Countries in News: Project Proposal

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

***Abstract*— With this project, we will use GDELT to investigate the portrayal of international events in the media based on information about the involved actors and publishers.**

I. Introduction

A. Objective

The primary objective of this project is to conduct a comprehensive analysis of the portrayal of countries in popular international media using sentiment analysis, topic modeling, and bias detection. By leveraging data from the Global Database of Events, Language, and Tone (GDELT), this project aims to assess how countries such as the USA, UK, India, China, and Russia are represented across various global news outlets. The analysis will provide insights into the underlying narratives, media biases, and the general tone of coverage, ultimately identifying patterns that influence public perception and international relations.

B. Motivation

In today's globalized world, international media plays a critical role in shaping public opinion and national images. Understanding how countries are portrayed in the media can reveal implicit biases and highlight disparities in representation, which can, in turn, affect international relations, policymaking, and cross-cultural understanding. This project seeks to contribute to ongoing efforts in media analysis by applying advanced computational techniques to extract meaningful insights from vast amounts of news data. The team is motivated by the potential to uncover significant findings that may inform journalists, policymakers, and researchers about the role of media in international perception and the dissemination of narratives. Additionally, the results of this study could foster better awareness of the ethical responsibilities media outlets have in portraying countries fairly and accurately.

II. Dataset

A. GDELT Dataset

The GDELT 1.0 dataset (Global Database of Events, Language, and Tone) is a comprehensive, open-access database that tracks global events and their associated metadata in real-time.[1] Covering events from 1979 to the present, GDELT 1.0 records interactions between individuals, organizations, and countries as reported in news media from across the globe. Each event has multiple attributes that code for various traits of the event and actors involved. Countries, organizations, and people involved with an event are provided, and religious and ethnic associations are also included using CAMEO taxonomy. This information will be extremely valuable for finding correlations in news coverage that aren't restricted to specific people, organizations, or countries. Religion and ethnicity labels can be used to investigate relationships between cultures and broad regions. Each event also includes a score on the Goldstein Scale to quantify the predicted effect of the event on a country's stability, which may help quantify the significance of events. Another included metric that could be used to determine importance and to analyze relationships is the tone score, which measures the tone of the articles that mention the event. GDELT 1.0 enables large-scale analysis of global political, social, and economic trends, making it an invaluable resource for international relations research, conflict studies, and media analysis.

B. BigQuery Extraction and Summary Statistics

As the complete dataset is huge, we will restrict our analysis to a subset of articles downloaded from Google BigQuery, which officially supports the GDELT dataset.[2] We will restrict our analysis to articles published between 2010 and 2020 because data before 2010 is sparse. We will then look for articles published from following major publishers worldwide: *The New York Times* and *The Washington Post* from United States; the *British Broadcasting Corporation (BBC)* from United Kingdom; *The Hindu* and *The Times of India* from India; *Xinhua* and the *South China Morning Post* from China. All of these media agencies are privately owned except for the *BBC* and *Xinhua*, which are state-owned. They were chosen based on their popularity in their respective locality. This results in a total of 1,868,683 articles. In later analysis, we may choose to take a random sample of this population.

Table I presents the number of articles published by each news source. Figure 1 presents news articles by each country and publisher.

C. Attributes or Features Used

In this analysis, we will utilize several key variables from the GDELT 1.0 dataset to examine global events involving the United States, India, and China. The `SQLDATE` field,

*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

[1]https://www.gdeltproject.org/

[2]https://tinyurl.com/gdelt-big-query

TABLE I

NUMBER OF ARTICLES PUBLISHED BY SOURCE

Publisher	Publisher Country	# of Articles
The Washington Post	USA	543,967
The Times of India	India	371,609
The New York Times	USA	291,462
The Hindu	India	252,163
Xinhua	China	224,152
South China Morning Post	China	148,051
BBC	UK	36,653

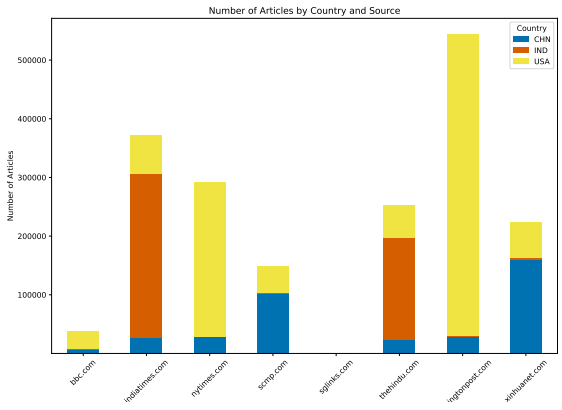

Fig. 1. Subject countries published by each news source.

representing the date of the event in `YYYYMMDD` format, enables us to filter events within a specified time frame. The `Actor1CountryCode` field captures the country affiliations of the primary actors involved in the events, allowing us to focus on specific countries. The `Goldstein Scale` provides a numeric measure, ranging from -10 to +10, which indicates the event's potential impact on the stability of the countries involved, with higher positive values signifying more stabilizing actions. Additionally, the `AvgTone` field reflects the average sentiment of the media coverage, with values ranging from -100 (extremely negative) to +100 (extremely positive), helping us gauge the overall tone of reported events. Finally, the `SOURCEURL` field identifies the news source reporting the event, allowing us to restrict the analysis to reputable publishers and assess the credibility of the data.

III. MILESTONES AND EXPECTED OUTCOME

This project is expected to yield several key outcomes that will provide valuable insights into media portrayal and sentiment across different countries. Specifically, we anticipate the following results:

1. Sentiment and Bias Analysis: A detailed report on the sentiment associated with news coverage of each country (USA, UK, India, China, and Russia), including positive, negative, or neutral trends. This analysis will highlight the presence of potential media biases, revealing patterns that may favor or marginalize specific nations.

2. Topic Modeling and Narrative Themes: Identification of the dominant topics and narratives in international media coverage for each country. This will provide insights into which issues (e.g., economy, politics, military) are most frequently associated with each nation and how they are framed across different regions.

3. Comparative Media Analysis: A comparative evaluation of how different media outlets and countries portray the same events or issues. This will offer a global perspective on divergent or convergent reporting styles and provide evidence of international media polarization.

4. Public Awareness and Policy Implications: A set of recommendations for media practitioners, policymakers, and researchers to enhance the fairness, accuracy, and diversity of country portrayals in news reporting. This includes suggestions for reducing biases in media coverage and encouraging more balanced international reporting.

Ultimately, this project aims to contribute to a broader understanding of the role media plays in shaping perceptions of countries, offering both academic and practical insights that could influence future media practices and international relations.

IV. TIMELINE AND MEMBER RESPONSIBILITIES

This project's success relies on each team member's collaborative efforts, with responsibilities divided across critical steps in the analysis.

The first step involves Colin and Harsh exploring, investigating, and cleaning the GDELT dataset to extract relevant, trustworthy, and consistent information. This complex task will require handling and cutting a vast amount of data, as described above, into more manageable and well-sourced information. We have attempted to represent the media across multiple vital countries accurately. This will help us retain important context and not skew analysis.

Once the data is ready, two primary analyses will be conducted. One will be Jeremiah, assisted by Harsh, performing the sentiment analysis to evaluate the tone of news and how it has shifted over time. This task will also present challenges in developing a system to manage regional and linguistic differences while ensuring an accurate reading of the sentiment expression. Another primary analysis will occur during this phase as Faith, with the help of Harsh, will conduct the topic modeling to identify common themes associated with selected countries. One of the significant challenges will be ensuring major economic and political topics are identified, as well as some subtler topics and narratives.

Finally, the team will collaborate to identify potential biases in different media's portrayal of countries. This step involves reviewing previous analyses and conducting possible extra research to help extract meaning from the data. It will be essential for all team members to understand each step of the process at this stage and thoroughly explain each of our findings. By structuring and planning these responsibilities ahead of time, the team hopes to ensure a smooth and predictable workflow, with each task building on the previous, leading to a comprehensive and unbiased analysis of the data.