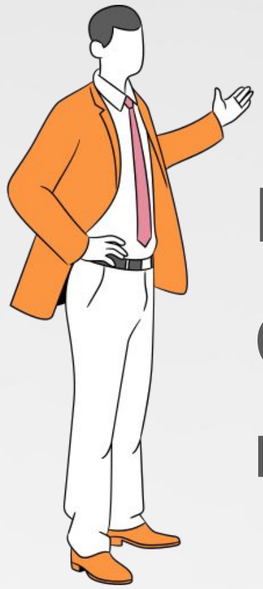
Countries in the News

FDAC 545



Faith Chernowski, Harshvardhan, Colin Canonaco, Jeremiah Augustine

December 3, 2024

Motivation

- Analyze the portrayal of countries in international media using sentiment analysis
- Identify patterns in news coverage that can help to understand their influence on global perception and international relations
- Reveal biases and disparities in news coverage

Dataset and Sampling

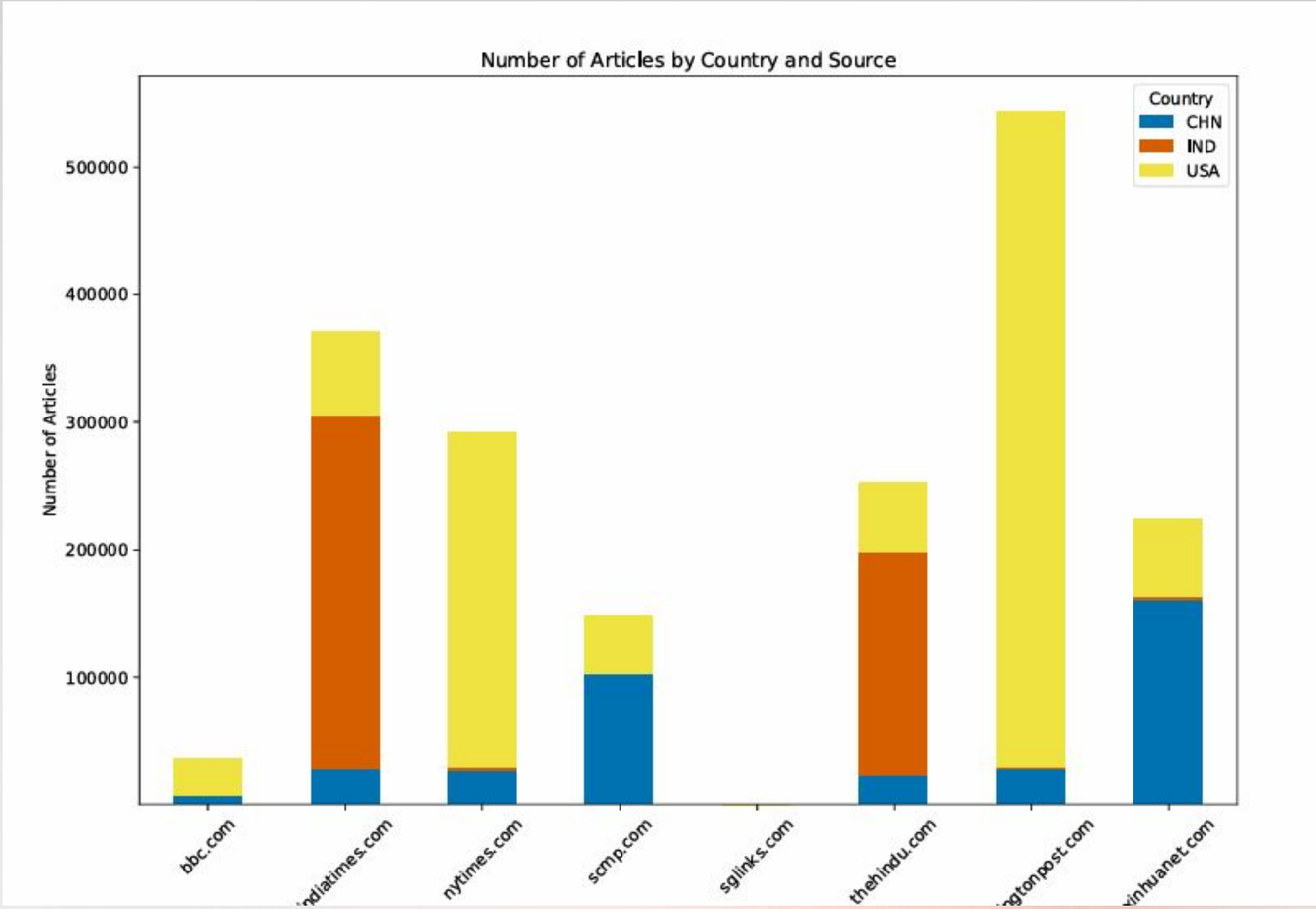
- GDELT 1.0: Open-access global database tracking events since 1979
  - 2013-2024 = 8 GB, 2.5% random sample -> 200 MB
  - Google BigQuery and Google Cloud Storage
- Attributes used:
  - Event date, country affiliation, stabilizing score, sentiment score, source
- Data Scope
  - Articles form 2013-2024 from publishers like The New York Times, BBC, Xinhua, etc
  - 1,868,683 articles included

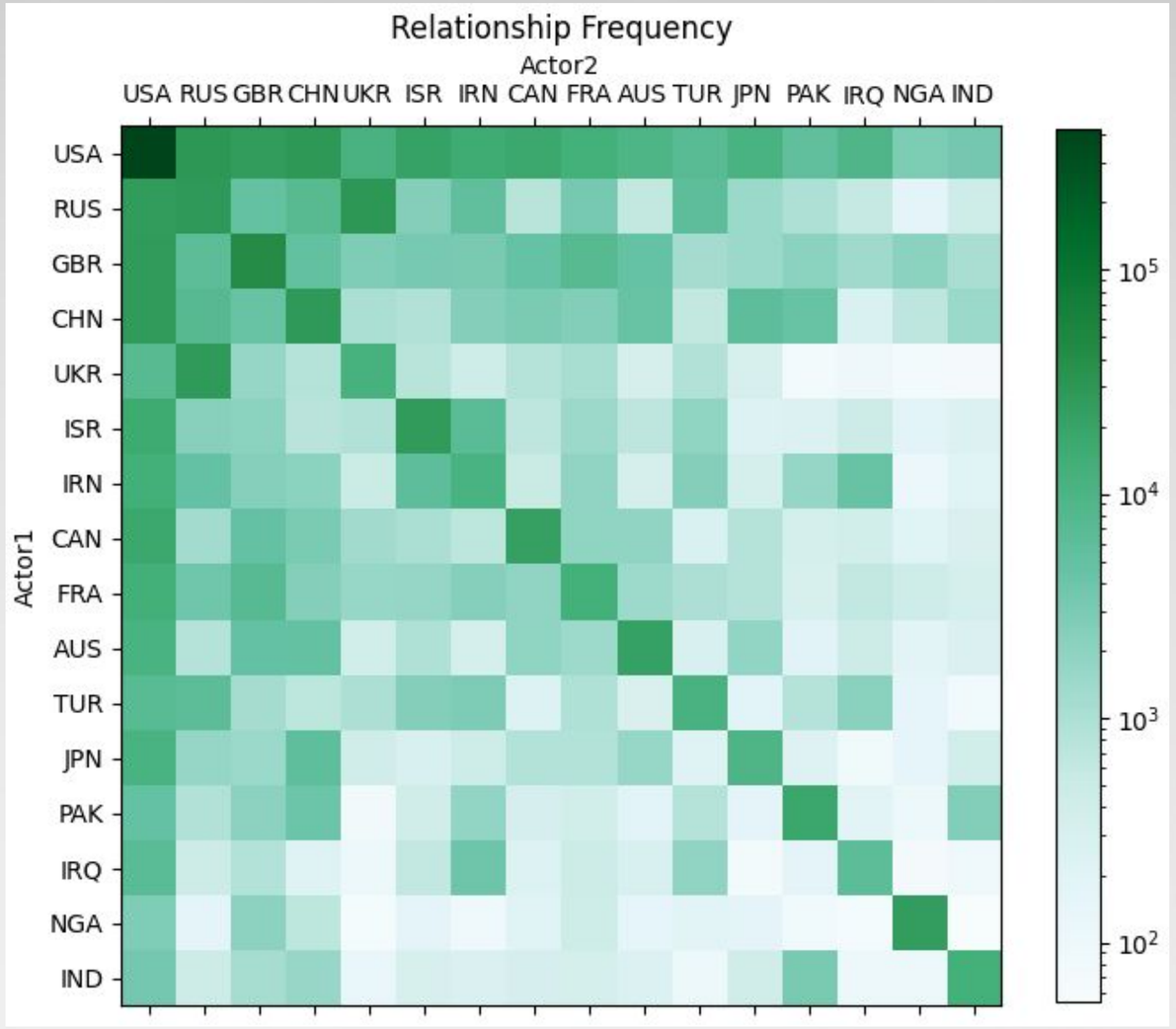
Methods

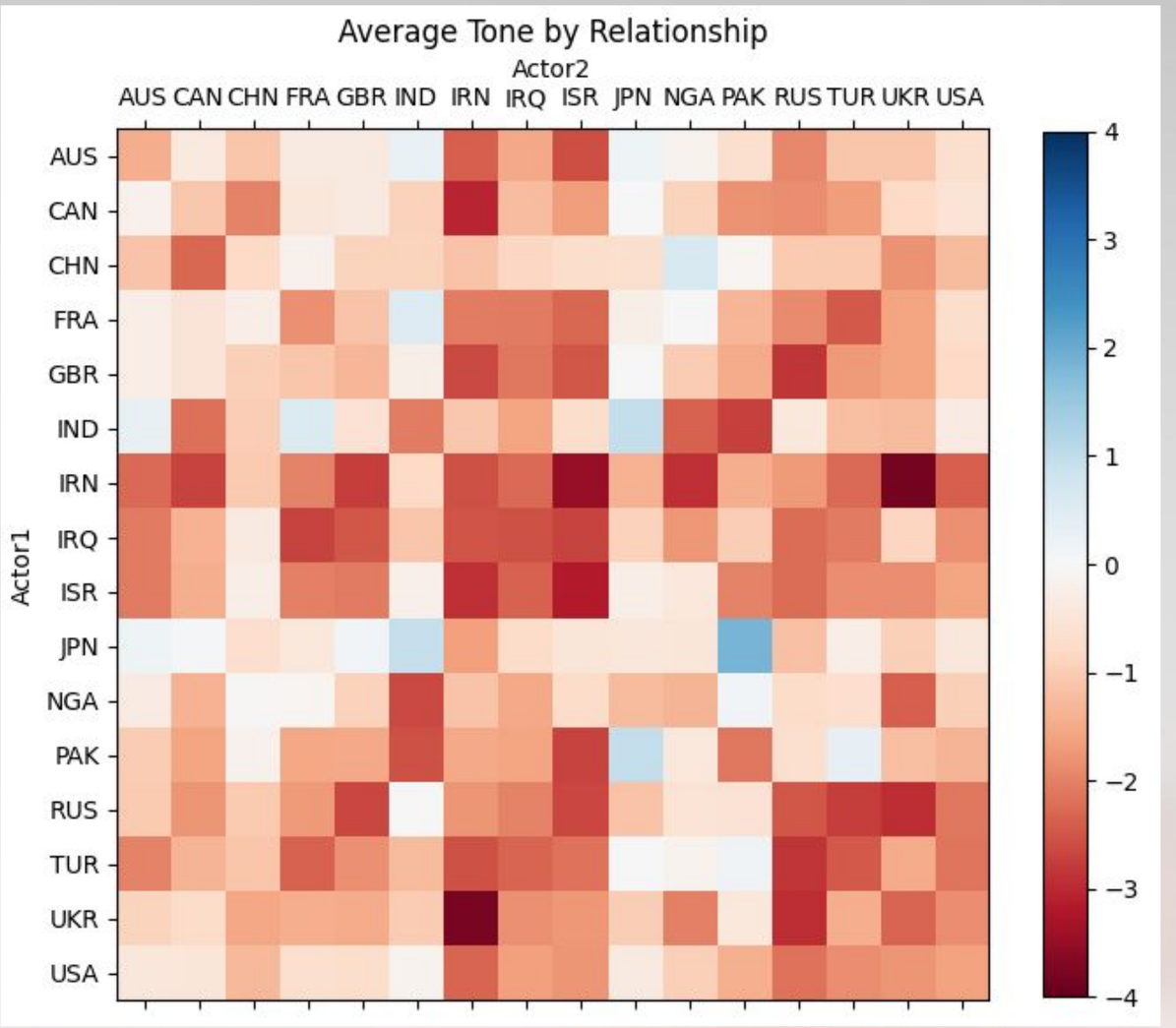
- Data Preparation:
  - SQL query using Google BigQuery
  - CSV with date, countries, tone, Goldstein, source
  - Filtered query to focus on specific sources
- Visualization
  - Color-coded matrices for individual relationships
  - Bar and line graphs for broader trends
  - Sorting axis labels to distinguish patterns

Results

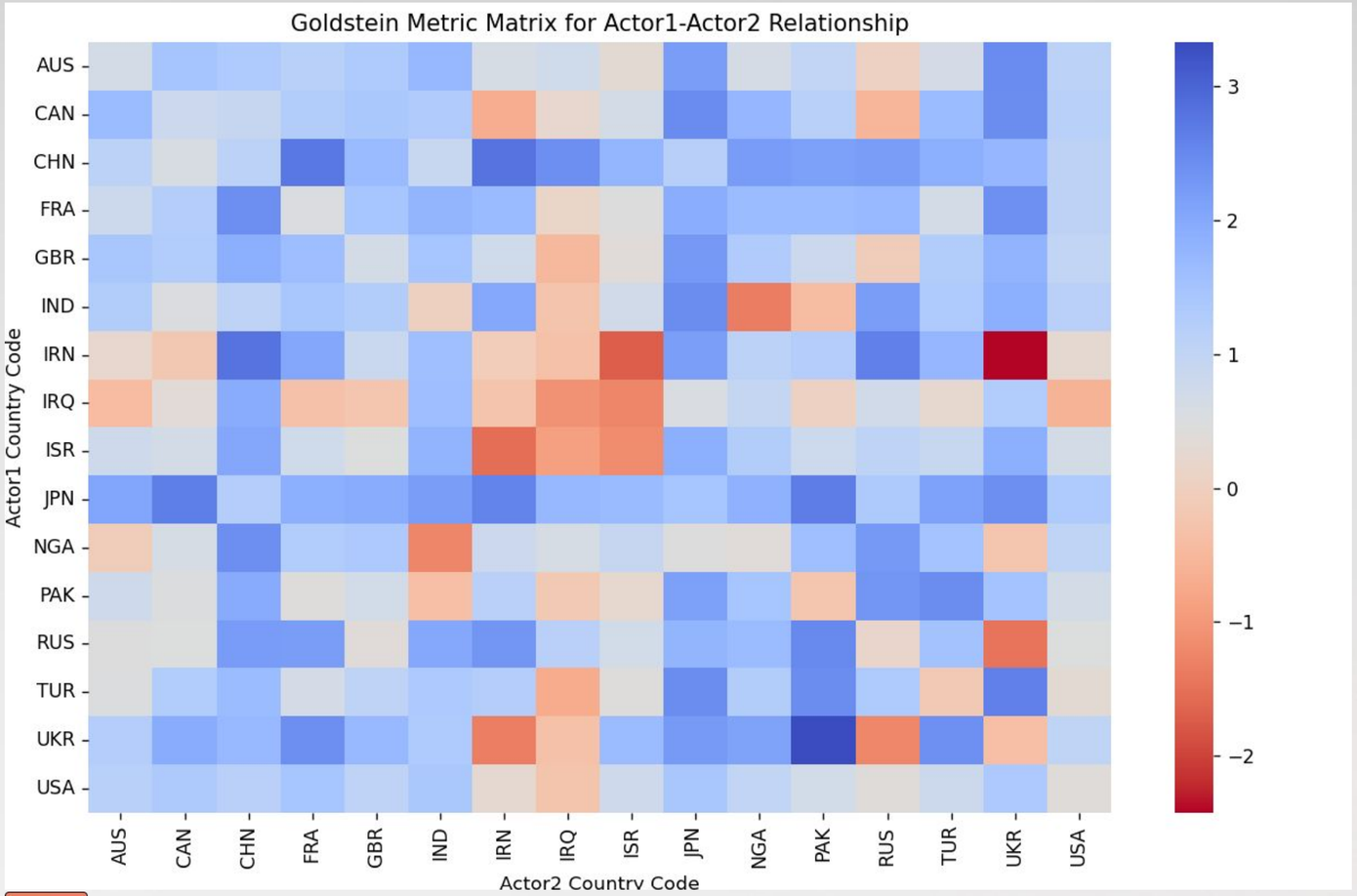
- Coverage differs greatly by news source
  - Quantity
  - Proportion
- Motivated analysis of country relationships for broader perspective

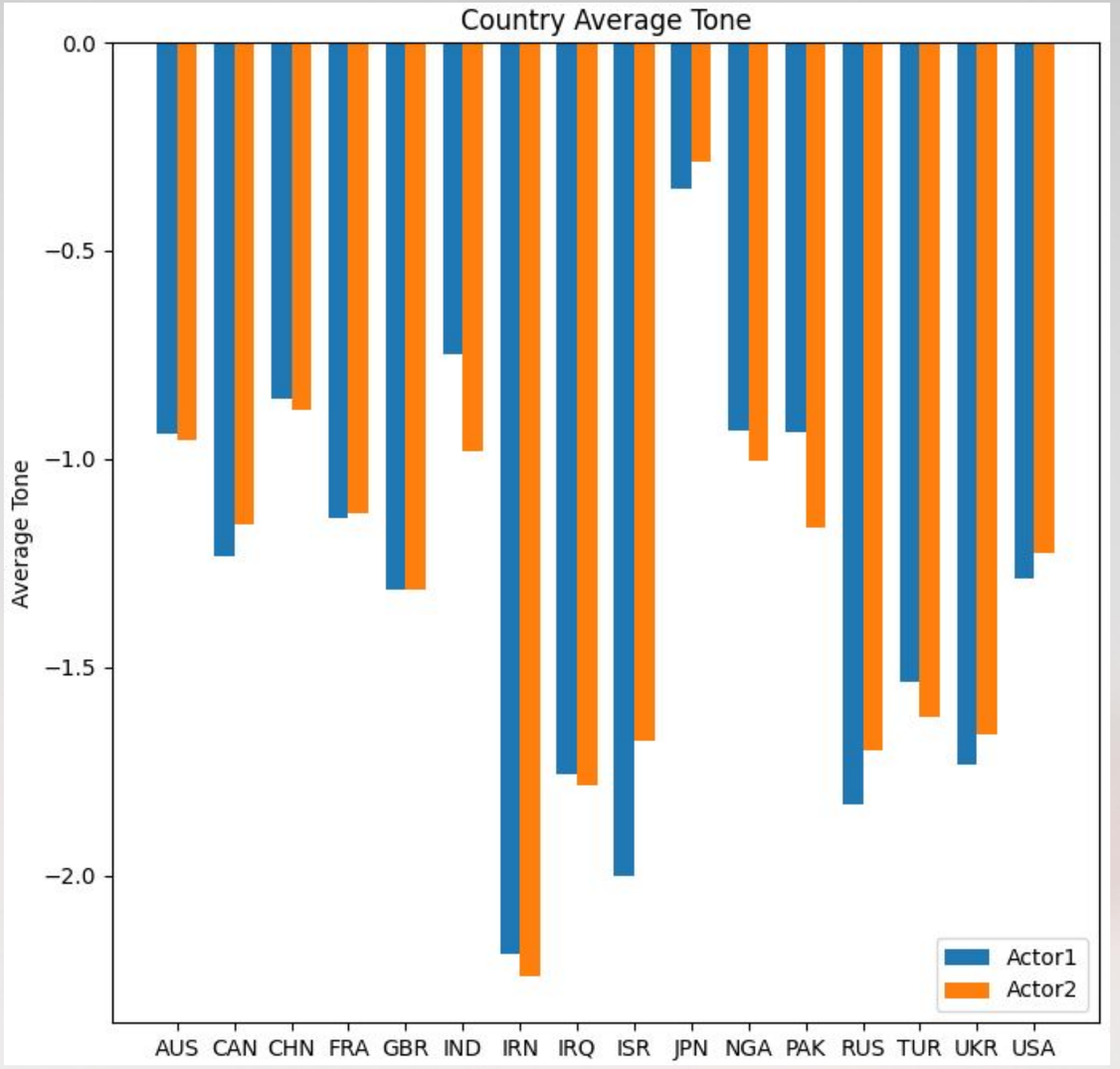



Relationship Frequency
Actor2
USA RUS GBR CHN UKR ISR IRN CAN FRA AUS TUR JPN PAK IRQ NGA IND
Actor1
USA
RUS
GBR
CHN
UKR
ISR
IRN
CAN
FRA
AUS
TUR
JPN
PAK
IRQ
NGA
IND
10^5
10^4
10^3
10^2

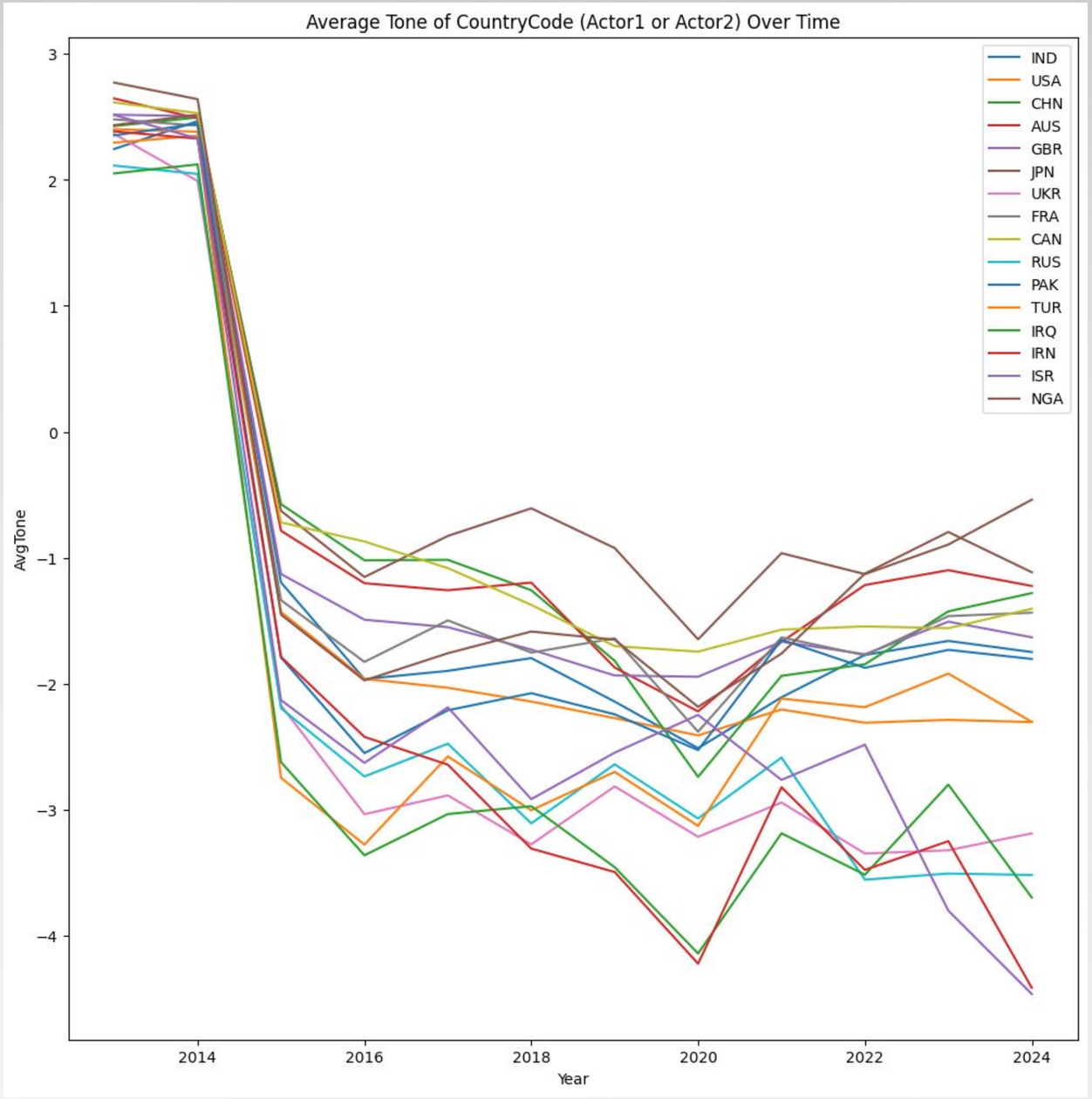

Average Tone by Relationship
Actor2
AUS CAN CHN FRA GBR IND IRN IRQ ISR JPN NGA PAK RUS TUR UKR USA
Actor1
AUS
CAN
CHN
FRA
GBR
IND
IRN
IRQ
ISR
JPN
NGA
PAK
RUS
TUR
UKR
USA
4
3
2
1
0
−1
−2
−3
−4

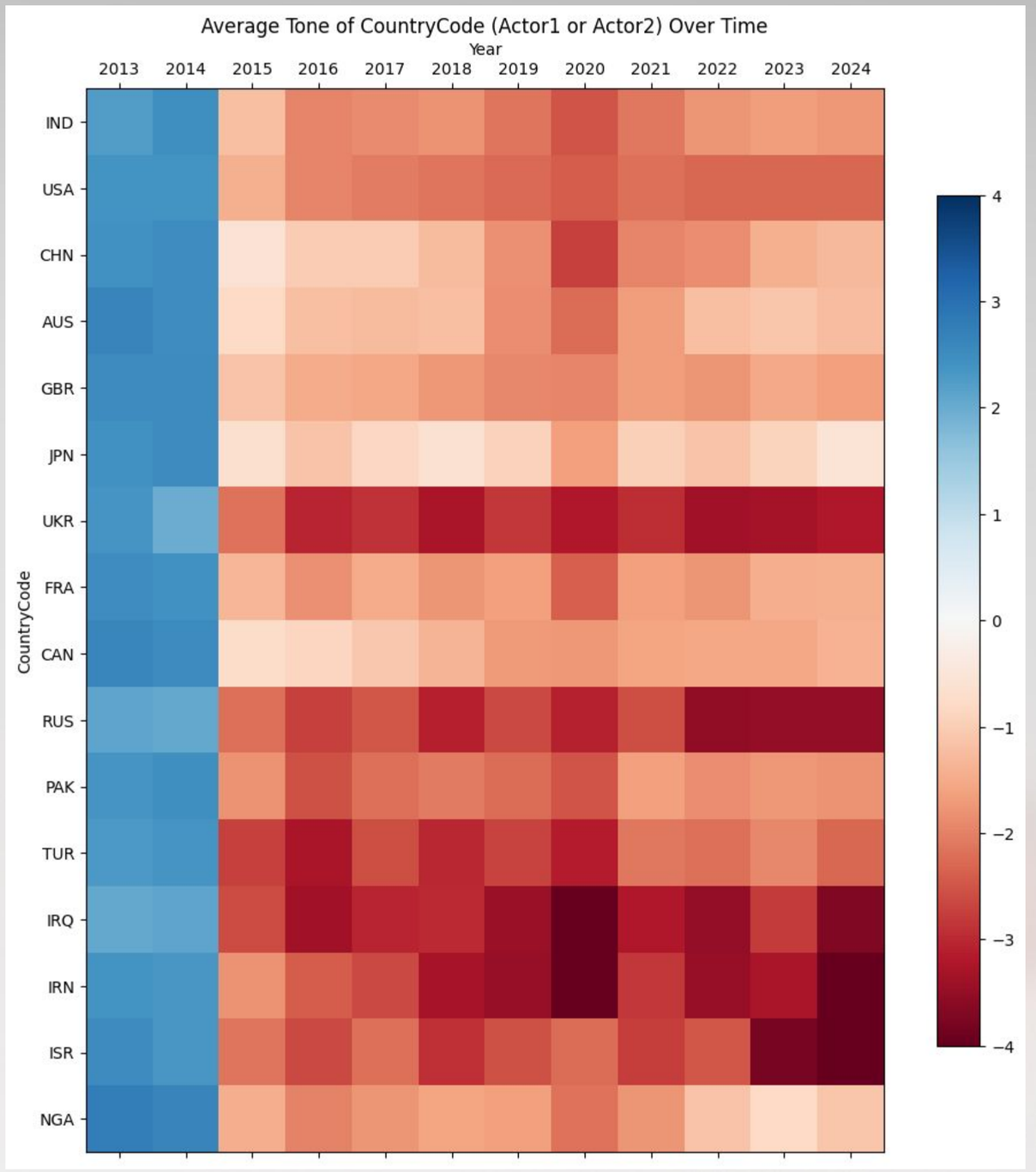
Results


Goldstein Metric Matrix for Actor1-Actor2 Relationship
Actor1 Country Code
AUS
CAN
CHN
FRA
GBR
IND
IRN
IRQ
ISR
JPN
NGA
PAK
RUS
TUR
UKR
USA
Actor2 Country Code
3
2
1
0
−1
−2


Country Average Tone
Average Tone
0.0
−0.5
−1.0
−1.5
−2.0
AUS CAN CHN FRA GBR IND IRN IRQ ISR JPN NGA PAK RUS TUR UKR USA
Actor1
Actor2

Avg Tone Over Time





Potential Biases

Analysis on Potential Biases:

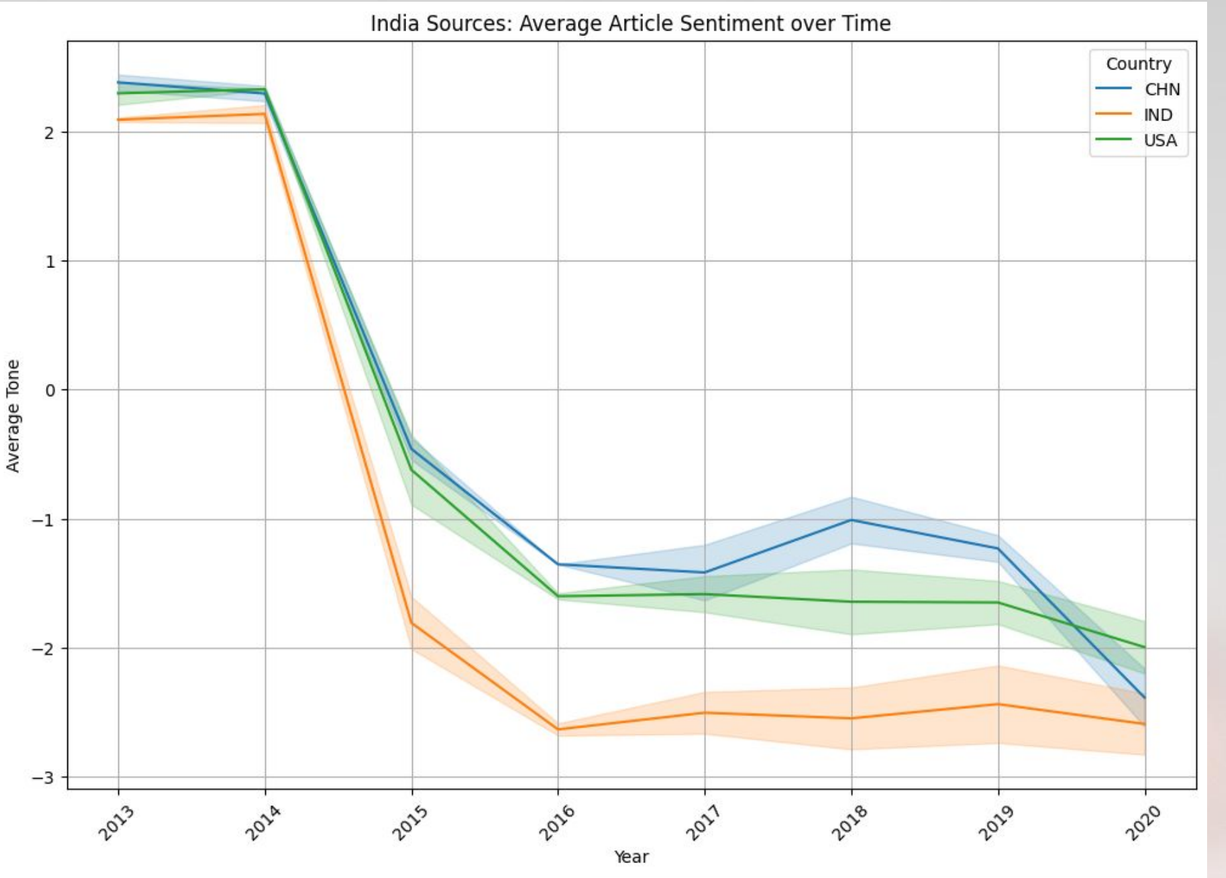
- Must know sources Country of Origin
- This allows us to see how other countries news bias

Chosen Sources:

- NYtimes & WashingtonPost: USA
- SCMP & XinhuaNet: China
- IndiaTimes & TheHindu: India

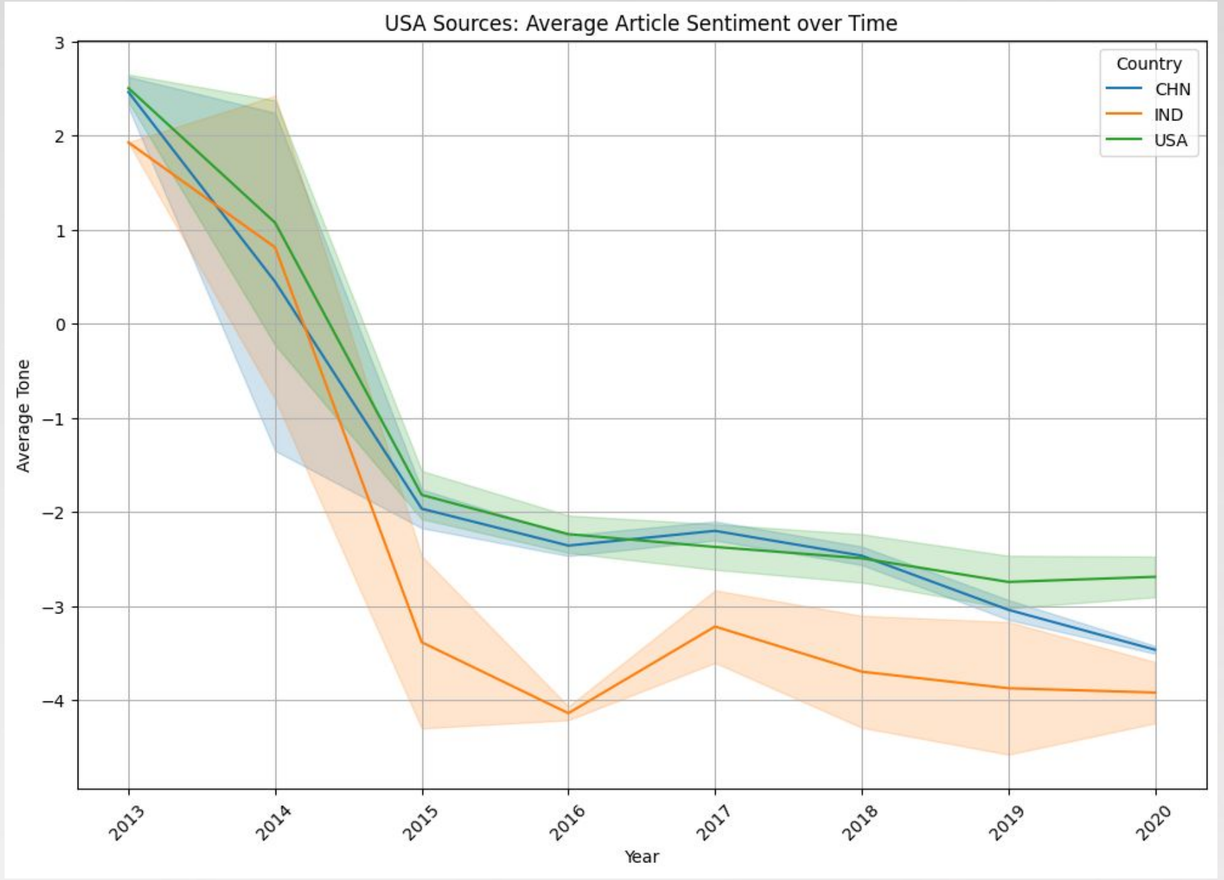
To the right is a good baseline of what we found

This graph was seen as fairly unbiased

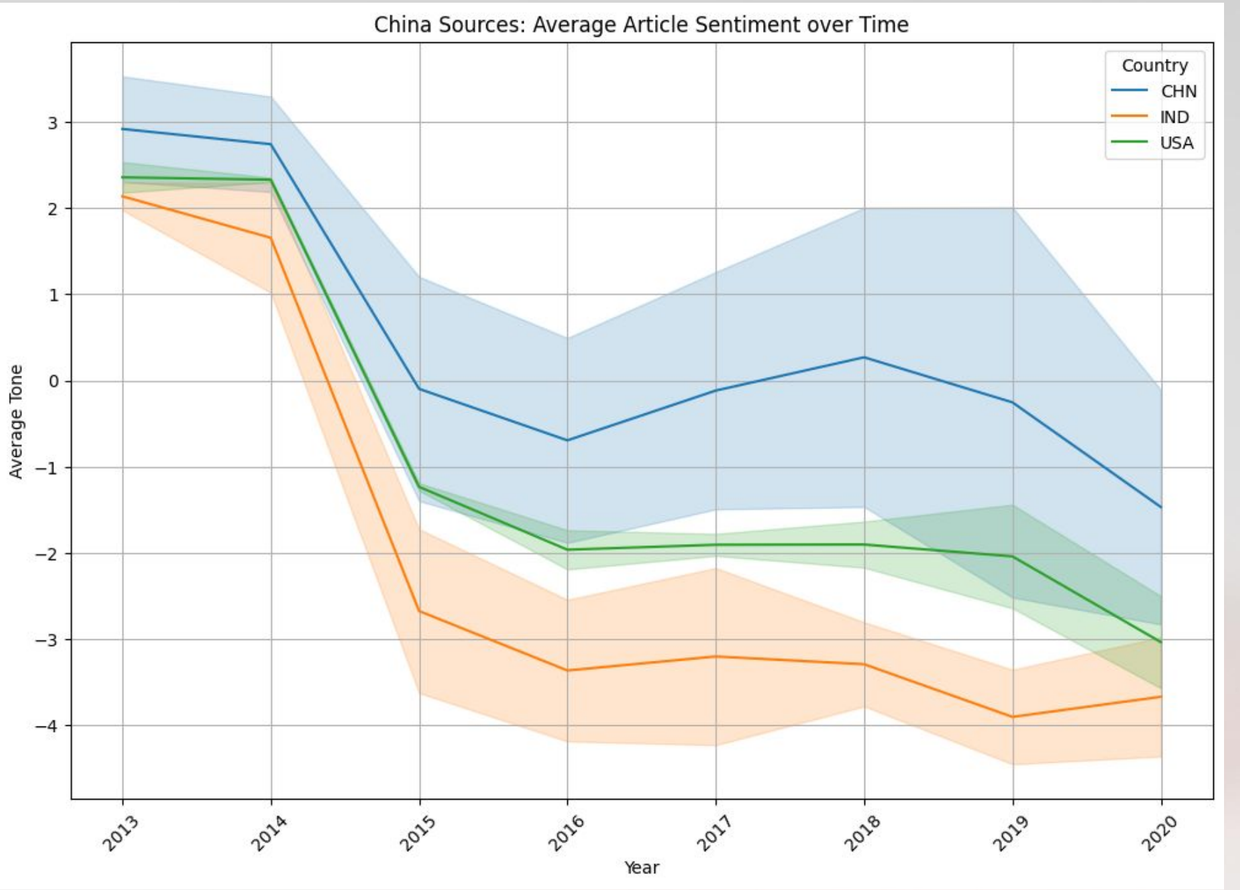


Potential Biases: Results

Fairly Unbiased



Clear Bias (China Sources -> China Avg Tone)



Word Analysis

Analyze Article Titles extracted from URLs to Identify words with the strongest correlation to the average tone of the articles.

- **Extract Article Titles**
- **Tokenize Titles:** Break down article titles into individual words
- **Pair Words with Their Article's Tones**
- **Calculate Correlation:** Group data and calculate the correlation between all of the word occurrences and the average tone of articles.

Word Analysis

Word	Correlation
pm	0.050511
modi	0.048074
to	0.047394
cooperation	0.046950
summit	0.046828
prince	0.046653
celebrates	0.045348
visit	0.044791
ties	0.039746
c90000	0.037597
meets	0.036833
president	0.036323
with	0.035501
china	0.035466
king	0.035315
celebrate	0.035041
charles	0.034916
harry	0.034842
11ef	0.034663
content	0.033035

Positive Correlation

Word	Correlation
israel	-0.112326
attack	-0.094831
killed	-0.086772
man	-0.085358
iran	-0.078577
israeli	-0.075639
gaza	-0.074471
murder	-0.073554
russian	-0.071784
police	-0.070379
hezbollah	-0.068671
hamas	-0.068544
war	-0.067689
shooting	-0.067227
attacks	-0.065330
kill	-0.063166
accused	-0.061525
charged	-0.060675
arrested	-0.060009
death	-0.058084

Negative Correlation

Primary Issues

- Size of data set was overwhelming so storing it was a pain and we had to pay for storage at one point
- We had to get a new dataset halfway through to use more recent articles
- Topic modeling didn’t work for the 1st GDELT data set because there was no text or content
  - Also unable to scrape all the websites for text despite numerous attempts (they didn’t like it)
  - Used an LDA algorithm for the topic modeling

Team Member	Responsibilities
Colin Canonaco	Exploring, investigating, and cleaning original GDELT data set and performing sentiment analysis. Use data/results to identify biases and potential conclusions.
Harshvardhan	Helping with data set cleaning and original analysis of news sources, then overseeing analyses. Use data/results to identify biases and potential conclusions.
Jeremiah Augustine	Worked with sentiment analysis and creating visuals to represent results and conclusions. Use data/results to identify biases and potential conclusions.
Faith Chernowski	Worked with other analysis methods such as topic modelling and clustering to represent scraped data. Use data/results to identify biases and potential conclusions.

Timeline

Sprint 1 - Sep 15 - Sep 30

Project proposal and original GDELT data scraping

Sprint 2 - Oct 1 - Oct 15

Topic modeling attempts and analysis by news source numbers

Sprint 3: Oct 15 - Oct 30

Secondary data set retrieval and original sentiment analysis averages

Sprint 4: Oct 30 - Nov 15

Tone visualization by country and over time (sentiment analysis)

Sprint 5: Nov 15 - Nov 30

Presentation and final report creation, finishing results discussion/analysis

Future Work

- Expand Dataset Scope: Include additional timeframes, more countries, and diverse publishers to capture broader media perspectives.
- Deepen Analysis: Investigate sub-national portrayals, cross-domain topics like environmental or human rights issues, and cultural influences on media narratives.
- Advance Techniques: Apply fine-grained sentiment analysis, network analysis for global media dynamics, and event chronology to track narrative shifts.
- Real-world Applications: Develop policy recommendations, media literacy tools, and interactive visualizations for public and academic use.
- Comparative Studies: Analyze platform-specific portrayals, cultural framing differences, and contrasts between traditional and modern media.

THANK YOU