PDRI | Penn Development Research Initiative   DevLab@Penn

High Frequency Tracking of Civic Space Utilizing Domestic News Scraping and Large Language Model Classification

Donald A. Moratz[1], Jeremy Springman[1], Serkant Adiguzel[2], Zung-Ru Lin[1], Diego Romero[3], Hanling Su[1], Jitender Swami[1], Rethis Togbedji Gansey[1], Mateo Villamizar-Chaparro[4], Erik Wibbels[1],

University of Pennsylvania[1], Sabanci University[2], Utah State University[3], Duke University[4]

Objective

Changes in politics and society - like protest movements, waves of arrests, corruption scandals, and legal changes- often happen quickly. Unfortunately, many of the standard indicators that researchers have to track civic space or regime characteristics are measured annually. In recent years, several big data approaches to measuring features of domestic and international politics at higher frequency have emerged, including GDELT and POLECAT. We present some of the advantages of tracking civic space using the Machine Learning for Peace (MLP) infrastructure. We apply customized scrapers to a curated sample of *national news sources* to provide a *monthly* measures of civic space activity. To process these articles, we apply large language model classification to track 20 different civic space event types.[a] In this report, we provide a clear use case by comparing national and international news coverage of civic space events in 62 countries, highlighting the shortcomings of data sources that rely solely on international coverage.

The Challenge

The overarching challenge to analyzing civic space is that the most common and careful measures, such as V-DEM, report on an annual basis. While those measures are useful for tracking overarching regime dynamics, they provide little information on the day-to-day political struggles that characterize the current era of democratic backsliding. The combination of big data and machine learning offer a solution. However, we show that any such attempt must overcome two challenges. First, international media does not provide an accurate picture of key civic space events in many countries. International sources provide sparse and inconsistent coverage of even highly salient events in many countries. Thus, big data projects and policy makers relying only on international news coverage garner a biased picture of civic space events as they unfold on the ground. Second, constructing a corpus of domestic news from countries around the world requires a great deal of human curation; improper scraping can introduce enormous error into data.

Our Approach

To track changes in civic space, we scrape news articles from several of the most widely-read national media sources for each of the 62 countries in our data. As a result, our corpus includes articles published in nearly 40 languages. Many domestic media outlets exhibit sudden shifts in publication volume and poor website architecture, requiring careful human monitoring and customized scraping and parsing to consistently capture all published articles. Distinguishing true changes to publishing rates from technical challenges in scraping and parsing articles precludes exclusive reliance on prepackaged web scrapers. As a result, our approach provides a much richer, more detailed picture of civic space in each country.

Findings

We have two key findings. First, data collection from national news sources requires careful human curation. We show that reliance on prepackaged scrapers like GDELT, Common Crawl, and the Internet Archive produce poorer coverage than our approach and introduce errors in the data. For example, we compared MLP's coverage of three Bandgladeshi news sources with that of GDELT and Internet Archive. MLP's coverage begins in 2013 for one source and in 2015 for the other two sources. By comparison, GDELT only has coverage from 2019 forward. Our process also generates much better coverage across all sources. For GDELT's best covered source, GDELT averages 2,100 articles per month compared to 2,500 per month from MLP.[a] Internet Archive performs even worse. The disorganized nature of Internet Archive means that it took nearly two weeks to collect URLs from a single source from 2019-2023, and the results included numerous irrelevant, broken, and duplicate links. Although Internet Archive delivered a large number of URLs, less than half were usable.

To further demonstrate the importance of human curation, Figure 1 shows two examples of coverage changes that can occur in web-scraped media. On the left, Figure 1a shows a spike in ghanaweb.com, a large media outlet in Ghana. The dramatic increase in publications in 2020 was related to a grant the outlet received from Google that allowed it to massively expand its coverage. On the right, Figure 1b shows a spike in lusakatimes.com in Zambia. This spike was driven by one article that the website mistakenly uploaded over 1.5 million times (each with a unique URL, making this error difficult to detect). Only careful human curation can distinguish between genuine (Ghana) and artificial (Zambia) changes in publication volume, and effectively guard against such risks to data quality.For more information about the process by which we curate our data, see this report.

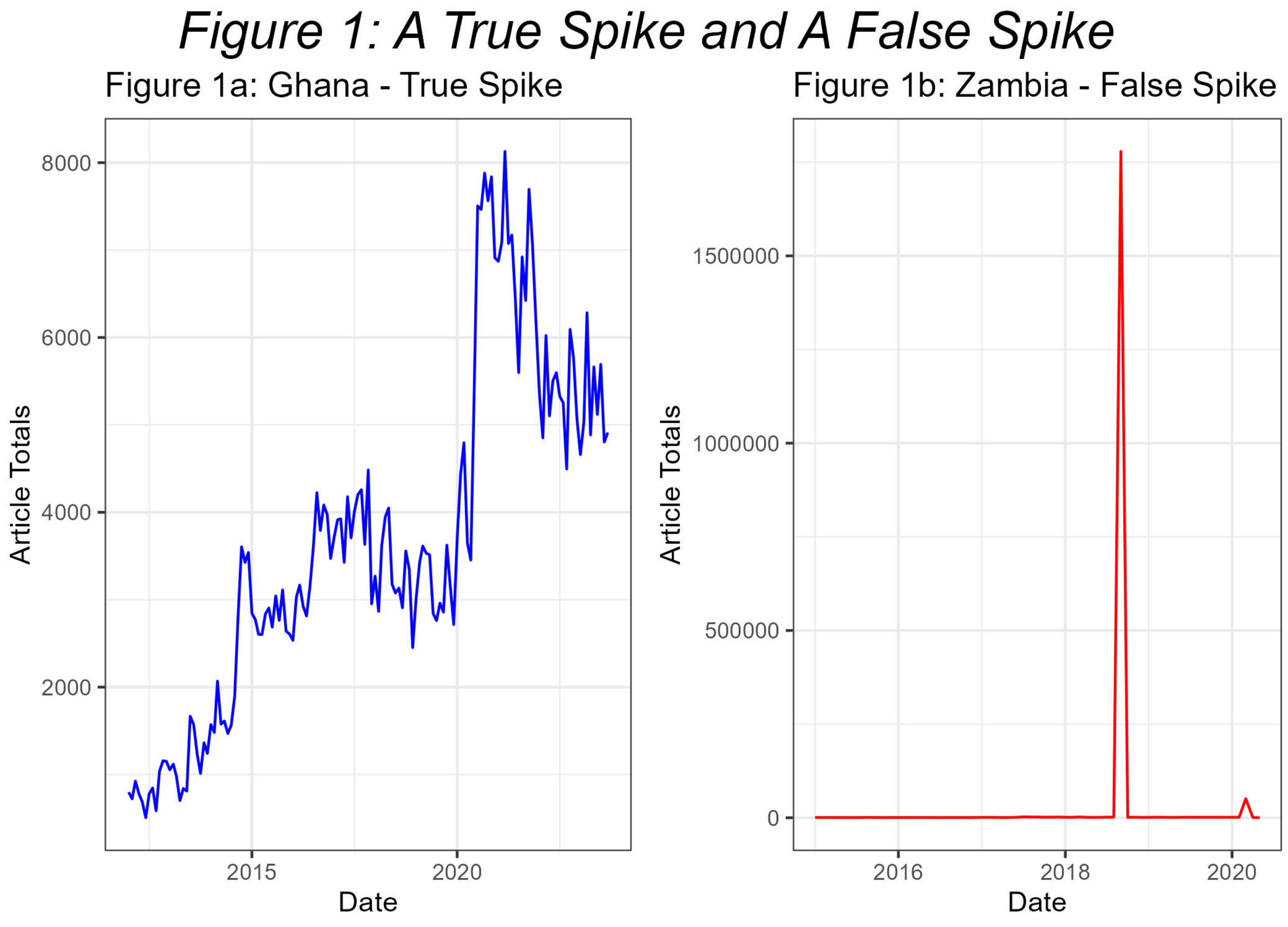

Figure 1: Changes in the volume of articles across two sources.

[a] See the Dimensions of Civic Space Codebook for definitions of all 20 event categories.

[a] GDELT also includes many broken links, redirects, duplicate articles, and advertising, all of which are fixed in our dataset. Additionally, GDELT restricts requests to one search every 5 seconds, so that scraping even a single source for the full time-period can take several days.

Our second key finding is that reliance on international sources–as is the norm with other big data projects–provides an incomplete and biased picture of events on the ground. We show that the correlation between international and domestic media coverage of civic space events is often low. Furthermore, this correlation is not driven by the extent of international reporting a country receives. It is also not the case that civic space events that are more frequently covered in international media correlate better with domestic reporting. We also compare the domestic media environment to coverage provided by international sources and find that there are significant differences in reporting.

In Figure 2, the blue bars indicate the share of our articles in the MLP corpus that are scraped from national sources. Clearly, the vast majority of our data comes from national (rather than international) sources. The dots in the figure show the correlation between the incidence of reporting on each type of civic space event. If international and national sources are reporting on the same events, albeit in different volumes, these correlations would be high. Yet the mean correlation across event types is only .23, and is only .51 for the most similarly reported category, election activity. This suggests fundamental differences in the type of events covered by domestic and international sources.

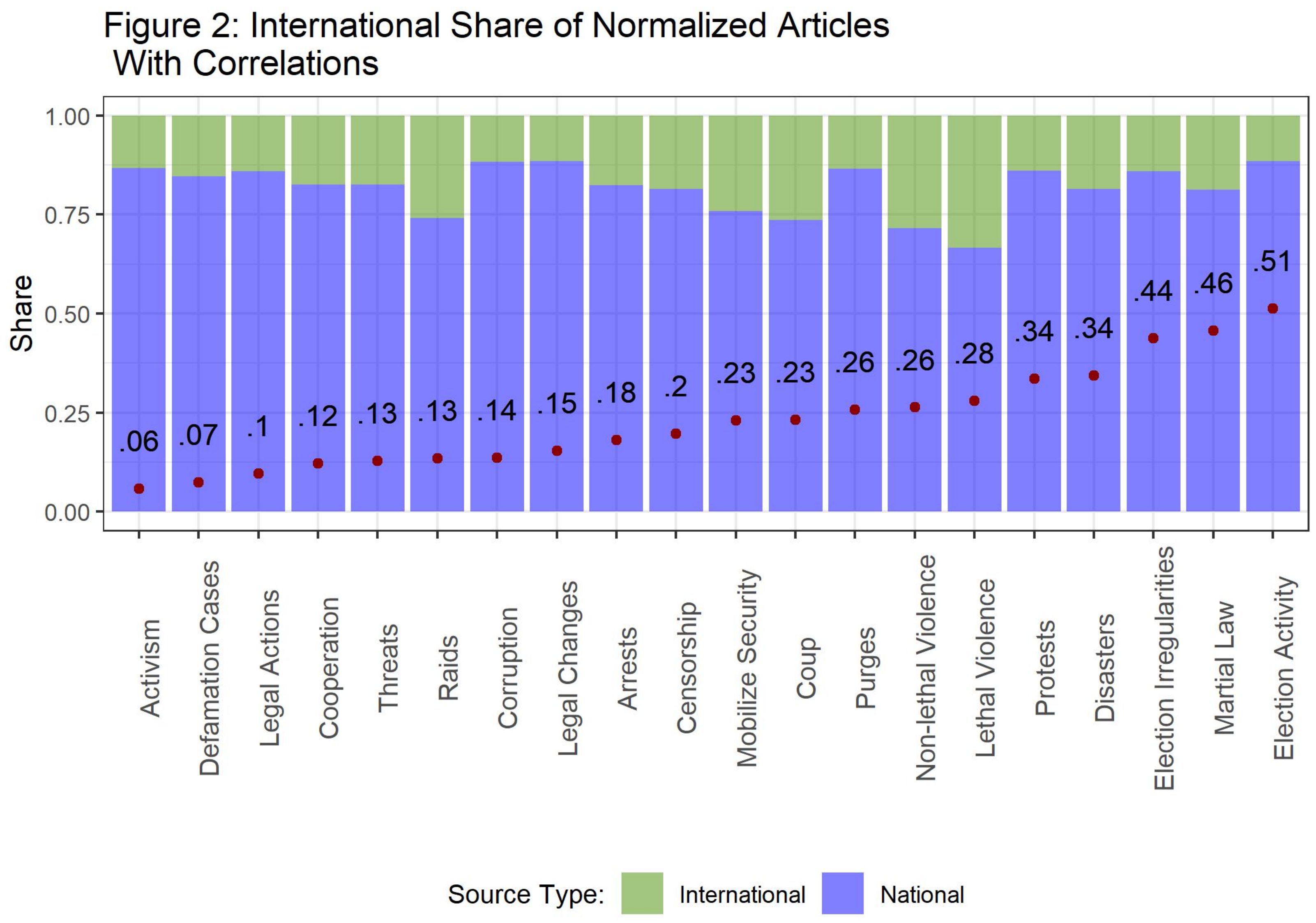

Figure 2: This figure shows the share of reporting on event types from international and national sources, and the correlation for each event type across all countries.

We also examine the rate at which different event types are covered by national and international news. Figure 3 shows event types sorted by the extent of international coverage, with national coverage stacked on top. We see that there are big differences in the types of events that are covered in international media compared to domestic media. International media is far more likely to cover acts of violence, whereas domestic media is far more likely to cover legal actions, election activity, corruption, protests, and cooperation. To the extent these are key features of civic space dynamics, reliance on international news will produce a heavily biased view of civic space.

Finally, we examine whether these trends are driven by the volume of international coverage a country receives. It is undeniably the case that international media covers some countries more extensively than others, and it could be that international coverage of civic space is more complete in countries with more coverage. Figure 4 shows the volume of international articles published about a country (x-axis) against the correlation in civic space coverage of events between international and national sources. The figure shows that while there is a slightly positive relationship between the volume of coverage and the correlation between domestic and international reporting, the correlation is low across all countries. For instance, in Turkey and India, two countries that receive a lot of international coverage, the correlation between national and international coverage is below 0.5.

There is almost no coverage of Timor Leste, and what little coverage exists has almost no correlation with what the domestic press is reporting.

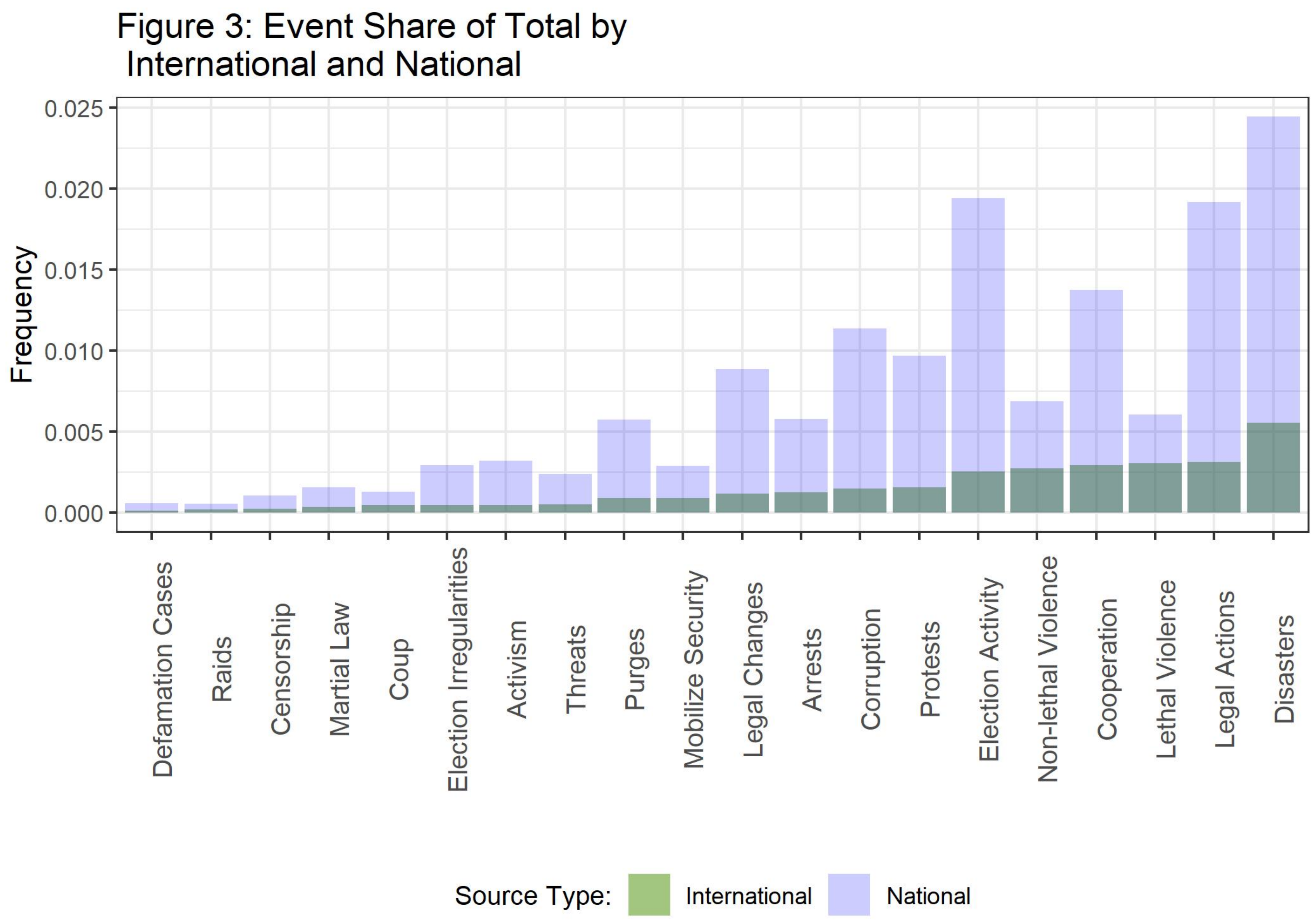

Figure 3: This figure shows the share of reporting of an event type from international and domestic coverage. It also displays the average correlation for each event type across all countries.

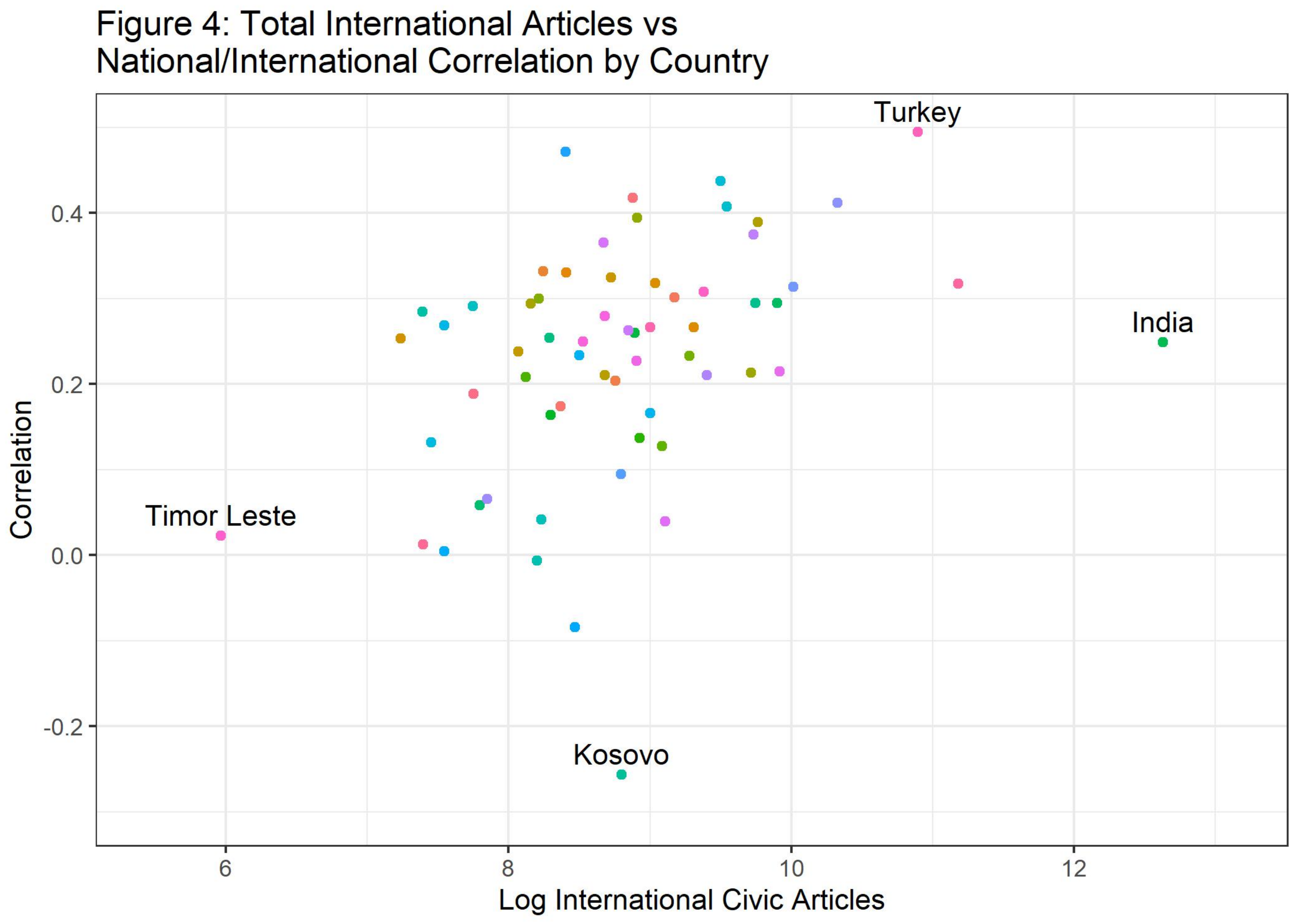

Figure 4: This figure shows the share of reporting of an event type from international and domestic coverage. It also displays the average correlation for each event type across all countries.

Policy Implications

Making US foreign policy decision-making more data-driven has tremendous potential to improve outcomes. Using media data to track changing political conditions in strategically important countries is critical to this goal. However, this report shows that policymakers should be cautious when using data that is overly reliant on international media or collected using automated tools. MLP combines direct, human-supervised data collect to ensure data quality with new tools, such as large language models and machine translation, to leverage much richer coverage from domestic outlets based in developing countries. Information sourced from media is one of the primary sources of information for the US government. Future investments in media data production should incorporate these insights into their design.

Acknowledgements

This study is funded under the Swift Expertise and Grounded Analytics Task Order by the United States Agency for International Development Democracy, Human Rights, and Governance Bureau.