Countries in News: Project Report

Colin Canonaco[1], Faith Chernowski[1], Harshvardhan[2], Jeremiah Augustine[1]

***Abstract*— This paper investigates the portrayal of countries in international media through sentiment analysis, bias detection, and topic modeling. Using the GDELT dataset and computational methods, the study identifies patterns and disparities in media coverage. The results provide insights into global perception, media biases, and their implications for international relations.**

I. Introduction

A. Objective

This project focuses on analyzing the portrayal of countries in international media. Our goal is to understand the patterns in news coverage and identify biases that might shape global perception. To achieve this, we used the GDELT dataset, which tracks media reports worldwide, and applied various techniques like sentiment analysis, bias detection, and topic modeling.

We aimed to understand how different countries are represented in the media and whether these portrayals vary based on the publisher's location or the nature of the events reported. By using computational analysis, we were able to explore trends in sentiment and the overall tone of international news, revealing both positive and negative biases. The findings provide a deeper understanding of how media shapes narratives about countries and influences global opinions.

B. Motivation

Media coverage plays a powerful role in shaping how people see the world. It influences public perception, affects international relations, and even impacts government policies. With so much information flowing from different sources, it's essential to understand whether media outlets are presenting countries in a fair and unbiased way. Our motivation was to uncover these portrayals and examine the differences in coverage across various countries and publishers.

We wanted to determine if certain countries receive consistently positive or negative portrayals, and how events like conflicts or political changes influence media sentiment. Understanding these biases can help policymakers, journalists, and the public become more critical consumers of news. By recognizing patterns in coverage, we can address imbalances and work towards more objective and informed reporting.

*This work was done for partial fulfilment of FDAC 2024 course requirements.

[1]Colin Canonaco, Faith Chernowski, and Jeremiah Augustine are with the Department of Computer Science, University of Tennessee, Knoxville {ccanonac, fchernow, jaugust4}@utk.edu

[2]Harshvardhan is with the Department of Business Analytics, University of Tennessee, Knoxville harshvar@utk.edu

II. Dataset

The GDELT 1.0 is an open-access database with records of events from around the world since 1979 [1]. It contains information about individual events as they were reported on. Each event entry is contributed to by articles that mention that event, and the first article to mention it is listed as the source. Event entry features relevant to our analysis also include the date, the countries involved and their roles, an average sentiment score of the language of the articles, and the average Goldstein scale value that measures the level of cooperation or conflict associated with an event.

A. Variables

- **Event Date:** The date when the event occurred.
- **Actor1 Country:** The primary country involved in the event.
- **Actor2 Country:** The secondary country involved in the event (if applicable).
- **Event Source:** The first news article that reported the event.
- **Average Sentiment Score:** A measure of the sentiment expressed in articles mentioning the event (positive or negative).
- **Goldstein Scale Value:** An intensity scale for measuring the impact of events, with positive, negative and neutral scores.
- **Event ID:** A unique identifier for each recorded event.
- **Event Type:** A classification of the event (e.g., protest, agreement).
- **Geolocation:** Latitude and longitude coordinates of the event's location.

B. Source

The GDELT 1.0 dataset compiles event records from diverse global news sources, including print, broadcast, and online media. Each event is linked to multiple articles, with the first article to report it designated as the primary source. This ensures traceability and offers a dynamic repository of global events, updated in real-time. The dataset's source diversity enhances its utility for analyzing geopolitical trends, sentiment patterns, and country-specific event dynamics. More detailed information is available at https://www.gdeltproject.org/.

Our analysis files are located in the Github repository https://github.com/harshvardhaniimi/countries-in-news.

III. Methodology

A. Data Preparation

The GDELT 1.0 dataset is very large with events between 2013 and 2024 being 8 GB in size. We took a 2.5% sample of the events from that period to get a smaller dataset that was easier to work with while still being representative of the population. This dataset was about 200 MB and contained 1,486,941 events. We sampled a second dataset by getting all events from specific news outlets that involved India, China, or the United States. This sampled dataset was about 330 MB and contained 1,868,066 events.

We acquired our data using SQL queries to the GDELT 1.0 database through Google BigQuery [2] and downloaded it through Google Cloud Storage. This provided us with CSV files to analyze.

B. Analysis Techniques

Our analytical approach focused primarily on dissecting the tone and relational dynamics present in the media coverage drawn from the GDELT dataset. One key method was sentiment analysis, where we used the AvgTone metric as a proxy to understand the positivity or negativity of media reporting. Higher AvgTone values indicated more favorable coverage, while lower or negative values signified critical or adverse portrayals. By grouping the data by publisher, country, and year, we extracted trends that highlighted how the media's tone evolved over time and across different geographical contexts.

Beyond examining numerical sentiment scores, we explored word correlations to gain insight into thematic underpinnings influencing coverage tone. By extracting keywords from article titles and correlating their frequency with AvgTone, we identified linguistic patterns. Certain words emerged as indicators of increased negativity — associated, for example, with conflicts or crises — while others aligned with more positive reporting that emphasized cooperation or progress. In Figure 1, you can see our findings for words with the most negative and positive correlations in news articles.

C. Visualization Approaches

To render our findings both accessible and interpretable, we employed a diverse set of visualization techniques, mainly in Python notebooks. Time-series line plots (Figure 2) depicted how the volume and sentiment of news coverage fluctuated over the examined period, enabling us to pinpoint years or months with notably positive or negative reportage. For example, line plots tracking average sentiment by publisher and country provided a visual narrative of evolving media perspectives. Stacked bar charts (Figure 3) illustrated how coverage distribution across different countries shifted among various news sources, showing the comparative attention each actor received.

We also used heat maps and matrices extensively. Frequency matrices visually represented the co-occurrence of countries in events, while sentiment matrices depicted average tone at the intersection of particular country pairs. By integrating color gradients, these heatmaps allowed for an

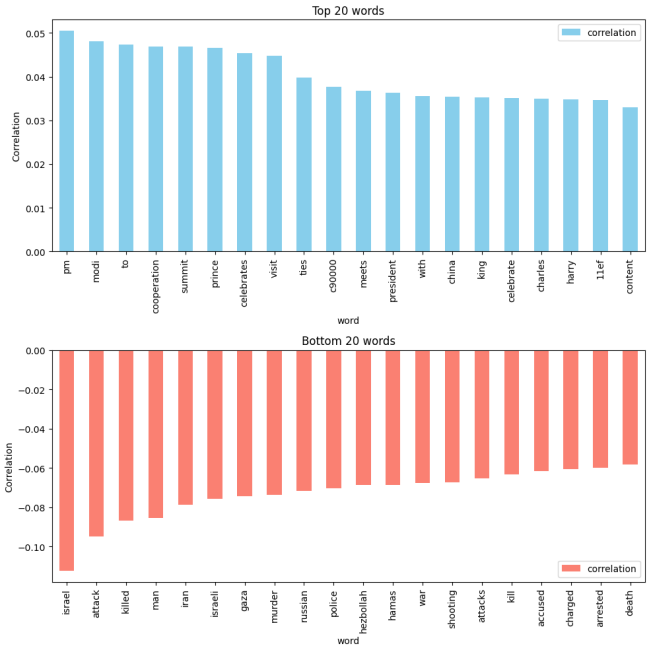

Fig. 1. Top Word Correlations

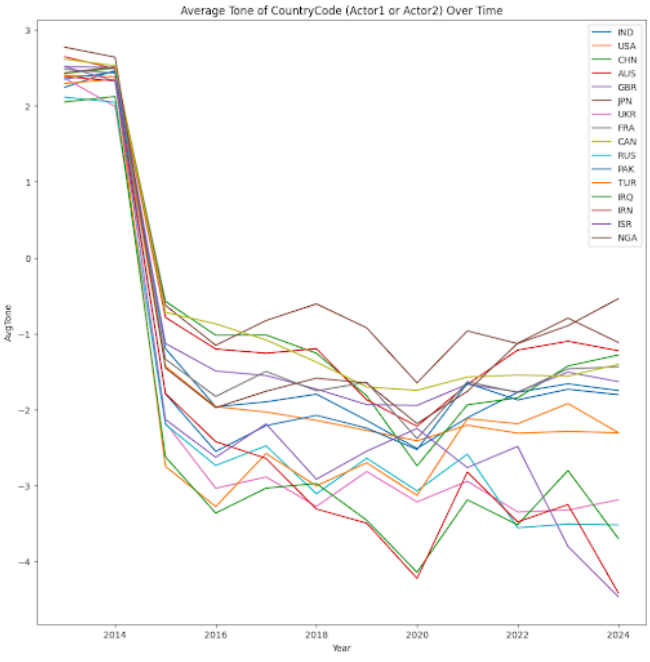

Fig. 2. AvgTone Over Time for Countries (Line Plot)

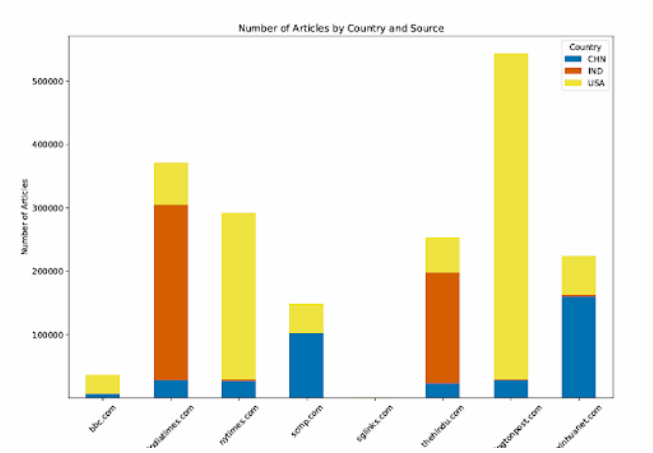

Fig. 3. Number of Articles by Source about Country

intuitive understanding of which relationships commanded frequent coverage and how that coverage skewed sentimentally. Additionally, correlation tables ranked words based on their association with tone, which could be displayed as sorted lists or bar plots. These visual tools, generated using Python's Matplotlib and Seaborn libraries, served as a foundation for telling a data-driven story, complementing textual explanations. See Figure 4 for co-occurrence frequency of different countries being reported together in the same article.

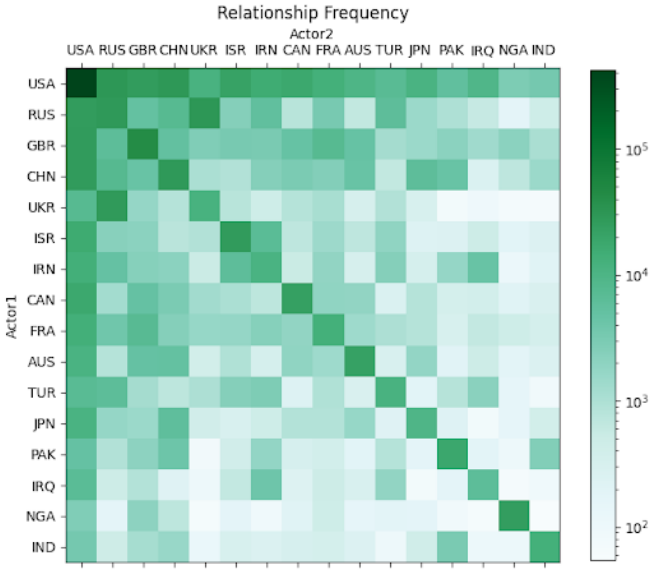

Fig. 4. Co-occurrence frequency of Actor1 and Actor2 countries in a single news article.

IV. Results

A. Sentiment and Bias Analysis

Our results revealed notable patterns of sentiment and bias. Overall, the tone of coverage displayed a gradual but consistent move toward negativity over the examined period. In particular, China's coverage of itself stood out as particularly biased, reflecting a more favorable narrative in contrast to the general trend of downward sentiment observed in other geopolitical contexts (see Figure 5). This finding suggests that some domestic media environments might selectively curate narratives to present their home country more positively than the global average tone would predict.

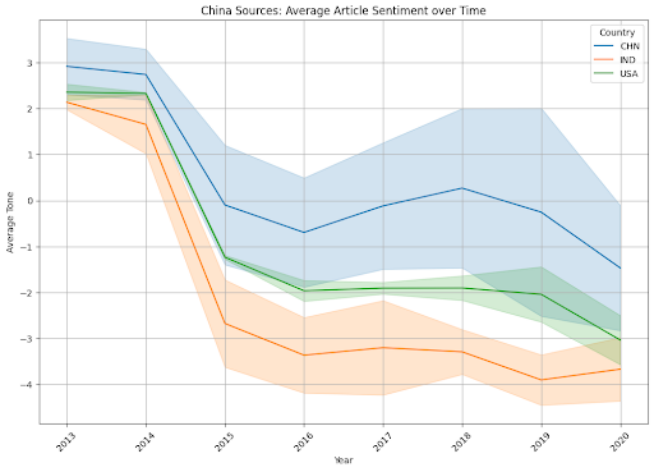

Fig. 5. Average tone of articles on China, India and USA reported by the Chinese media

We also observed that significant geopolitical events, especially wartime engagements, were marked by sudden spikes in mentions and negative sentiment. For instance, as Figure 6 shows, we observe clear sentiment dips and heightened coverage surrounding Russia and Ukraine in 2022 as well as Israel and Iran in 2024. These sharp changes in AvgTone indicate that conflict-related events catalyzed more negative language, reinforcing the idea that war and instability drive intense and often critical media narratives.

B. Comparative Media Analysis

When comparing coverage across various publishers, it became evident that the political environment and editorial freedoms of the media outlets strongly influenced narrative construction. Outlets based in less democratic countries tended to project a distinctly favorable image of their own nations. This positive bias toward self-coverage stood in stark contrast to the general negative trend in global sentiment, suggesting that state-aligned or tightly controlled media ecosystems present a more sanitized or optimistic portrayal of their home country's actions and policies.

Meanwhile, wartime events consistently dominated the global news cycle, exerting a powerful influence over both the language and sentiment of reporting. Increases in hostilities coincided with spikes in mentions of conflict-related words—terms that intrinsically carried negative connotations and intensified the prevailing mood of the coverage. Such terms clustered around violence, aggression, and crisis, reinforcing a downward shift in sentiment for all actors

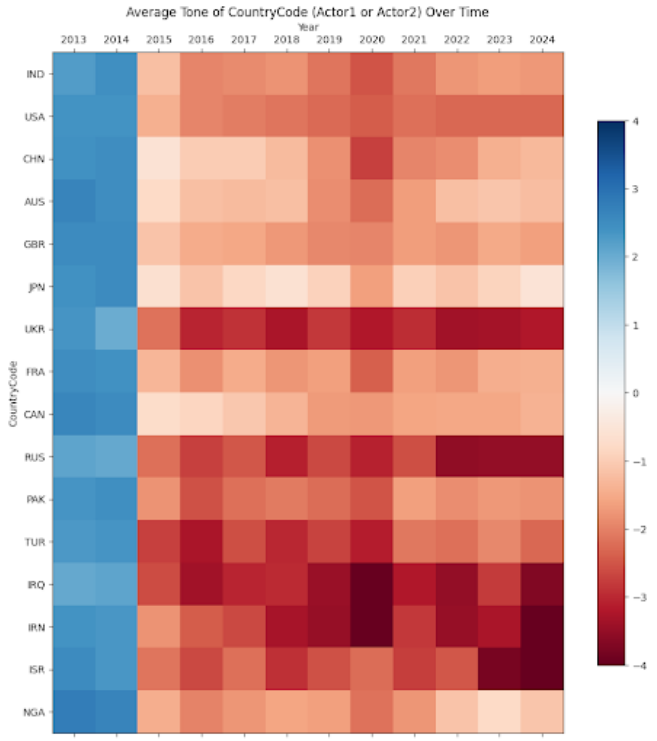

Fig. 6. Heatmap showing the average tone of articles about various countries over time

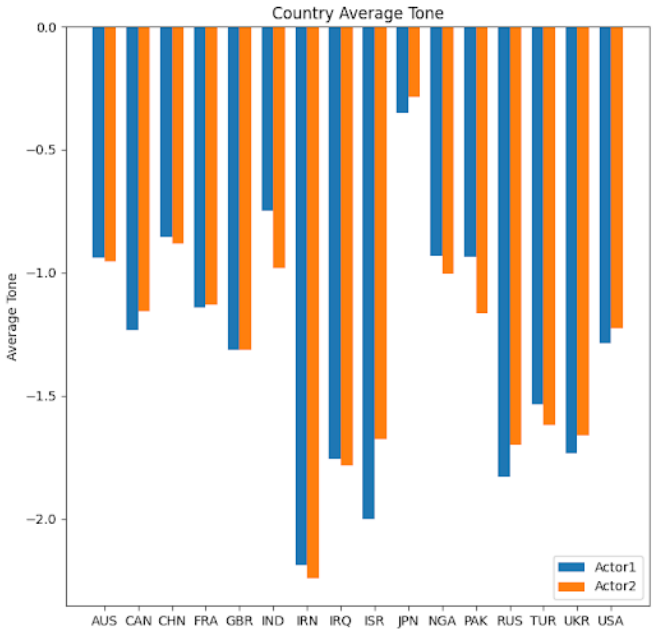

Fig. 7. Average tone of news coverage for countries as Actor1 and Actor2

involved (see Figure 7). As conflicts involving countries like Russia and Ukraine in 2022 and Israel and Iran in 2024 flared, reporting became more critical, painting a grim picture of international relations during turbulent periods. These dynamics highlight that, while certain state-controlled outlets may strive to maintain a positive self-image, the gravitational pull of wartime coverage—replete with negatively correlated keywords—ultimately shaped much of the global newscape with a distinctly negative hue. This can be seen in examining countries' spikes in media coverage during wartime and wartime countries AvgTone's shown in Figure 7.

C. Key Challenges

Throughout the analysis, we faced several challenges that shaped our approach. Managing the sheer scale of GDELT's event database demanded careful data filtering and chunk-wise processing to prevent computational bottlenecks. Standardizing data fields—such as ensuring consistent parsing of country codes and stable identification of reputable publishers—was essential for producing reliable comparisons.

Interpreting sentiment solely through the AvgTone measure had inherent limitations. While useful for broad trend detection, it could not capture the full depth of editorial choices, cultural contexts, or the subtle nuances of diplomatic language. Similarly, word correlations provided valuable but incomplete guidance; understanding the reasoning behind certain linguistic patterns would require more sophisticated natural language processing techniques or topic modeling. These expansions could further illuminate why certain

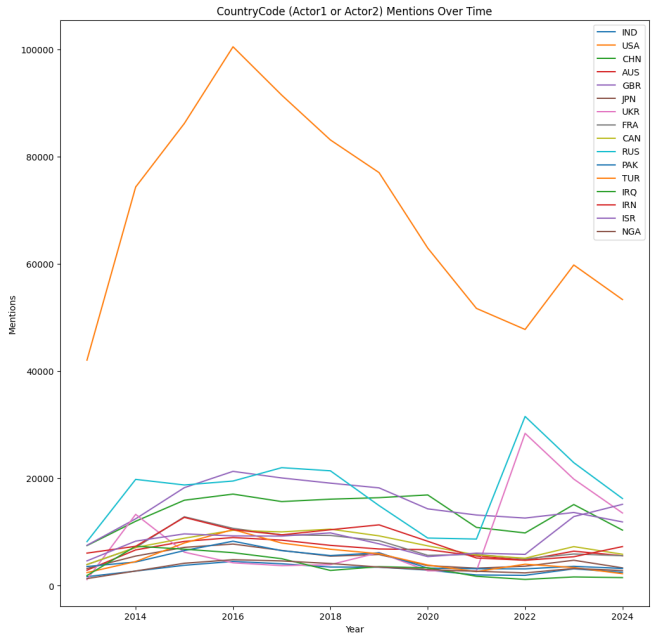

Fig. 8. Country Mentions over Time

phrases correlate strongly with negative tone during wartime or why specific actors, like China, maintain a positive image of themselves even amid global negativity.

V. Discussion

A. Implications of Findings

Our findings reveal important patterns in how media portrayals influence public perception, shape policymaking, and highlight the need for media literacy. One key observation is the strong presence of bias in state-controlled media outlets. For example, Chinese media consistently portrays the country in a more favorable light compared to global averages. These biases directly impact how people perceive their own nations and others, often boosting nationalism while potentially alienating international audiences. For global readers, understanding and identifying these biases is critical to forming a well-rounded view of world affairs.

The study also shows how the media amplifies geopolitical tensions, especially during major global events. Coverage of the Russia-Ukraine conflict in 2022 and the Israel-Iran tensions in 2024 used language heavily centered on conflict and crisis. This reporting significantly influences how nations and events are perceived, often framing them in a way that fosters hostility or fear. Policymakers should account for this slant when interpreting media narratives, ensuring they rely on a variety of perspectives to guide decision-making.

Media coverage also plays a role in escalating public focus on crises. Frequent use of terms like "war," "attack," and "death" during conflicts drives the narrative toward negativity, often overshadowing stories of diplomacy or resolution. While it is essential to report on conflicts, media outlets could strike a balance by also highlighting cooperative efforts or peaceful outcomes. This approach could help shift the global narrative toward constructive discussions and reduce the polarizing effects of constant negative coverage.

For individuals, these findings highlight the growing need for critical media literacy. Readers need tools and strategies to assess the reliability of their information sources, identify potential biases, and explore diverse viewpoints. Media literacy programs and education initiatives can help people better navigate the complexities of today's media landscape, reducing the influence of one-sided or biased reporting.

The research also points to practical applications for governments and international organizations. Governments can use these findings to assess how their actions are portrayed globally and adjust their communication strategies to improve their international image. International organizations could also use this work to identify and counter misinformation while encouraging balanced reporting. These applications could help create a media environment that better reflects diverse and accurate perspectives.

While the study provides valuable insights, it also faced challenges. Managing the large scale of the GDELT dataset required significant computational resources, and the reliance on a 2.5 percent sample may not fully capture the complexities of global media trends. The use of AvgTone as a sentiment measure, while helpful for identifying broad patterns, lacked the nuance to interpret sarcasm, cultural differences, or context-specific language. Similarly, analyzing word correlations provided useful insights into sentiment but did not fully explain the reasons behind the patterns. These limitations highlight the need for more sophisticated methods, such as advanced natural language processing and contextual analysis.

B. Future Work

Looking forward, this research could expand by incorporating more years, countries, and media outlets to give a broader perspective on global narratives. Advanced techniques like fine-grained sentiment analysis, network mapping of media relationships, and topic modeling could offer deeper insights into the forces shaping international news coverage. This work could also inspire practical tools for media literacy education, interactive dashboards for policymakers, and frameworks for journalists to evaluate biases in reporting. Future research could explore how media covers specific issues, such as climate change or human rights, and compare portrayals across traditional and digital platforms to understand the evolving role of media in the digital age.

This study shows how media portrayals can influence global perception and relations, providing a foundation for more balanced and informed approaches to reporting and decision-making. By addressing biases and fostering more diverse narratives, this work can support efforts to improve public understanding and promote fair representation on the global stage.

VI. Project Timeline

The timeline below presents a chronological overview of the key milestones and sprints for the project:

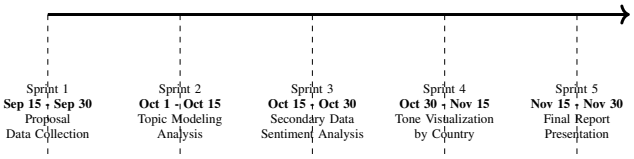

Fig. 9. Project Timeline

The Project Timeline provides an overview of the major phases and key milestones for our project. The timeline spans five sprints, each representing a distinct stage in the research and development process. In Sprint 1 (Sep 15 - Sep 30), we focused on project proposal development and initial data collection to set the foundation for the research. During Sprint 2 (Oct 1 - Oct 15), we performed topic modeling and analyzed different news sources to understand media coverage patterns.

In Sprint 3 (Oct 15 - Oct 30), we carried out secondary data collection and conducted sentiment analysis to identify general trends in news portrayals of countries. This stage allowed us to refine our dataset and enhance the robustness of our analysis. Sprint 4 (Oct 30 - Nov 15) involved visualizing

the tone of media coverage over time and by country, highlighting any noticeable biases or shifts in sentiment. Finally, in Sprint 5 (Nov 15 - Nov 30), we compiled our findings, created our final report, and prepared for the presentation.

The timeline illustrates the structured approach we adopted, ensuring each milestone built on the previous one to develop a comprehensive analysis of media portrayals. This phased strategy helped us maintain a clear focus on our objectives while allowing for systematic data exploration and insightful visualization of the results.

VII. Team Structure

In this section, we outline the roles and responsibilities of each team member and provide a visualization of the project organization.

A. Organizational Chart

Below is the organizational chart showing the structure of our project team:

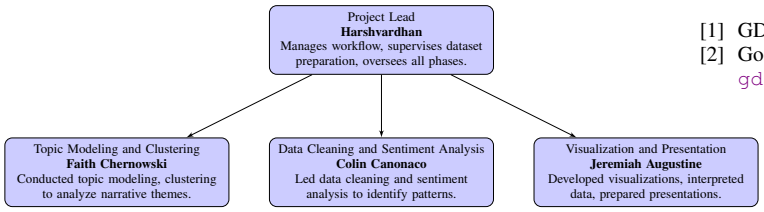

Fig. 10. Project Organizational Chart

B. Team Responsibilities

Each team member was responsible for specific aspects of the project to ensure efficient workflow and optimal use of individual expertise:

- Colin Canonaco: Led data cleaning and conducted sentiment analysis to identify patterns.
- Faith Chernowski: Handled topic modeling and clustering analysis.
- Harshvardhan: Project Lead, responsible for managing project workflow and overseeing all project phases.
- Jeremiah Augustine: Created visualizations, found word correlations, and helped interpret the results.

VIII. Conclusion

This study examines how international media shapes global perceptions through sentiment analysis, bias detection, and topic modeling. Using the GDELT dataset, we identified patterns in how different countries are portrayed, revealing significant disparities and biases in reporting. Our analysis highlights the strong influence of state-controlled media in shaping narratives that favor their home countries, as seen in the case of Chinese outlets, and the media's tendency to amplify negativity during geopolitical conflicts.

The findings emphasize the importance of critically assessing media content, both for the public and policymakers. Media biases and the focus on conflict-driven narratives can shape opinions, influence diplomatic decisions, and even escalate tensions. Recognizing these trends allows readers, leaders, and organizations to approach media with a more informed and balanced perspective.

While the study provided valuable insights, challenges such as the dataset's size and limitations in sentiment analysis pointed to areas for improvement. Future work can address these gaps by using more advanced techniques and broader datasets to capture a fuller picture of global media trends. Expanding the scope to include platform-specific analysis, cultural contexts, and issue-specific portrayals, such as human rights or climate change, could further deepen our understanding of how media narratives shape the world.

In a rapidly evolving media landscape, this research lays the groundwork for more transparent and equitable reporting. By identifying biases and highlighting opportunities for more balanced storytelling, this work contributes to ongoing efforts to improve media literacy, foster fair representation, and encourage constructive global dialogue.

References

[1] GDELT Project, https://www.gdeltproject.org/.
[2] Google BigQuery, https://tinyurl.com/gdelt-big-query.