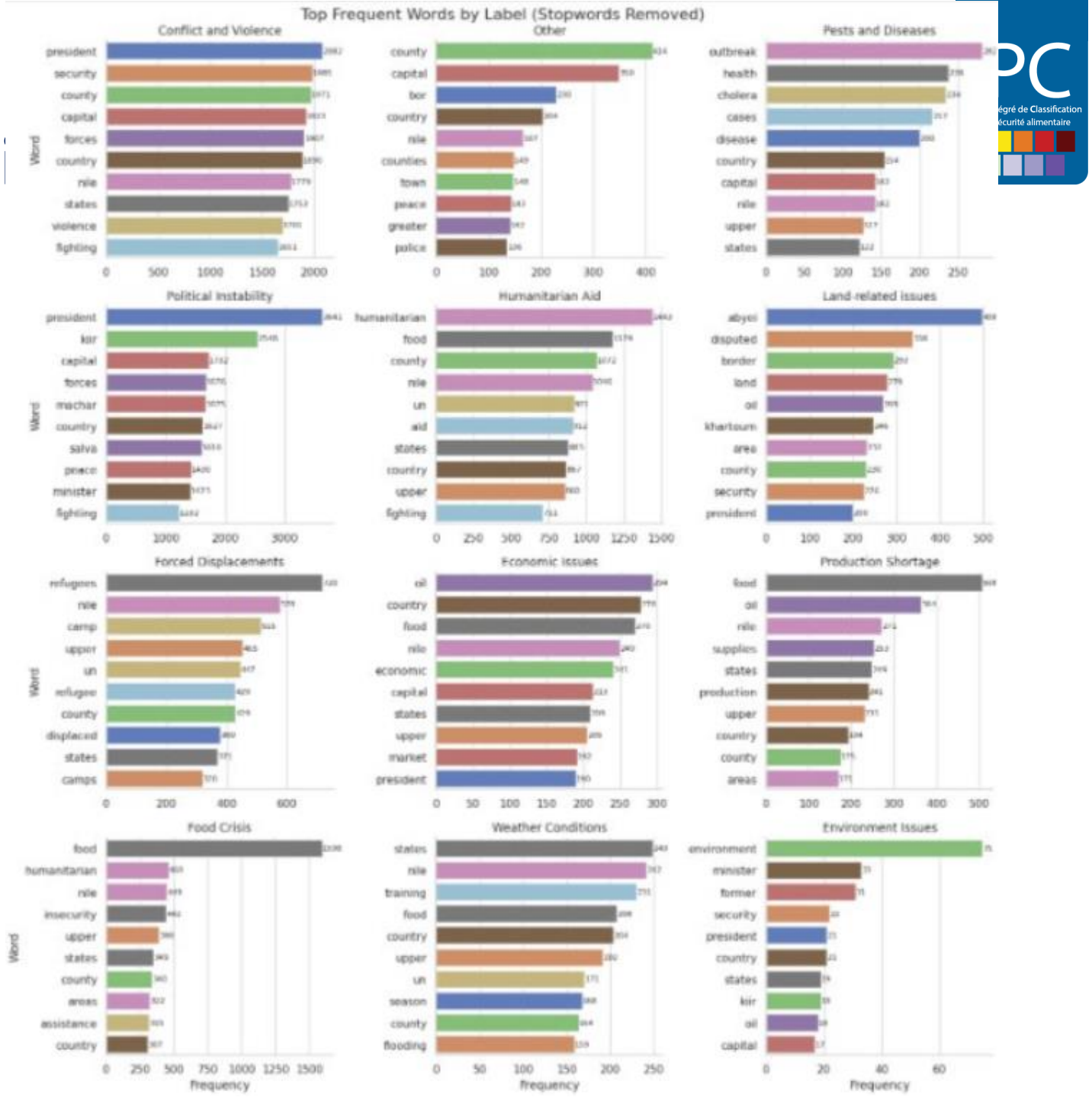


Results: Topic Taxonomy (Balashankar et al, 2023)

Category	Keywords
Conflict and Violence	conflict, war, fighting, battle, violence, clash, attack, military, armed, rebel, soldier, security, bomb, shell, shooting, insurgent, terror, terrorism, casualty, hostage, airstrike
Political Instability	political, government, protest, demonstration, election, coup, instability, corruption, parliament, opposition, governance, policy, minister, president, cabinet, regime, referendum
Economic Issues	price, inflation, economy, economic, market, trade, currency, poverty, unemployment, growth, debt, finance, livelihood, wage, budget, gdp, cost, commodity, imports, exports
Weather Conditions	drought, rain, rainfall, storm, cyclone, hurricane, typhoon, flood, flooding, weather, climate, temperature, heatwave, monsoon
Production Shortage	harvest, production, yield, crop, crops, planting, farming, agriculture, livestock, pasture, supply, shortage, output, seed, fertilizer
Humanitarian Aid	aid, relief, assistance, humanitarian, donor, funding, wfp, unhcr, unicef, ngo, distribution, support, msf, red cross, icrc
Food Crisis	food, famine, hunger, nutrition, malnutrition, insecurity, ipc, starvation, acute, hungry
Land-related issues	land, tenure, dispute, boundary, eviction, pastoralist, grazing, farmland, property, encroachment
Forced Displacements	displacement, displaced, refugee, refugees, idp, idps, migrant, migration, camp, camps, asylum, relocation, returnee, returnees
Pests and Diseases	locust, pest, pests, disease, diseases, outbreak, cholera, malaria, ebola, measles, covid, virus, infection, armyworm, pandemic
Environment Issues	environment, environmental, deforestation, erosion, pollution, biodiversity, conservation, desertification, degradation, wildfire, climate change, greenhouse

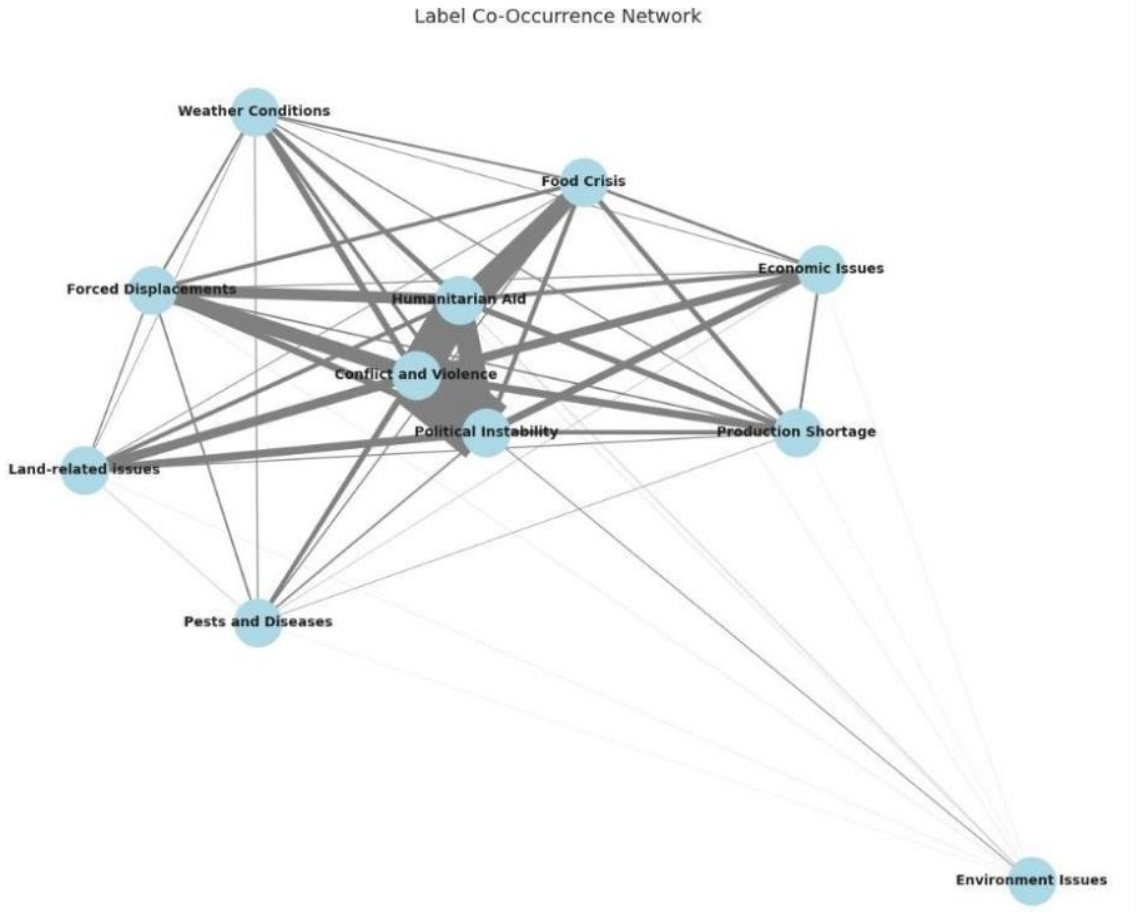
Results: Topic Model

- Keyword counts by topic



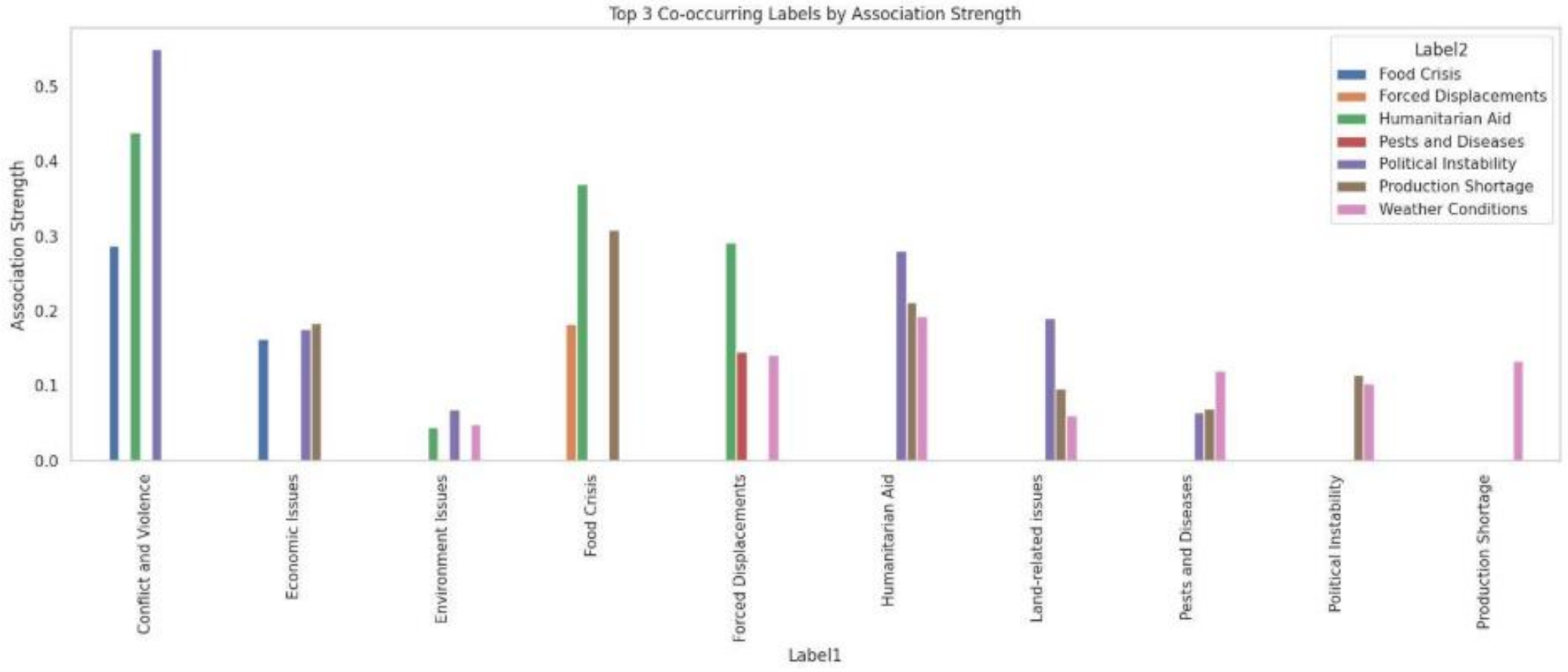
Results: Topic Modeling Outcomes

- Co-occurrence networks revealed theme interconnections (e.g., Conflict and Violence with Political Instability)



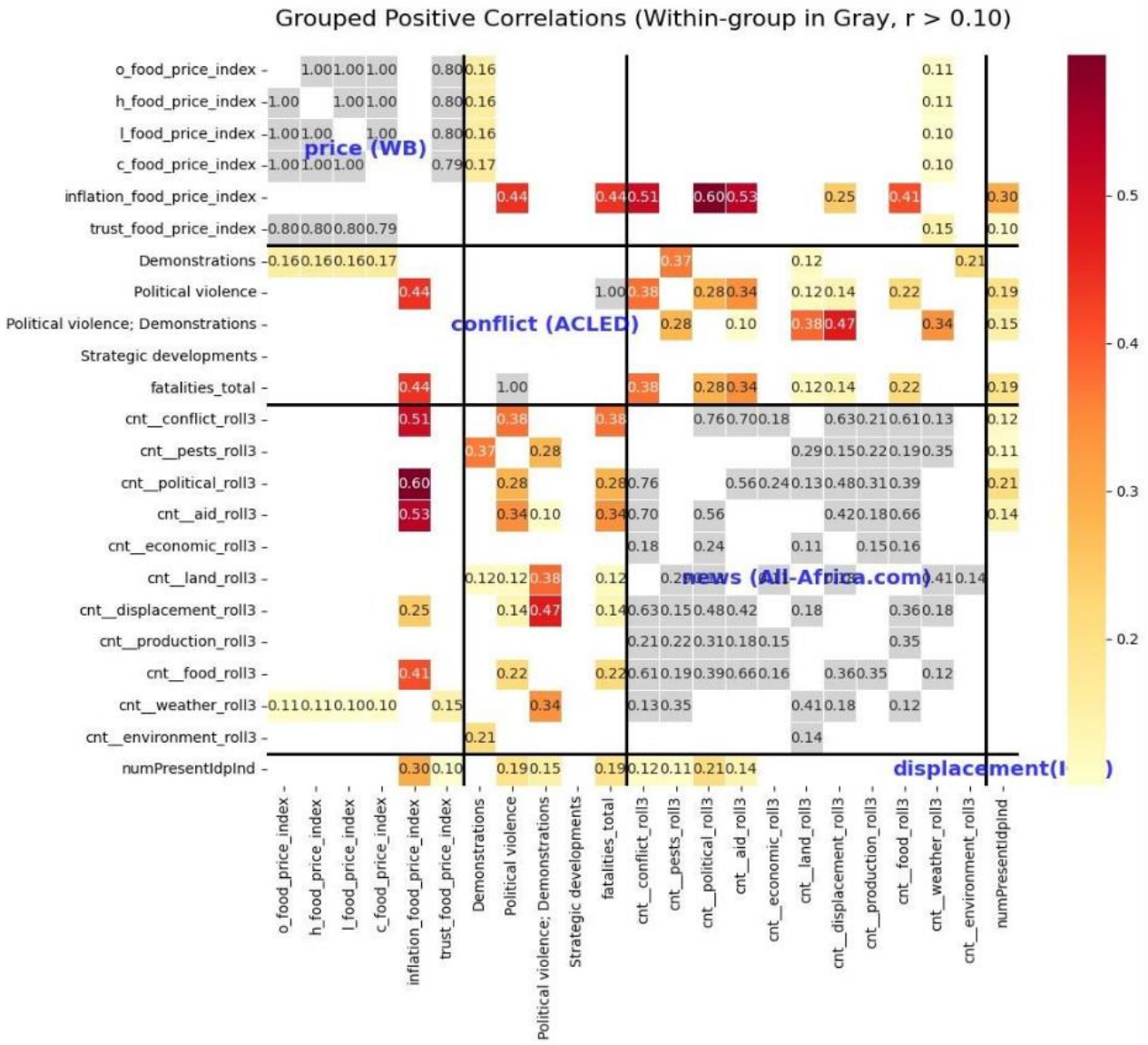
Results: Topic Modeling Outcomes

- Top 3 Co-occurring Labels by Pearson-Correlation Coefficient (or Association Strength)



Results: Linkages to External Variables

- Food Prices: World Bank food price inflation index
- Conflict: ACLED fatalities
- Displacement: IOM displacement data

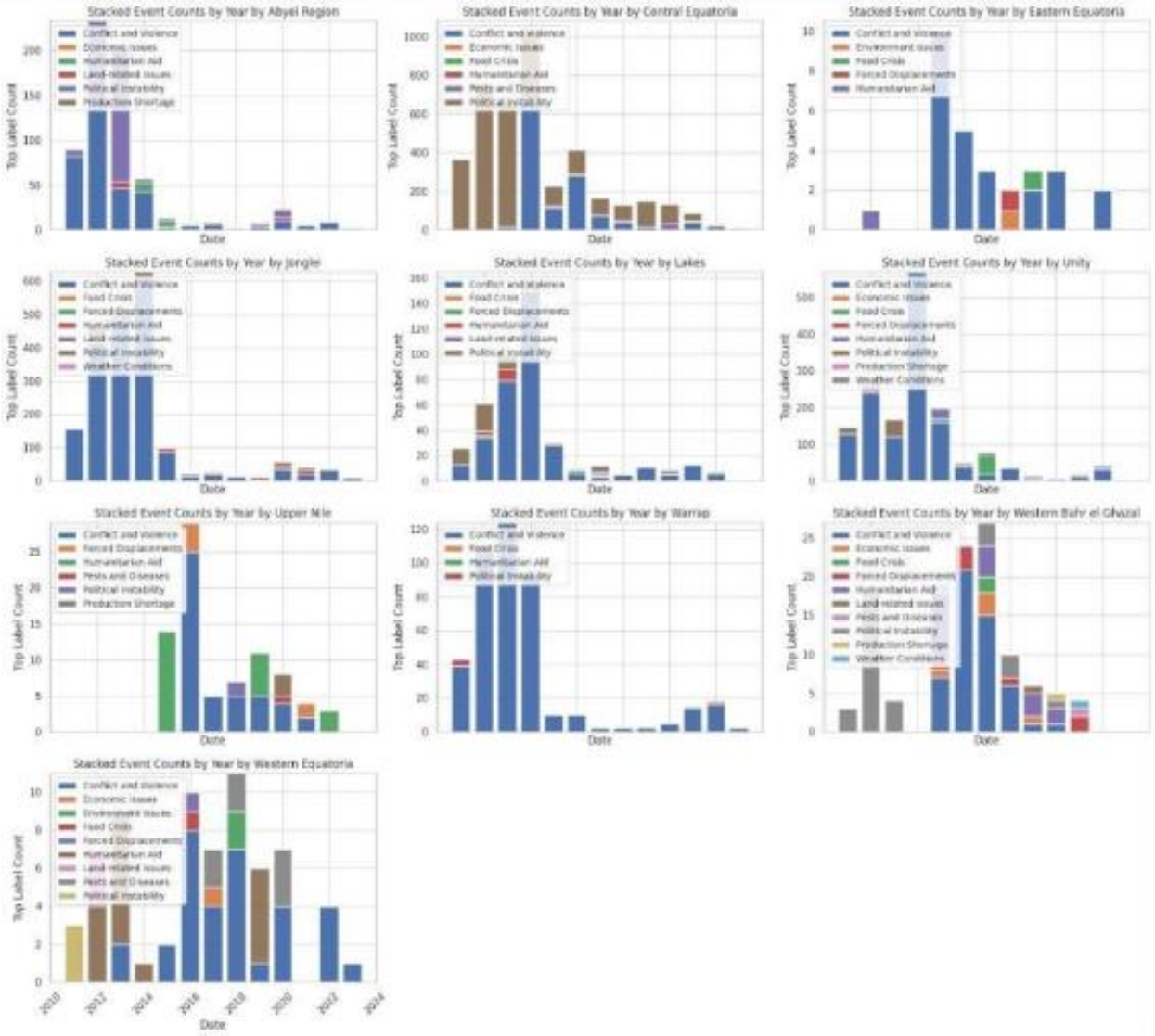


Use Cases: What Analysts Can Do With It?

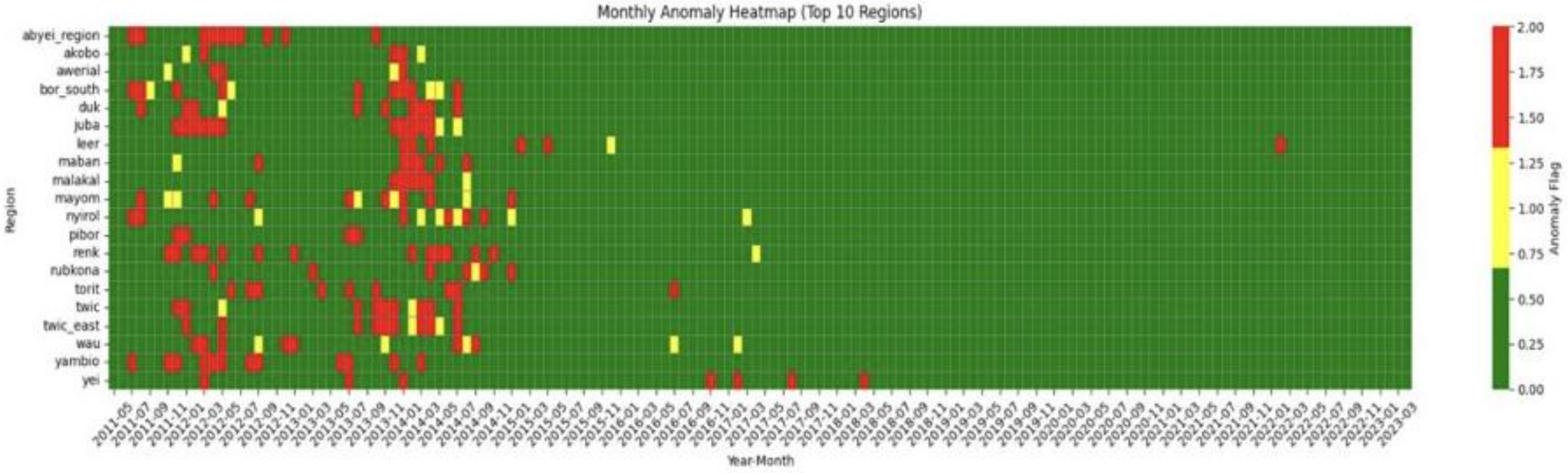
- Now we have transformed news articles into subnational indicators
  - Counts = media attention by topic and region
  - Sentiment = severity of events
- How can we use this?
  - Track thematic composition over time to see how risks cluster and evolve regionally (risk analytics)
  - Detect anomalies in article counts or sentiment for each topic and region (early alerts)
  - Use LLMs to generate summaries that add situational insights to complement quantitative indicators
  - Incorporate indicators into models predicting IPC outcomes

Use Case 1: Track thematic composition over time

- See how events cluster by region
- Compare trajectories across areas



Use Case 2: Anomaly Detection



Use Case 3: LLM-Powered Event Summaries

- Summarize key events by region and time using LLMs

You are a factual summarizer. Based only on the passages below, extract two things:

1. A one-sentence summary of the specific events related to **"{theme[0][0]} or {theme[0][1]}"** in or around the **{region_name}** region or associated ADM2 region name **{region2_name[0]}** of South Sudan between **{time_start}** and **{time_end}.** It must mention **"{region_name} or {region2_name[0]}".** Include a short quote or exact sentence fragment from the source that directly supports the summary.

2. A one-sentence summary of the potential implications of these events for food security in the {region_name} region, if any are clearly mentioned. Also include a direct quote or exact sentence fragment that supports this statement.

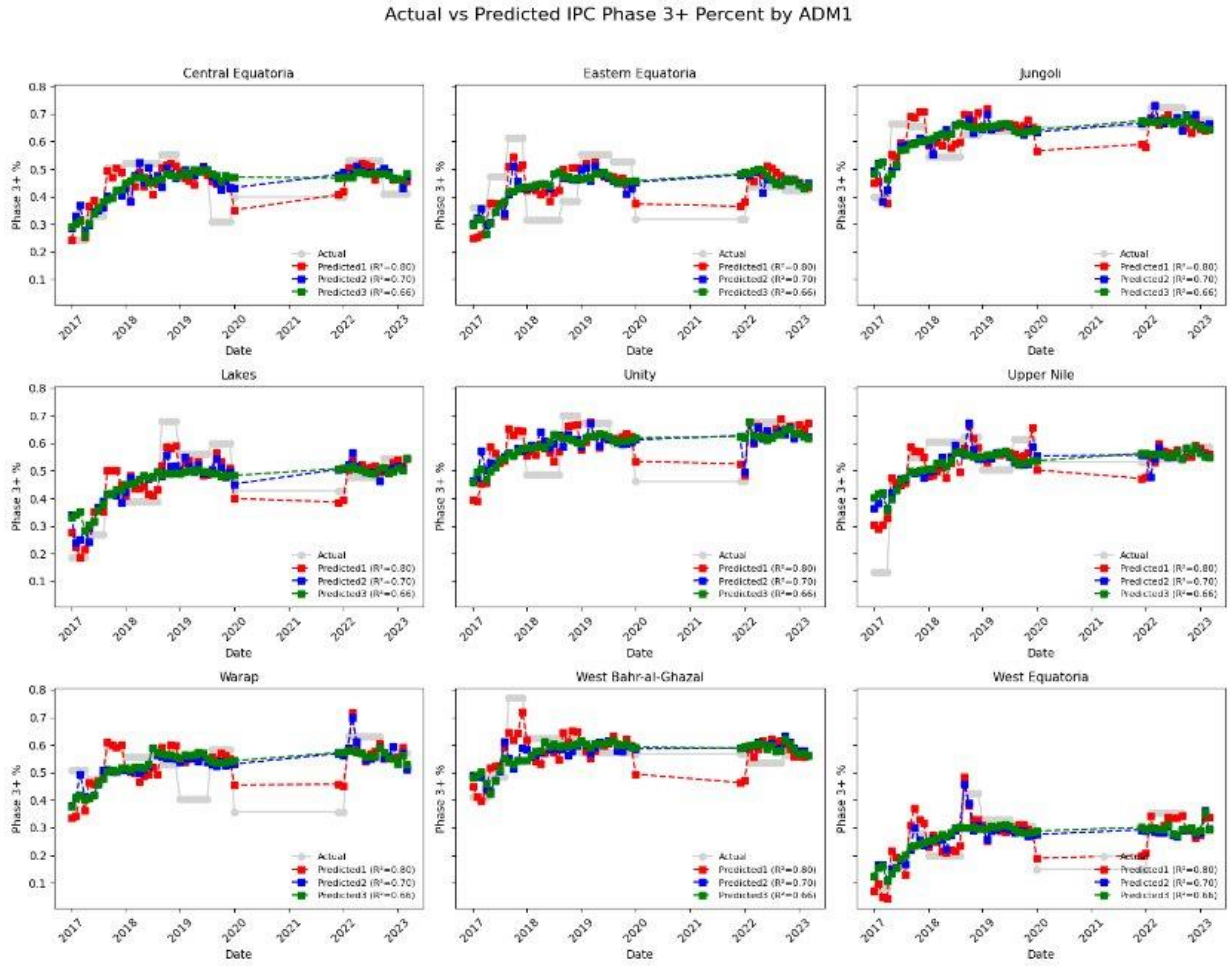
If the source text does not contain sufficient information to support either summary, write `"null"` for that field.

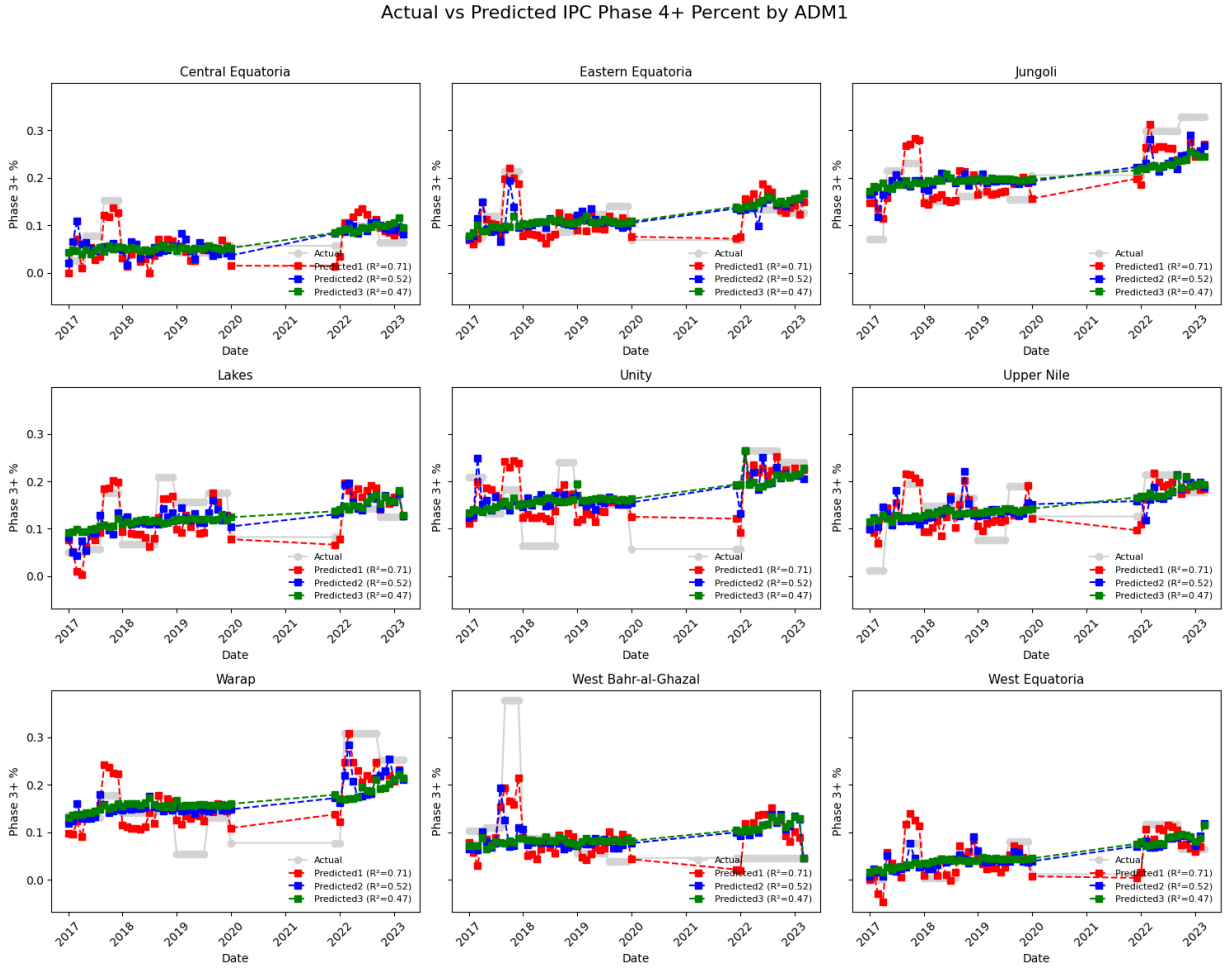
Return your answer in this exact format:

```
{{
  "summary": "...",
  "summary_source": "...",
  "impact": "...",
  "impact_source": "..."
}}
```

```
{'summary': "Between January and April
2023, South Sudan's Juba region experienced
political shock due to continued fighting,
accusations against Kenya of territorial
theft, and renewed clashes leading to an
influx of weapon-wounded patients in
hospitals.",
'impact': 'The political unrest and
violence in the Juba region could
potentially exacerbate issues of conflict,
flooding, and widespread famine, further
threatening food security in the area.',
'source': '}
```

Use Case 4: Incorporate indicators into models predicting IPC outcomes (in-sample prediction of 3+ and 4+ population share)





Limitations: Not a silver bullet

- **Media bias and coverage gaps**
  - News articles are unevenly distributed across regions and topics, with heavy concentration in **Juba (ADM2) and Central Equatoria (ADM1)**
  - Remote and insecure areas maybe often under-reported, leading to bias
- **Ambiguity in geoparsing**
  - Despite using NER and LLM-assisted matching, location might be imperfect
- **Noise in sentiment analysis**
  - Transformer-based sentiment models are not trained specifically on humanitarian text, and can misclassify neutral or technical reporting
  - Negative sentiment may reflect reporting style rather than actual severity of events
- **Topic classification constraints**
  - Dictionary-based labeling risks oversimplification of complex narratives
  - Overlapping crises (e.g., political instability linked with conflict) are difficult to disentangle
- **Temporal alignment issues**
  - News reporting lags or anticipates events inconsistently compared with ground conditions

Conclusion

- **Complementary data:** News analytics offers a *relatively* low-cost, near real-time supplement to surveys, especially valuable in fragile and conflict-affected contexts.
- **Direct pipeline:** Links original text → quantifiable indicators → anomaly detection → contextual insights
- **Integration:** Complements JRC AI-RAG (disasters) and DFS NLP Pipeline (structured evidence) to cover anticipation, diagnosis, and convergence.
- **Scalable & actionable:** Expandable to other regions; may support anticipatory action and provide digestible, evidence-based narratives for our anlaysts.

Test the Pipeline in South Sudan (post-2023)

- Refine & Apply the pipeline to recent data (2024 onward)
- Validate indicators against IPC outcomes, prices, and conflict
- Deliver outputs:
  - ***News Analytics Dashboard: scrap articles (original text) → quantifiable indicators (article counts & sentiment analysis) → anomaly detection → contextual insights (using RAG & LLM)***

Scaling to High-Profile Countries on a Monthly Basis

- **Pipeline automation:** monthly scraping, geocoding, topic and sentiment scoring
- **Data platform:** centralized storage with version control
- **Delivery:** analyst dashboards (Streamlit/Power BI), APIs, and monthly summaries
- **Quality assurance:** validation checks & coverage monitoring