Tracking Civic Space in Developing Countries with a High-Quality Corpus of Domestic Media and Transformer Models*

Donald A. Moratz[†,1,2] Jeremy Springman[†,1,2] Erik Wibbels[†,1,2]
Serkant Adiguzel[3] Mateo Villamizar-Chaparro[4] Zung-Ru Lin[1]
Diego Romero[5] Mahda Soltani[6] Hanling Su[1] Jitender Swami[7]

July 18, 2025

Civic space - the fundamental freedoms necessary for citizens to influence politics - is under constant contestation. Despite the importance of day-to-day contestation over these rights, there is very little data allowing us to study the events and processes that constitute this struggle. We introduce new data that captures civic space activity across 65 developing countries from 2012 to 2024. Using an original corpus of over 120 million articles from nearly 350 high-quality domestic media outlets and 30 international and regional outlets, we use human-supervised web scraping and open-source computational tools to track monthly variation in media attention across 20 civic space events. Our approach achieves unprecedented coverage of reporting by developing country media outlets, addressing biases in other media event data. We demonstrate the utility of this data for identifying and forecasting major political events and discuss applications for research on regime dynamics during a time of democratic backsliding.

[†] These authors contributed equally to this work.

[1] PDRI-DevLab, University of Pennsylvania
[2] Department of Political Science, University of Pennsylvania
[3] Sabanci University, Turkiye
[4] Universidad Católica del Uruguay, Uruguay
[5] Utah State University
[6] Stanford University
[7] Temple University

*This study was funded by the United States Agency for International Development (USAID) Bureau for Democracy, Human Rights, and Governance and the Open Society Foundations. We would like to thank many partners in the NGO and policy world who have helped in the development of this work, including Laura McKechnie, Dan Spealman, Asta Zinbo, Daniel Sabet, Erin McCarthy, and David Jacobstein. We also thank several researchers who were instrumental in the origins of this project, including Scott de Marchi and Spencer Dorsey, and a number of others who made critical contributions along the way, including Rethis Togbedji Gansey, Andreas Beger, Tim McDade, Akanksha Bhattacharyya, and Joan Timoneda.

Introduction

In 2016, 3.5 billion people lived under autocracy; by 2021, this number surged to over 5.4 billion (Boese-Schlosser et al. 2022). Concentrated in the global south, this "third wave of autocratization" is constricting civic space and limiting the ability of citizens to advocate for better governance (Lührmann and Lindberg 2019; Waldner and Lust 2018).[1] Nevertheless, citizens around the world continue to challenge these authoritarian movements.

Despite the importance of this day-to-day push-and-pull over political liberties and state control, data to study the events and processes that constitute this struggle is limited. Existing measures of civic space come largely from annual, expert-coded indicators classifying the nature of political regimes. While these regime indices have opened-up new domains of research to rigorous investigation, they are not designed to provide insight into the quotidian politics where battles over civic space take place.

This article introduces the Machine Learning for Peace (ML4P) dataset, which provides monthly data on 20 civic space events across 65 developing countries from January 2012 through December 2024. *ML4P* measures civic space activity by capturing monthly variation in levels of media attention across 20 civic events, providing a dynamic view of where and when civic space events are happening and their level of political salience. *ML4P* represents a major advance in our ability to understand civic space dynamics by providing a higher-frequency measure of a broad range of events bearing on civic space.

Importantly, *ML4P* is constructed from articles collected by the High-Quality Media from Aid Receiving Countries (*HQMARC*) corpus. *HQMARC* is an original corpus of articles scraped from 348 prominent *domestic* media outlets based across our sample of 65 countries and publishing in 36 languages. We supplement these domestic outlets with content scraped from 12 regional and 15 global outlets (henceforth, we refer to the combination of regional and global outsets as "international"). In sharp contrast to many other sources of event data, more than 95% of the articles in *HQMARC* are scraped from domestic media outlets based in the countries covered by our dataset.

HQMARC employs a human-supervised, source-specific scraping methodology that prioritizes data quality and comprehensiveness over the broad but shallow coverage typical of automated web crawlers. Rather than deploying generic crawling tools that blindly follow hyperlinks across domains, *HQMARC* develops customized harvesting workflows for each media source. This targeted approach enables comprehensive or near-comprehensive capture of all relevant articles published by each source from 2012 through 2024. This process proves particularly valuable for domestic news sources, whose websites are less stable than international outlets.

Our data collection yields significant advantages over both "big data" media repositories like GDELT, Internet Archive, and Common Crawl and expensive commercial databases like Factive and LexisNexis, delivering a stable, well-understood corpus composition with superior linguistic diversity and coverage of high-quality developing-country sources. However, *HQMARC's* size and linguistic diversity makes human classification prohibitively expensive. To produce *ML4P's* structured data on civic space events, we apply free, open-source computational tools to translate and

[1] Following Brechenmacher and Carothers (2019), we define civic space as the fundamental freedoms that allow people to gather, communicate, and take part in groups to influence society and politics.

extract information from each article, identifying the main event being reported on and the country in which the event occurs.

This paper proceeds in six parts. In Section , we discuss how *ML4P* complements existing data on regimes and opens up new avenues for research. In Section , we describe the data production process and the advantages of our approach. We compare the coverage of *HQMARC* to other major media repositories, showing superior accuracy and stability compared to alternatives. Section presents results from exploratory analyses and validation exercises. First, we compare event coverage from international and regional sources to that of domestic sources, finding systematic differences in the types of events being covered and documenting the frequency with which significant events are ignored by non-domestic sources. Our findings have implications for event data that is generated from predominantly international or regional media outlets, a common practice in the social sciences. Second, we conduct an AI-assisted qualitative audit of major events detected across six developing countries. We confirm that events detected by *ML4P* correspond to real events in the world. This measures *ML4P's* ability to generate true positives. Finally, we assess false negatives by conducting a series of case studies to identify major political events in recent years, demonstrating that *ML4P* consistently captures these major events.

In Section we provide a use case, documenting how civic space events are predictive of independently measured instances of severe political instability. Section discusses limitations. Section concludes, discussing why *ML4P* represents a valuable new resource and recommending future applications for research on autocratization and democratic backsliding, political accountability and contentious politics, media behavior, crisis response, and program evaluation.

Democratic Erosion, Annual Indices, and the Need for Civic Space Data

The "third wave of autocratization" has brought renewed attention to the study of regime type, political transitions, and democratic backsliding (Lührmann and Lindberg 2019). This attention has been accompanied by a proliferation of measures of regime type, including the Varieties of Democracy project (Coppedge et al. 2023), the Civil Society Organization Sustainability Index (U.S. Agency for International Development 2022) and the World Justice Project's Rule of Law Index (World Justice Project 2024), among many others. These indices are designed to provide information about levels of democracy over time and space and to capture distinct features of regimes, ranging from freedom of the press, rule of law, the ease of civic organizing and beyond. Though VDEM, in particular, has helped make annual indices more rigorous, they are not designed to provide insight into the quotidian politics where battles over civic space take place.

Ultimately, these slow-moving changes in the nature of regimes are the result of specific actions and events occurring at specific moments in time. Existing measures attempt to capture the cumulative impact of these actions and events over 12-month periods. Our project complements those efforts by tracking the events – often occurring over days or weeks – that contribute to broad, prolonged processes of changes captured by annual indices. Take, for instance, Hungary's systematic dismantling of democratic institutions since 2010. Each stage of this process involved a host of important events: the 2010 media law that brought most outlets under government control, the 2011 constitutional changes that packed the Constitutional Court, the 2012 electoral law that gerrymandered districts, the 2017 targeting of Central European University, and the

2018 "Stop Soros" laws criminalizing aid to asylum seekers. While each of these individual events was meaningful for Hungary's democratic trajectory, annual democracy indices smooth over these discrete moments of institutional change, obscuring the specific mechanisms and timing through which democratic backsliding actually occurs. *ML4P* is designed to shift analytical focus away from these slow-moving summary indices and toward the fast-paced civic space events underlying broader changes in the nature of countries' political governance.

Several existing event data projects produce high-quality data bearing on civic space. Among the most notable are the Armed Conflict Location Event Data Project (ACLED; Raleigh et al. (2010)), the Uppsala Conflict Data Project Georeferenced Event Dataset (UCDP GED; sundberg2013introducing), the Political Event Classification, Attributes, and Types (POLECAT; Halterman et al. (2023)) dataset (formerly the Integrated Crisis Early Warning System dataset; Boschee, Natarajan, and Weischedel (2012)), and the Global Database of Events, Language, and Tone (GDELT; Leetaru and Schrodt (2013)). While each of these datasets have advanced social science research, each is limited in their ability to drive research on civic space.

ACLED and UCDP are focused on violence and protest, and thus cover a narrow slice of the events that signal changes in or struggles over civic space. Alternatively, GDELT relies on the Conflict and Mediation Event Observations (CAMEO) coding ontology, which covers a broad range of events but focuses on inter-state disputes and strategic interactions (P. A. Schrodt, Gerner, and Yilmaz 2012). While CAMEO classifies events using an overly complex and rigid system and uses actor dictionaries that have limited coverage and are out of date, POLECAT relies on the PLOVER ontology, which is designed to capture similar events to CAMEO but with more flexibility (Halterman et al. 2023).

ML4P is the first event data source focused specifically on events that bear on civic space. While we have a rich body of theory about 'regimes', the literature on 'civic space' and 'civil society' is spread across varied bodies of work on protest, social capital, legal studies, and election studies. Our solution is to collect data on a broad range of events that expand, contract, or contest civic space. In consultation with academic research and policy practitioners, we define 20 civic space event types, ranging from political arrests and censorship to corruption, legal actions, and legal changes (see Table 1 for a complete list of event categories). We also code articles reporting on disasters and election activity, given the propensity of aspiring autocrats to use these events as justification for restricting civil society. While some of the events we cover, such as protests and lethal violence, are subject to systematic data collection elsewhere, *ML4P* produces the first systematic data on many of these events. Together, these event types provide a rich picture of the monthly contest over civic space and offer the potential for new research on civic dynamics.

These limitations of existing data sources underscore the need for ML4P's approach to civic space measurement. By capturing 20 distinct civic space events at monthly frequency across 65 developing countries, ML4P fills critical gaps in our understanding of the day-to-day political dynamics that drive broader regime changes. Unlike annual indices that smooth over discrete moments of institutional change, or event datasets that focus narrowly on conflict, *ML4P* provides comprehensive coverage of a large spectrum of civic space activities. This approach enables researchers to study not just the outcomes of democratic backsliding, but the specific mechanisms and temporal dynamics through which civic space is contested and transformed. Furthermore, our approach allows researchers to measure the salience of these events in domestic media and allows for future changes to our coding criteria to be quickly applied to the entire corpus.

Constructing ML4P

Social scientists rely heavily on media reports to produce event data (P. Schrodt and Yonamine 2013). While the shortcomings of this approach are well documented Earl et al. (2004), monitoring media reports is the best means available to track the occurrence of many events across a wide range of national contexts. Evidence suggests that access to traditional media effectively increases citizen knowledge of major events and government behavior, even in repressive political environments (Besley and Burgess 2002; Arendt 2024). While platforms like radio and social media are important, they often rely on content originally produced by traditional news outlets (Quartey et al. 2023; Study of Journalism 2019), which are generally more trusted (Fotopoulos 2023; Bridges 2019) and provide more comprehensive coverage of political events (Lee, Diehl, and Valenzuela 2022; Schäfer and Schemer 2024).

Historically, efforts to create event data from media have faced two persistent challenges. First, extracting information from unstructured text has traditionally required human coders fluent in relevant languages, which limited the volume of material that could be processed and often introduced lags between an event's occurrence and its inclusion in datasets. This manual approach also made it expensive to revise category definitions or coding procedures in a way that maintains backward compatibility of the data (ACLED 2023). To maintain backward compatibility, any significant changes to what or how information is extracted from text would require a human to re-code every previously classified article. Fortunately, recent advances in machine learning have made it possible to automate coding quickly and cheaply while maintaining accuracy. While the debate around the relative accuracy of human and machine coding are ongoing, many teams have reported accuracy levels in predicting human labels that rival or exceed that of humans across diverse domains Mueller and Rauh (2018).

Second, reliable repositories of high-quality media reports are surprisingly difficult to create and remain extremely rare. Many prominent political event datasets rely heavily on data sourced from international and regional rather than country-specific domestic sources, which is reflected in their limited linguistic diversity (Raleigh, Kishi, and Linke 2023). Furthermore, this sourcing is often done by private media aggregators, such as Factiva and Lexis Nexis, that provide limited information about how they select sources, collect data, or identify which news articles are politically relevant. For example, POLECAT uses Factiva to source politically relevant news stories from over 1,000 sources, but these sources publish in only seven languages and there is no human curation of which sources to include (Halterman et al. 2023). Similarly, our analysis of all sources available from the Lexis Nexis University archive shows that for six *ML4P* countries, Lexis Nexis has zero local sources, and across the *ML4P* countries where Lexis Nexis has at least one local source, their sources publish in 17 languages compared to *ML4P's* 34. As we show in Section , the relative emphasis on different types of events covered by international and regional outlets is systematically different than that of domestic outlets when reporting on the same country.

Alternatively, "big data" media repositories like GDELT, Internet Archive, and Common Crawl use automated crawlers to collect news articles from huge numbers of sources with impressive linguistic diversity, but fail to achieve comprehensive or consistent capture from many of the most important news sources. For example, while GDELT sources data by crawling a massive number of sources publishing in more than 100 languages, this crawling approach means that the list of sources they pull from changes constantly and new sources are included without human oversight (Raleigh, Kishi, and Linke 2023). In Section , we show that the large-scale, indiscriminate crawling by GDELT, Internet Archive, and Common Crawl all capture only a fraction of the total articles

being published by most sources. Even when these disparate big data corpora are combined, the coverage of prominent, high-quality outlets based in developing countries is extremely sparse. We further document that, due to their use of general scrapers and parsers to extract article metadata, these repositories are plagued by widespread inaccuracies in critical fields, such as the year in which articles are published.

ACLED stands alone in maintaining human review of sources while achieving broad coverage, employing more than 200 local human researchers to monitor more than 13,600 sources in over 100 languages. However, they rely on humans to manually monitor each source and pick-out reporting on relevant events. Given the huge volume of sources and the relatively small number of humans, it is likely that many of the articles published by these sources are never reviewed by ACLED's coders. Alternatively, UCDP does not provide specific information about the number of sources or languages they publish in, but they rely primarily on human monitoring of news collected by media aggregators such as Lexis Nexis (Raleigh, Kishi, and Linke 2023).

Importantly, none of these datasets or media repositories allow researchers to understand how coverage of events relate to the broader media environment. Projects that source relevant articles from media aggregators (POLECAT, UGDP) and projects rely on human monitoring of news (ACLED) do not document or retain irrelevant articles. Similarly, none of the crawler-based big data repositories successfully capture all articles published by the sources they crawl. Consequently, researchers cannot calculate how much attention was devoted to coverage of relevant events relative to other events and topics in the news. This is a major limitation, preventing researchers from determining the salience of events within a country's media ecosystem. Crawler-based repositories also leave researchers blind about why some articles published by a specific source are captured while other articles are not. To the extent that human monitoring by projects like ACLED fail to have humans check every article published by the sources they cover, they will face a similar limitation

While the vastness of the crawler-based repositories are appealing for researchers looking to report large samples and broad coverage, they also add and drop sources indiscriminately. Adding or dropping sources introduces the possibility that trends in the volume of reporting dedicated to specific events are artifacts of changes in source material rather than true changes in the frequency of these events, threatening the ability to measure trends over time. Private media aggregators also face this challenge due to frequent and erratic changes in licensing aggreements that are not accounted for by researchers when reporting an increase in the number of events being reported over time. ACLED acknowledges this challenge in their methodology, requiring that new sources only be added when resources are available to have humans scour the source's historical archives and code events for the entire time period over which ACLED's existing sources have been coded.[2]

To address these issues, *ML4P* combines recent advances in automated text analysis with *HQ-MARC's* curated corpus of news scraped directed from high-quality domestic outlets. The core of *HQMARC's* approach is to identify a curated list of critical domestic sources for each country and then design a customized harvesting workflow that can achieve comprehensive capture of everything published by those sources. This targeted "medium data" approach enables comprehensive capture from each source, allowing researchers to calculate the share of all articles published by a

[2]ACLED's documentation notes that "… the addition of such a source in an ad hoc fashion risks the integrity of historical trends as it will introduce an 'artificial spike' in the data. This refers to the phenomenon where if that same source was first back-coded before being introduced into the data, the 'spike' that its inclusion introduces in the data would be gone (or minimized) — suggesting that the spike does not reflect a 'true spike' in disorder on the ground" (ACLED 2023).

given source that were covering a specific type of event. Critically, *HQMARC's* human-supervised scraping results in a corpus with a more stable, well-understood composition than the widely-used alternatives. Because the composition is understandable and stable, this corpus can be used to measure the salience of topics or events in a source's coverage at any given moment.

This process ensures that we capture a broad range of high-quality media from countries that often go underreported in the international press. The result is a highly flexible research infrastructure that balances breadth of coverage, source quality, and processing scalability. Figure 1 provides a graphic representation of the *ML4P* data production pipeline. In the remainder of this section, we describe each steps in this pipeline.

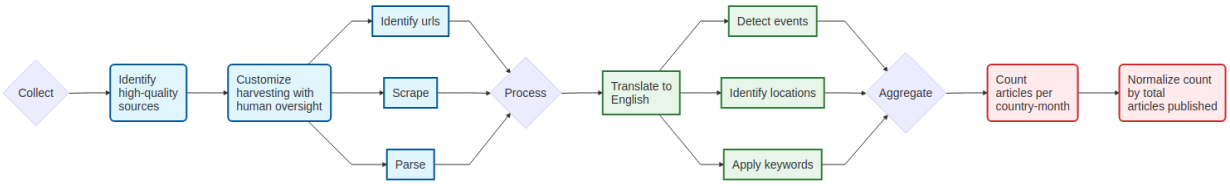

Figure 1: ML4P data production pipeline. Blue nodes captures steps in the construction of the *HQMARC* media corpus. Green and Red notes captures the data processing and aggregation steps in the construction of the *ML4P* event data.

Building the HQMARC Corpus

ML4P is constructed by processing articles from the *HQMARC* corpus. A key advantage of *HQMARC* is its unprecedented accuracy and granularity in capturing the publication history of critical domestic media outlets. To overcome the composition challenges discussed above, we developed a data-collection infrastructure designed to (1) comprehensively capture local sources' full publication history and (2) maintain accurate metadata. This process involves three main steps:

1. **Identify High-Quality Local Newspapers:** We begin by compiling a list of local newspapers with machine-scrapable websites. We consult publicly available directories of each country's media market (e.g., university library guides, Reporters Without Borders) as well as partners working in international NGOs, USAID country offices, and local civil society organizations to identify reputable publications that produce and publish original news content. Importantly, we conduct a detailed desk review of each source's partisan affiliation by consulting reports on media ownership and press freedom in the outlet's country (see Appendix A Section 4). In very repressive countries, we occasionally include newspapers based outside their home countries if they remain a leading voice on domestic affairs. For example, we include *El Faro*, a Salvadoran independent newspaper that relocated its headquarters to Costa Rica due to government persecution.

   From this initial list, we select newspapers whose online archives extend as far back as possible, preferably to 2012. We aim for at least 3–5 local sources per country, collectively yielding several thousand articles per month. In cases where a source's publication volume declines drastically or ceases entirely, we follow standardized replacement procedures. We then supplement these local outlets with articles from international and regional sources to ensure comprehensive coverage.

2. **URL Discovery:** Second, we identify urls for all articles published by a source by looking for a structured entry-point. Typically, this is a public sitemap. If the sitemap is incomplete or missing, we switch to site-specific search strategies (pagination through section indexes, keyword queries, RSS feeds, etc.). Only when those programmatic methods fail do we fall back on more intensive tools, such as simulated infinite clicking/scrolling with Selenium. Even in these cases, the goal is to retrieve clean article links, not to crawl arbitrary pages. In order to avoid storing the same article multiple times, we de-duplicate based on URL and title similarity for articles published on the same day.
3. **Develop Custom Scrapers and Parsers:** Next, we create tailored scrapers and parsers for each website's unique structure and publishing practices. These tools can bypass common barriers such as robot blockers (e.g., Cloudflare), which affect roughly 15% of our sources. By designing source-specific scrapers, we minimize data loss and ensure that critical metadata (e.g., publication date, author, section) is accurately captured.
4. **Monitor and Update Quarterly:** Finally, we evaluate scraper and parser performance every 90 days, adapting to changes in website architecture or operational status. This regular monitoring helps us detect when a source reduces its publication frequency or shuts down entirely.

Appendix A provides comprehensive documentation of the media sources and linguistic diversity underlying the *HQMARC* corpus. Section 1 documents the geographic distribution of domestic and regional media outlets across our sample, presenting visualizations of source counts by country and temporal patterns of source activity over the 2012-2024 period. Section 2 catalogs the linguistic diversity of the corpus, documenting the languages published by media outlets in each country using ISO language codes. Section 3 provides the complete inventory of digital news sources included in *HQMARC* by country, offering full transparency about the specific media outlets that comprise our dataset.

To demonstrate the importance of our custom harvesting workflows, we conduct a case study comparing *HQMARC's* coverage with that of several "big data" media corpora, demonstrating that *HQMARC* captures a significantly larger share of articles from high-quality domestic news outlets. We then demonstrate the pitfalls of relying on generalized scraping and parsing tools without human oversight, showing how our customized harvesting workflow with human oversight mitigates data-reliability issues that affect automated mass crawlers.

To demonstrate that relying on crawler-based big data archives results in poor coverage from critical sources in developing countries, we compare *HQMARC's* coverage of three prominent Bangladeshi news outlets to that of GDELT and Internet Archive. We focused our case study on Bangladesh for three reasons. First, Bangladeshi outlets publish a high volume of articles relative to other countries, making them more likely to attract and be captured by automated crawlers. Second, the website architecture for each outlet is relatively straightforward, maximizing the likelihood that cralwers, combined with generalized scrapers and parsers, should be able to accurately retrieve articles. Website architecture varies widely across sources. For Bangladeshi outlets, relatively minimal customization was necessary for *HQMARC's* scrapers and parsers, implying that general, large-scale crawlers should achieve good coverage. Third, many Bangladeshi sources publish primarily in English, reducing the additional hurdles of multilingual parsing. As a result, we regard these outlets as a "best-case scenario" for large-scale media repositories.

Despite favorable conditions, we find notable differences between the results achieved by *HQMARC's* curated approach and those of GDELT and Internet Archive. *HQMARC's* coverage begins in 2013 for one source and in 2015 for the other two. However, GDELT does not have any articles published before 2019 for any of the three sources. Even within the overlapping years beginning 2019, GDELT captured many fewer articles than *HQMARC*. For the source with the smallest disparity between *HQMARC* and GDELT, GDELT retrieves an average of 2,100 articles per month, compared to 2,500 in *HQMARC*. GDELT also includes numerous broken links, redirects, duplicate articles, and advertisements that were flagged by *HQMARC's* human review and removed. In addition, GDELT enforces a five-second delay per query, making it extremely time-consuming to scrape a full historical archive of this size. Across these three sources, Internet Archive achieved coverage similar to that of *HQMARC*, but more than half of these urls were broken and no longer pointed to the a webpage that contained the article text. Furthermore, collecting URLs from Internet Archive for 2019–2023 required roughly two weeks from a single source and returned many irrelevant and duplicate links not contained in *HQMARC*.

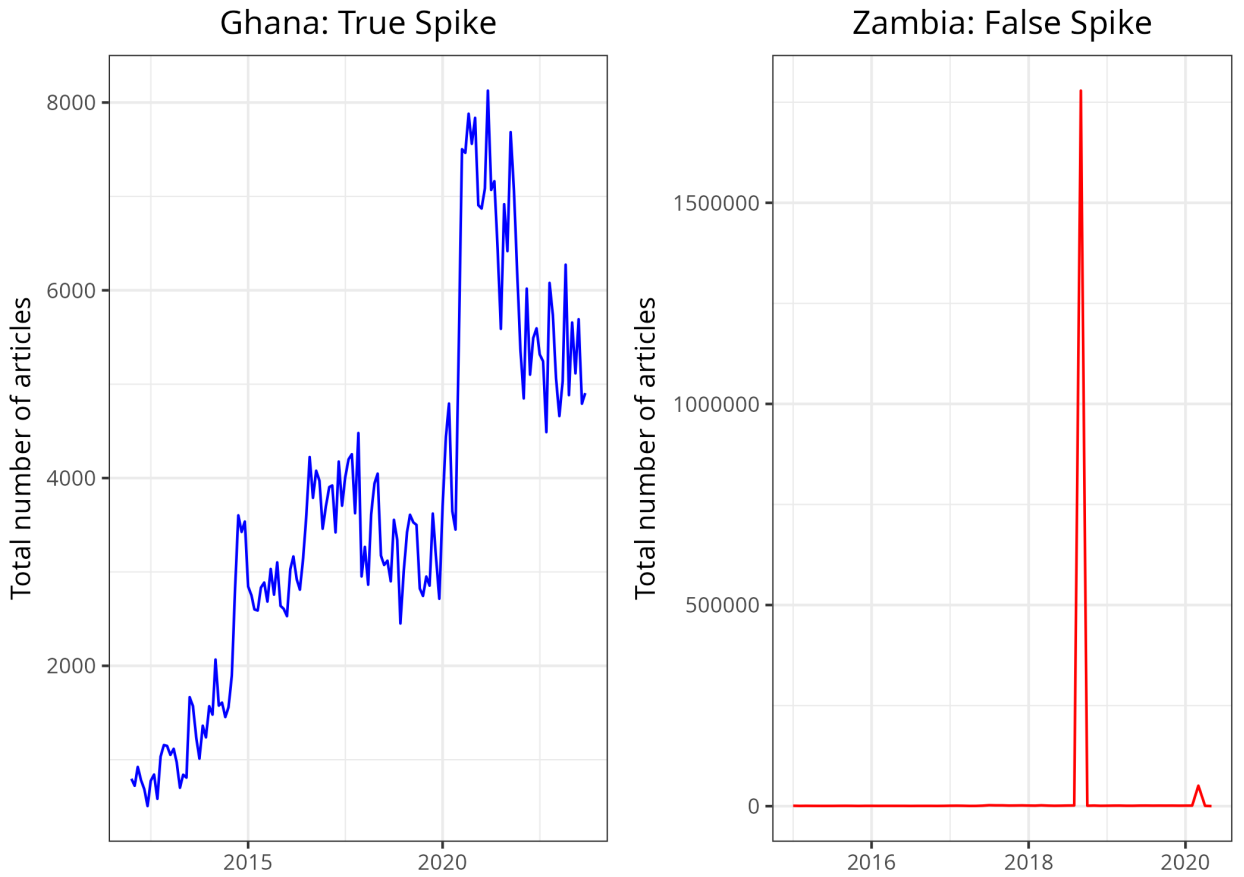

Figure 2: Changes in the volume of articles across two sources. In Ghana, the sudden in shift in volume was driven by a grant that reflected a real change in the total articles being published. In Zambia, the sudden shift in volume was driven by a single article duplicated hundreds of thousands of times.

Our investigation shows that large-scale crawlers often fail to collect all available urls, with *HQMARC's* customized workflow resulting in much more comprehensive coverage from critical sources. However, once urls are collected, information must be extracted from these urls with scrapers and

parsers. When scraping many thousands of sources, big data media repositories must rely on generalized scraping and parsing tools without human oversight. Figure 2 highlights one of the many ways that this can introduce errors and the necessity of human oversight in mitigating data-reliability issues. When scraping articles without human oversight, we see a large spike in the volume of articles being published by major outlets in Ghana and Zambia. On the left, this spike captures a genuine increase in articles published by `ghanaweb.com` (Ghana), which wwas triggered by a grant from Google that enabled the outlet to expand its reporting capacity. On the right, we see an artificial spike in the number of articles published by `lusakatimes.com` (Zambia) driven by a single article being hosted at more than 1.5 million *unique* urls on the source's website, making the duplication difficult to detect. In both cases, human overseers noticed a suspicions spike in publication volume and investigated the cause. Our human-in-the-loop approach effectively guards against such errors, enhancing the overall reliability and quality of *HQMARC*. Importantly, such errors can be caused by a wide range of scraping and parsing failures, including dates that are incorrectly formatted or incorrect dates or other tags accidentally embedded in an articles html.

Capturing Civic Space Events in ML4P

To generate the *ML4P* data tracking civic space events, we use open-source computation tools to extract information the text of articles stored in the *HQMARC* corpus. We describe each step in this process, including translating article text, identifying the primary locations mentioned, detecting articles that cover relevant events, ensuring the political relevance of detected events, measuring the salience of events at the country-month level, and detecting months with high levels of activity across event-types.

Translating Non-English Text

It is well-documented that reporting by international and regional media outlets on political events in developing countries contains significant biases that can affect data quality (Baum and Zhukov 2015), even when covering relatively uncontroversial topics like natural disasters (Brimicombe 2022). To address this, *HQMARC* focuses on collecting data from a curated list 348 prominent *domestic* media outlets publishing in 36 languages. In fact, more than 95% of the articles in *HQMARC* are scraped from domestic media outlets based in the countries covered by our dataset.

After scraping the text of each article, we translate the first 600 characters of all non-English articles into English using neural machine translations (NMT) through Hugging Face or OpenNMT.[3]. To select a translation model, we test the efficacy of all models by extracting sample text from articles published in each language and running the text through all available translation models on the Hugging Face open database. We then assess whether the translations are sufficiently clear to identify the main event being reported on. If multiple models produce satisfactory results, we select the model that yields the clearest sentence-to-sentence translations, comparing results to that of Google translate to assess whether there is a significant loss of contextual detail. If they are not, we test commercial translation services accessible through the `deep-translator` Python

[3]While multilingual transformer models capable of classifying events directly in multiple languages exist, these models are not currently able to support the diversity of languages present in *HQMARC*. However, once more capable open-source models become available, *ML4P's* flexible infrastructure will allow us to quickly apply these models to the entire *HQMARC* corpus.

package. For every language in our corpus, we were able to find translation models that generated translations clear enough for humans to identify the relevant event being reported.[4]

Identifying Locations

Both international and domestic outlets report on events taking place across a broad range of countries. To ensure that the events we capture events in the country of interest, we identify all locations mentioned in the first 600 characters of text. If no country is found in the text, we assign the article to the country in which the publishing outlet is based. For international and regional outlets, articles are only assigned to a country if they explicitly mention a location within that country in the first 600 characters.

To locate events and identify those happening within a target country, we use the CLIFF-CLAVIN geoparser[5] with the GeoNames ontological gazetteer to identify geographic entities (e.g., states, cities, towns) mentioned in the text. CLIFF-CLAVIN integrates the GeoNames database, which is a free, online directory containing over 12 million place names across 250 countries (D'Ignazio et al. 2014). GeoNames is one of the most comprehensive and actively maintained sources of geographic data, making it an ideal reference for matching entity mentions to specific global coordinates. For each location detected, we use the CLIFF API to retrieve the location's country code and assign the article to that country. We implement several corrections to the underlying CLIFF system, including overriding an error that assigns mentions of "West Africa" to Angola and the assignment of "Gaza" to locations named "Ghaza" in Algeria and Pakistan.

Classifying Civic Space Events

To identify articles reporting on one of our 20 civic space event categories, we fine-tuned an open source, transformer-based RoBERTa language model (Liu et al. 2019).[6] To fine-tune this model, we constructed a double human-coded training dataset consisting of 6,475 newspaper articles (4,982 reporting on one of our 20 events and 1,493 reporting on irrelevant events) originally published in both English and non-English languages. Average out-of-sample classification accuracy is above 80%,[7] with many misses coming from the presence of multiple events in a single entry or from partially overlapping event categories. To reduce noise associated with background and contextual content, we apply this classification to the article title and first 600 characters of main text.[8]

Table 1 summarizes the out-of-sample performance of our fine-tuned model. The model achieves an overall accuracy of approximately 0.82 (for civic space events), with many misclassifications arising

[4]Tetum is the only language for which we could not initially find an acceptable translation model. After waiting several months, an open-source model became available.

[5]For technical details on CLIFF, see: CLIFF Annotator

[6]Recent research has shown that costly, closed-source LLMs only perform moderately better at even complicated tasks relative to first-generation models like RoBERTa, and usually require much more costly fine-tuning [@de-andrade2024]. Moreover, RoBERTa performs well for most common applications in Political Science [@timo-neda2024roberta].

[7]This is comparable to intercoder reliability. During model training, we adopted an iterative approach: after each training round, we gathered new examples for categories with lower accuracy, retrained the model, and repeated until performance stabilized.

[8]Typically, this corresponds to the article title plus the first two sentences of text. Extensive testing suggests that providing addition text from articles decreases classifier performance by including irrelevant contextual information that reduces the model's ability to identify the main event.

from articles that mention multiple overlapping events. Column-specific metrics (precision, recall, F1) are provided for each event category. See Appendix B Section 1 for a definition and examples for each event category.

Table 1: Performance metrics for fine-tuned RoBERTa classification model. -999

Event Category	Precision	Recall	F1
Arrest	0.91	0.88	0.89
Protest	0.85	0.98	0.91
Legal action	0.77	0.75	0.76
Disaster	0.87	0.86	0.86
Censor	0.76	0.95	0.84
Election activity	0.78	0.84	0.81
Election irregularities	0.72	0.68	0.70
Activism	0.95	0.83	0.88
State of Emergency	0.92	0.90	0.91
Cooperate	0.50	0.67	0.57
Coup	0.68	0.83	0.75
Non-lethal violence	0.79	0.81	0.80
Lethal violence	0.90	0.82	0.86
Corruption	0.74	0.71	0.73
Legal change	0.84	0.80	0.82
Security mobilization	0.83	0.77	0.80
Purge	0.91	0.86	0.88
Threats	1.00	0.78	0.88
Raid	1.00	0.83	0.91
Irrelevant events	0.81	0.79	0.80

To address articles containing multiple overlapping events, we permit dual classifications for certain event types. In particular, events such as corruption often occur concurrently with arrests or legal proceedings. For several categories, we further apply a targeted keyword filter to eliminate common false-positives. See Appendix B Section 2 for a description of keyword filtering used during classification.

Distinguishing Politically Relevant Events

In addition to our event classification model, we developed an additional classifier to catch and exclude articles reporting on events that match one of our event categories but are not politically relevant. For several event categories, we encountered false positives resulting from close overlap between our event categories and newsworthy but irrelevant events. For example, our Arrest category seeks to identify politically relevant arrests, such as the arrest of a politically relevant figure. However, our model would occasionally include articles reporting on arrests for criminal activity that did not meet our category definition. This political relevance classifier was built using transfer learning from our fine-tuned RoBERTa model, fine-tuned using a double human-coded training dataset of 2,938 articles and achieving an overall accuracy of 0.87. For each article the main classifier flags as an event, this secondary model provides a binary (0/1) output indicating

its civic relevance. Our final dataset includes articles that were classified according to one of our event categories, but flagged as not politically relevant, in separate columns capturing politically irrelevant events.

Measuring Event Salience

Finally, we aggregate these data to the country-month level, normalizing the count of articles reporting on each event by the total number of articles published in that country-month. The final *ML4P* measures correspond to the monthly share of all news articles reporting on a country that are reporting on each *ML4P* event category. Importantly, this is made possible by *HQMARC* custom scraping to capture outlets' full publication history, which allows us to measure the true number of articles published by constituent sources. This ratio tells us how frequently each *ML4P* event type is reported-on relative to the total volume of news in a given month. While this method does not directly allow us to code individual *ML4P* events, it does provide information on the *relative importance* of each type of event-category in a given month. This approach also enhances the ability to measure trends over time by avoiding the risk that increases or decreases in our event measures are artifacts of changes in volume of overall reporting driven by sources entering or leaving the database or changes in the publication volume of sources over time (ACLED 2023).

Detecting Major Event Shocks

For each of our 20 civic space event categories, we generate a measure of how much media attention that event type received on a month-to-month basis. We supplement these measures by identifying months in which major civic space events occurred. To detect months with major events, we developed an ensemble algorithm to detect sharp increases in the share of reporting dedicated to each event category. We refer to these sharp increases as *shocks*.

Our approach begins with winsorization of the data, which curbs the influence of extreme outliers by replacing values beyond a specified percentile threshold with the nearest boundary value. Next, we apply a 25-month rolling window to smooth the normalized event counts and perform a grid search to tune various parameters. These include the multipliers for weighted means and weighted standard deviations, as well as the binning weights and decay functions that govern how observations in the window are weighted. To capture shocks accurately, we employ two distinct weighting schemes for the historical (left-hand side, LHS) and future (right-hand side, RHS) segments of the rolling window. For the LHS window, we use a non-linear decay weighting that places progressively less emphasis on more distant historical months, enabling the detection of rapid changes in recent data. For the RHS window, we apply binning weights that decay linearly over time, preventing overestimation of peaks when the underlying data structure shifts. Combining winsorization with context-sensitive decay and binning weights enables early detection of significant increases in civic space activity.

Next, we trained neural network model to detect spikes from a human-labeled dataset covering the full time-series for 30 country-event pairs. Human-labeling was conducted by asking humans to identify months with visually distinct, sharp increases in our event measures. Human labelers were instructed to identify no more than 15% of overall months as shocks, ensuring that peaks are not overly frequent in highly variable data while still capturing meaningful shifts in lower-variance

event types. When either the statistical or neural network model detect a shock, we label that month as a shock in the data.

Data Description and Validation

In this section, we present the data and results from two data validation exercises. Figure 3 and Figure 4 show cross-national variation in the data across four annual snapshots. Figure 3 measures the share of all articles classified as one of our 20 civic space events across four years. For example, the share of Ukraine's coverage dedicated to civic space events jumped from 10% in 2012 to 30% in 2024, . Complementing this, Figure 4 examines the most frequently reported-on civic space event type across countries. Several temporal shifts are evident: in 2020, Natural Disasters dominated coverage in most coverage, driven by coverage of COVID-19, while in 2024, several countries with high-profile national elections see Election Activity coverage dominate. Maps showing annual averages for all years in the data can be found in Appendix C.

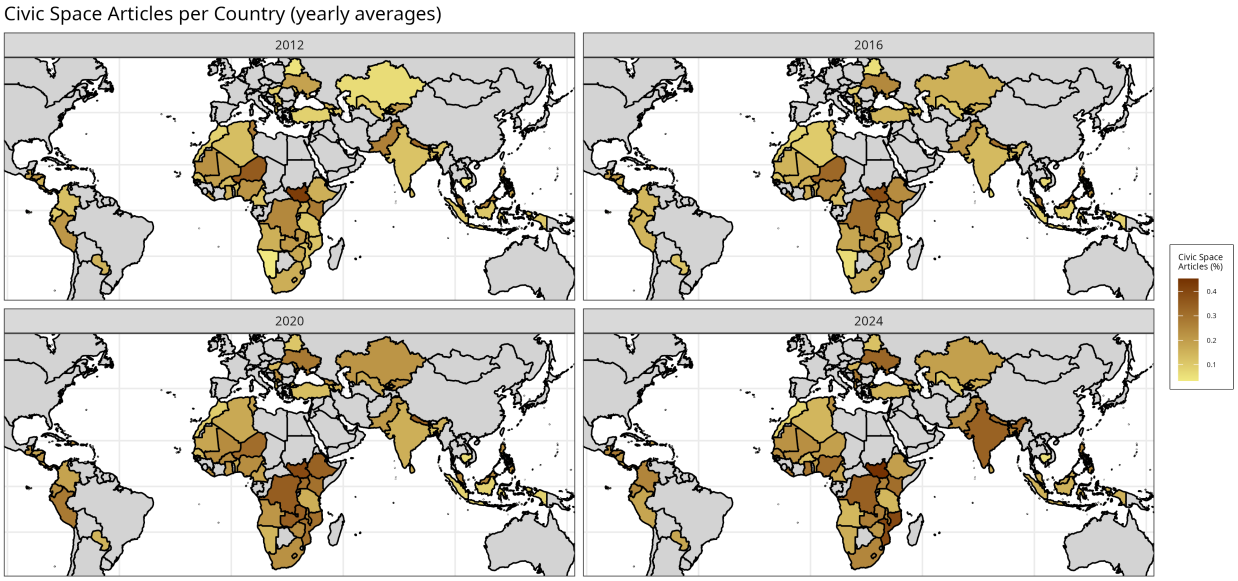

Figure 3: Civic space articles as percentage of total coverage by country-year for 2012, 2016, 2020, 2024. Countries are colored according to the proportion of their total news coverage dedicated to civic space events. Darker values show larger percentages, revealing significant variation in civic space reporting intensity both across countries and over time.

Comparing International and Local Media

While the previous section demonstrates how *ML4P* provides reliable data on civic space dynamics, we now show that this data relies heavily on the domestic sources targeted by *HQMARC*. Compare domestic and international news coverage of civic space events across our sample of developing countries, we demonstrate that relying solely on international media often yields an incomplete and potentially biased view of civic space.

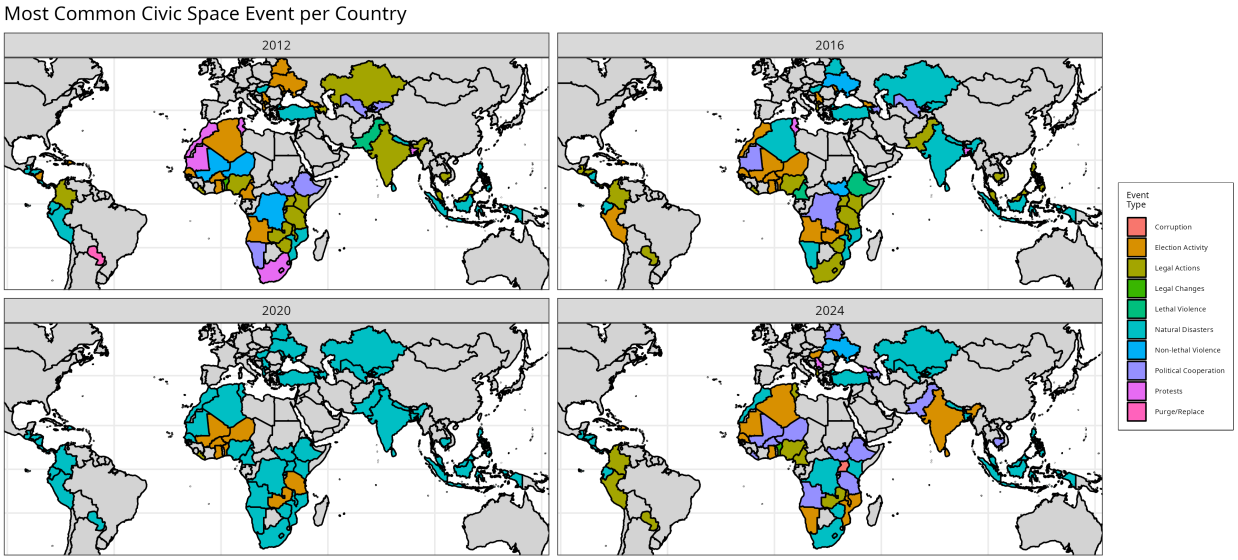

Figure 4: Most frequently reported civic event type per country-year for 2012, 2016, 2020, 2024. Countries are colored according to the civic event category with the highest reporting frequency in each year, showing the diversity of civic space concerns across different national contexts and revealing temporal shifts in civic space priorities.

For all articles in *HQMARC* reporting on a each event category, Figure 5 presents the proportion of those articles coming from domestic versus international sources averaging across countries in our sample. The stacked bars show the percentage of total articles about each civic event type that come from domestic (blue) versus international (green) sources. The red points labeled with correlation coefficient represent the correlation between domestic and international reporting for each civic space event category within countries and over time. If domestic and international sources were covering the same events at the same times, we would expect them to increase their reporting on event categories roughly simultaneously, resulting in a high correlation. Yet the correlation in reporting is consistently weak or moderate across our 20 event categories. Furthermore, this correlation is not consistently higher for categories that receive more international attention, such as Lethal and Non-Lethal Violence.

This correlation in the focus of coverage between international and domestic sources does increase significantly for countries that receive more international media attention.Rather than reporting the average correlation within event categories across countries, Figure 6 shows the correlation within countries and across event categories. Figure 6 plots a country's volume of international articles (x-axis) against the correlation in civic space event coverage between international and national sources (y-axis). Although the correlation increases substantially for countries that receive more iternational coverage, the correlation remains low in nearly all cases. For example, Turkey and India receive substantial international attention yet exhibit domestic-international correlation values below 0.5.

We see similar variation when considering where international and domestic outlets focus their attention. For all articles in *HQMARC* coming from domestic versus international sources, Figure 7 investigates how intensely these sources focus on each event category. Overlapping bars show the

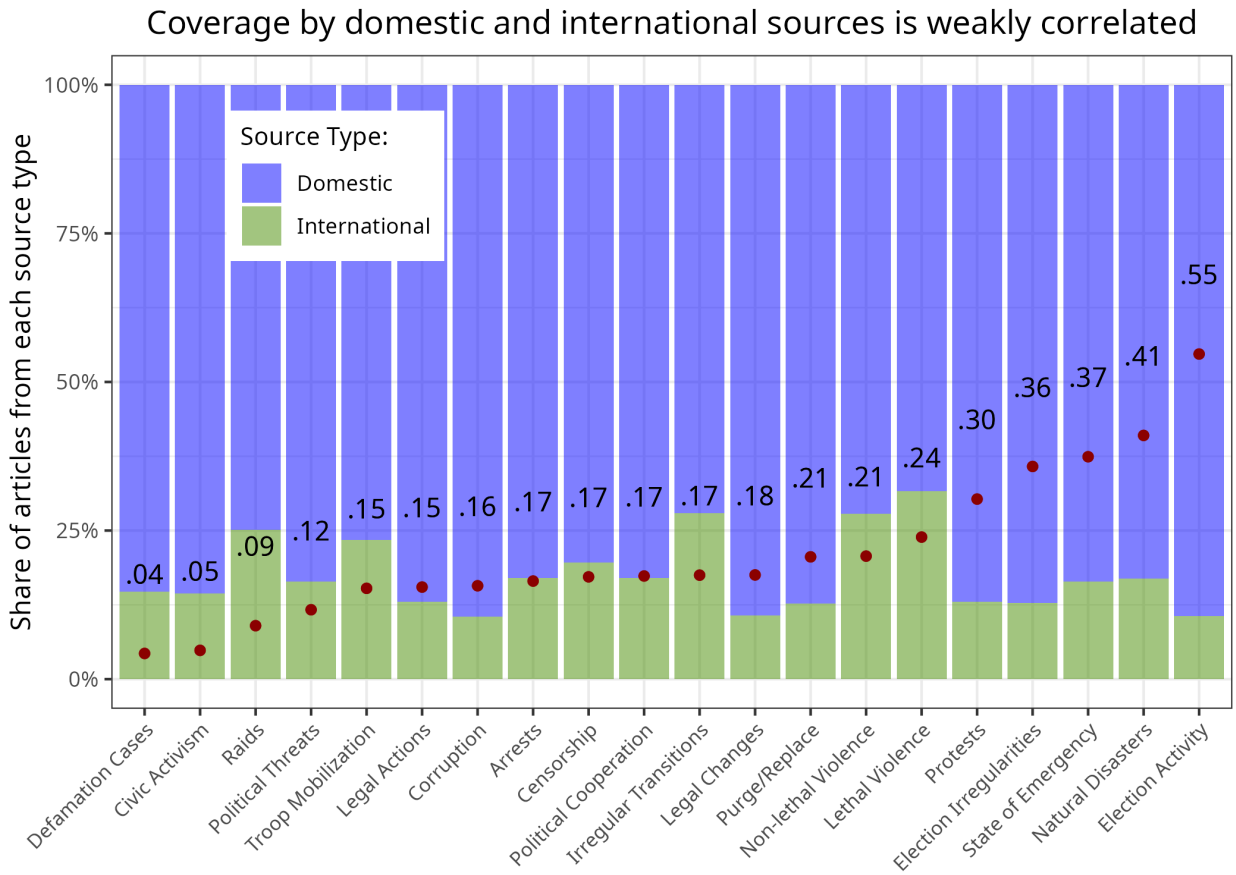

Figure 5: Proportional distribution of civic event coverage between domestic and international sources. Stacked bars show the percentage of total articles about each civic event type that come from domestic (blue) versus international (green) sources. Events are ordered by correlation strength between domestic and international coverage (red points labeled with correlation coefficient). Domestic and international sources show weak correlations in their coverage patterns, regardless of the share of articles coming from international sources. Example: Of all Arrest articles, 83% come from domestic sources and 17% from international sources. The correlation between domestic and international reporting on arrests is 0.17.

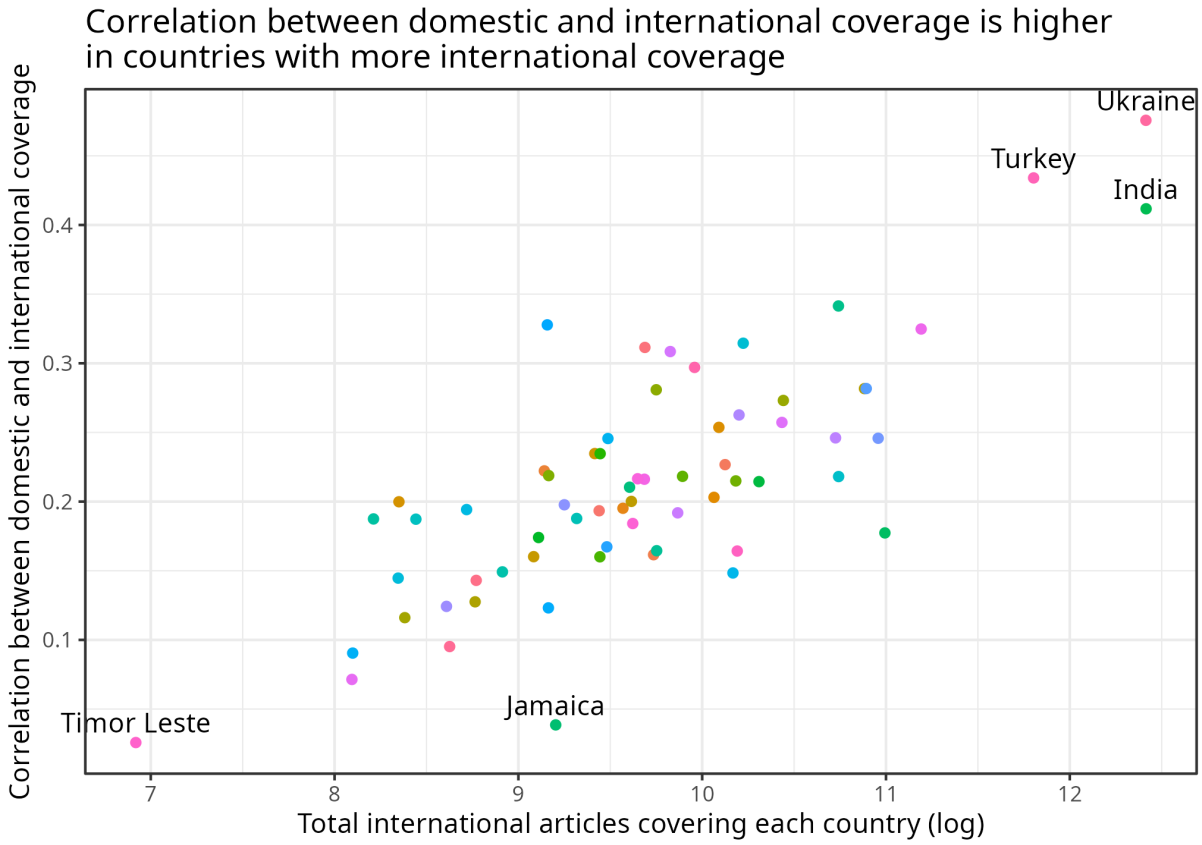

Figure 6: Relationship between international media attention and domestic-international coverage correlation. Each point represents one country, plotting the correlation between domestic and international civic event coverage (y-axis) against the total volume of international articles about that country (x-axis, log scale). Countries with higher international media attention tend to show stronger positive correlations between domestic and international coverage patterns, suggesting that sustained international focus may lead to more synchronized reporting priorities. Selected countries are labeled to illustrate this pattern: Ukraine, Turkey, and India (high international attention, strong positive correlation) versus Timor Leste and Jamaica (low international attention, weak correlation). Even for high-attention countries, this correlation is surprisingly weak.

rate at which each source type covers different civic events, expressed as articles per 10,000 total articles published by that source type. Consistently higher bars for international sources suggest that international sources devote a larger share of their total articles to civic space events, relative to domestic outlets, whose coverage includes a greater share of articles reporting on events that do not fall into one of our 20 event categories. We also see that, among civic space events, international sources exhibit higher relative focus on Lethal and Non-lethal Violence, while domestic sources exhibit higher relative focus on Election Activity. These patterns reflect different editorial priorities of international versus domestic media.

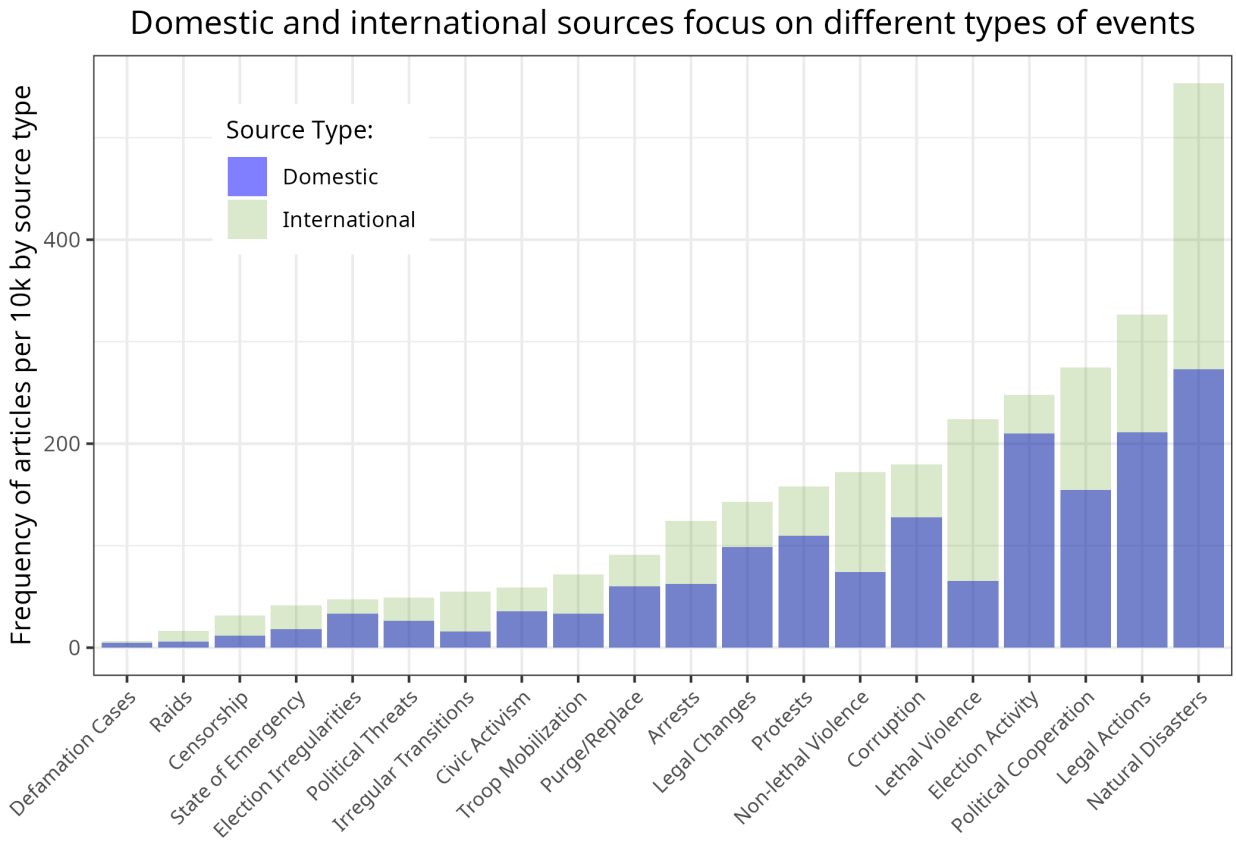

Figure 7: Frequency of civic event coverage within domestic versus international source portfolios. Overlapping bars show the rate at which each source type covers different civic events, expressed as articles per 10,000 total articles published by that source type. Events are ordered by international coverage frequency (lowest to highest). International sources (green) show more intense focus across all civic event types, suggesting their International sources exhibit higher relative focus on Lethal and Non-lethal Violence, while domestic sources (blue) exhibit higher relative focus on Election Activity. These patterns reflect different editorial priorities of international versus domestic media. Data normalized by total articles within each source type per country-month. Example: Domestic sources publish 65 lethal violence articles per 10,000 articles, while international sources publish 326 lethal violence articles per 10,000 articles.

We illustrate the broader point with a case study of reporting on corruption in Indonesia. Between March and November, a series of major corruption scandals broke out, including the PT Timah state

enterprise corruption case in March, the Tom Lembong sugar import case and the Sahbirin Noor South Kalimantan Governor case in October, and the Rohidin Mersyah electoral case in November. In May 2024 alone, domestic outlets published over 970 articles covering corruption revelations and 107 articles covering political arrests, many related to the PT Timah state enterprise case. Despite the gravity of these events, our regional and international sources carried *zero* relevant articles. This discrepancy suggests that a reliance on international and regional sources not only provide an incomplete view of the salience of different events in the domestic political environment, but also the possibility that major events may be entirely overlooked. Across our dataset, we identified 59 cases across 23 countries where we detected shocks indicative of major events but found zero relevant articles published by our regional or international sources.

Major Event Validation

To validate the ability of our data to detect major political events, we adopt two approaches. First, focusing on months designated by our model as containing major events (shocks), we examine whether observed spikes in coverage correspond to real-world developments (capturing true positives). We randomly selected five countries from our database—Kosovo, Morocco, Angola, Mauritania, and Ukraine—which together span three regions (North Africa, Sub-Saharan Africa, and Eastern Europe) and four languages (Serbian, Albanian, French, and Ukrainian). Focusing on the three most recent months of data available at the time of the analysis (April–June 2024), there were 300 total possible events that our shock detection model could have detected (3 months $\times$ 5 countries $\times$ 20 events). Of these 300 possible event-months, our model flagged 40 of these event-months as exhibiting a shocks.

For each of the 40 country-event-month combinations, we retrieve all relevant articles published by domestic outlets in the *HQMARC* corpus to determine whether these shocks are true or false positives. Because some of these 40 country-event-months with shocks contained hundreds of articles,[9] we used GPT-4o[10] to generate brief summaries of the five most important events being reported on in the sample of articles. A research assistant then reviewed both (a) the original articles (or a random subset of 50, if more than 50 were available) and (b) the GPT-4o summaries. They evaluated each of the top five summarized events on four dimensions:

- How many summarized events are accurately described by GPT-4o (i.e., factually correct)?
- How many summarized events are indeed the most important events, according to the underlying articles?
- How many occurred in the assigned country?
- How many match the assigned event category?

In 34 out of 40 country-months (85%), *all* top-five events identified by GPT-4o occurred in the correct country, and in 38 out of 40 (95%), *all* top-five events belonged to the correct event category. Overall, out of 200 summarized events, 181 were true positives. These findings strongly suggest that our measure reliably detects months with major political developments.

Our second approach begins by identifying major events in the world that we would expect to generate measurable shocks in the *ML4P* data and confirming that these events are captured as

[9] Across detected events, the number of relevant domestic articles ranged from 1 to 1,002. For rare event categories (e.g., *Defamation Case*), a single article can define an event.

[10] The full GPT-4o prompts, GPT-4o generated summaries, human-coding instructions, and validation results can be found in the `shock-validation` subfolder of the Git repository. GPT-4o was accessed through the OpenAI API.

shocks in the data (avoiding false negatives). We apply this approach in three ways: (1) identifying a single historical event likely to generate media attention on one of our *ML4P* event categories across multiple countries, (2) assessing our ability to detect both frequent and rare political events through spikes in relevant *ML4P* event categories, and (3) analyzing events within a single country that are expected to trigger shocks across multiple *ML4P* event categories.

We first examine the onset of the COVID-19 pandemic and government responses to it, particularly the widespread implementation of social confinement measures, such as lockdowns. These measures often included school closures, curfews, and restrictions on non-essential businesses and government services (Cheng et al. 2020). These lockdown measures should be associated with shocks in the State of Emergency event category. As shown in Appendix D, we detect shocks in State of Emergency counts across all countries in our dataset starting in March 2020.

Next, we analyze the detection of key political events across multiple countries, focusing on a relatively frequent event—elections and electoral activities—and a rare event—coup d'état. To do so, we identify the most recent electoral event (e.g., general, parliamentary, or presidential election) for each country in our dataset. Figure 8 demonstrates that our shock detection models detect periods of heightened electoral activity in the months immediately preceding these elections in all 10 Latin American and Caribbean countries we examined. Notably, our approach also detects electoral activity in electoral autocracies, such as Nicaragua under Daniel Ortega (Thaler and Mosinger 2022), highlighting its ability to track political events across different regime types.

We now focus our attention on coups d'état. We identify all nine countries in our sample—Burkina Faso, DR Congo, Ethiopia, Mali, Niger, Peru, Tunisia, Turkey, and Zimbabwe—where a coup or self-coup has been attempted or succeeded in the past decade. We then assess whether these events correspond to shocks in the Irregular Transition event category. As Figure 9 illustrates, all 14 successful or attempted coups and self-coups across these nine countries are associated with a shocks in the Irregular Transition event measure.

Finally, we examine events expected to generate shocks in multiple event categories. We focus on the 2023 Guatemalan general elections, where opposition candidate Bernardo Arévalo and his party, *Movimiento Semilla*, secured a surprise victory despite institutional attempts to undermine their candidacy. The electoral period spanned from January to August 2023, concluding with the run-off election. As expected, the upper-left panel of Figure 10 shows that our shock detection models identify a shock in the Election Activity measure during this period. Efforts to disqualify Arévalo intensified between the first-round election on June 25 and the run-off on August 20, followed immediately by attempts to prevent him from taking office once the vote count was finalized. The upper-right panel of Figure 10 captures these attempts, detecting shocks in Election Irregularities between the general and run-off elections, as well as in the post-election period leading up to the inauguration.

The lower-left panel of Figure 10 provides further validation of *ML4P's* ability to capture critical events in Guatemala's political history. Between the run-off and the inauguration, we detect a shock to reporting on Legal Actions in September, corresponding with the Public Prosecutor's direct efforts to nullify the election results, and again in December 2023, when the Organization of American States (OAS) officially condemned the ongoing and intensifying power grab.[11] Finally, the lower-right panel of Figure 10 shows that our shock detection models detect a shock in reporting on the massive protests that broke-out in October, led by indigenous groups and civil society

[11]The press release by the OAS can be found here: https://www.oas.org/en/media_center/press_release.asp?sCodigo=E-084/23

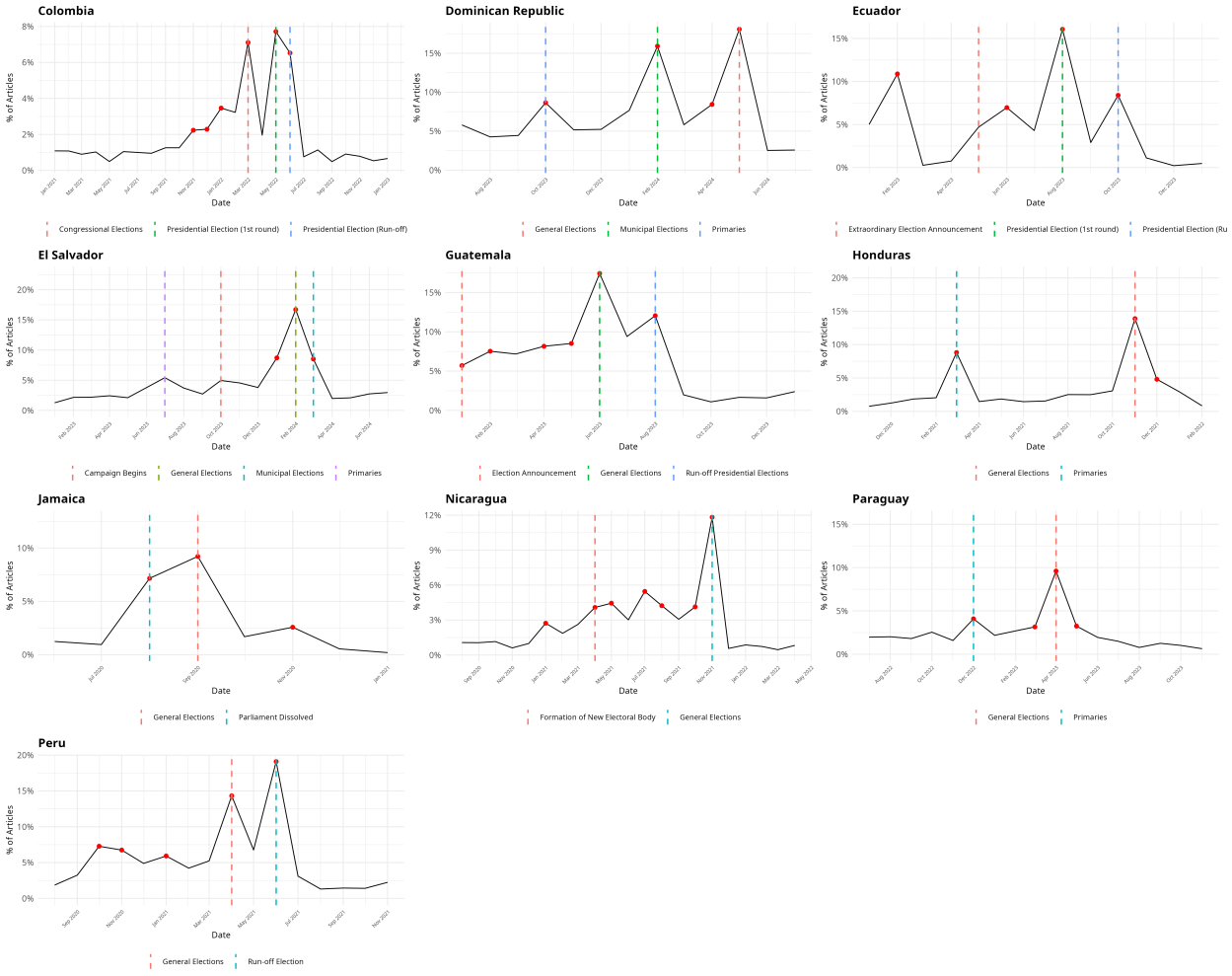

Figure 8: Elections and Electoral Activity (Latin America and the Caribbean). Notes: The vertical lines in each panel indicate key milestones in each country's electoral cycle, such as primary elections, congressional elections, or presidential elections.

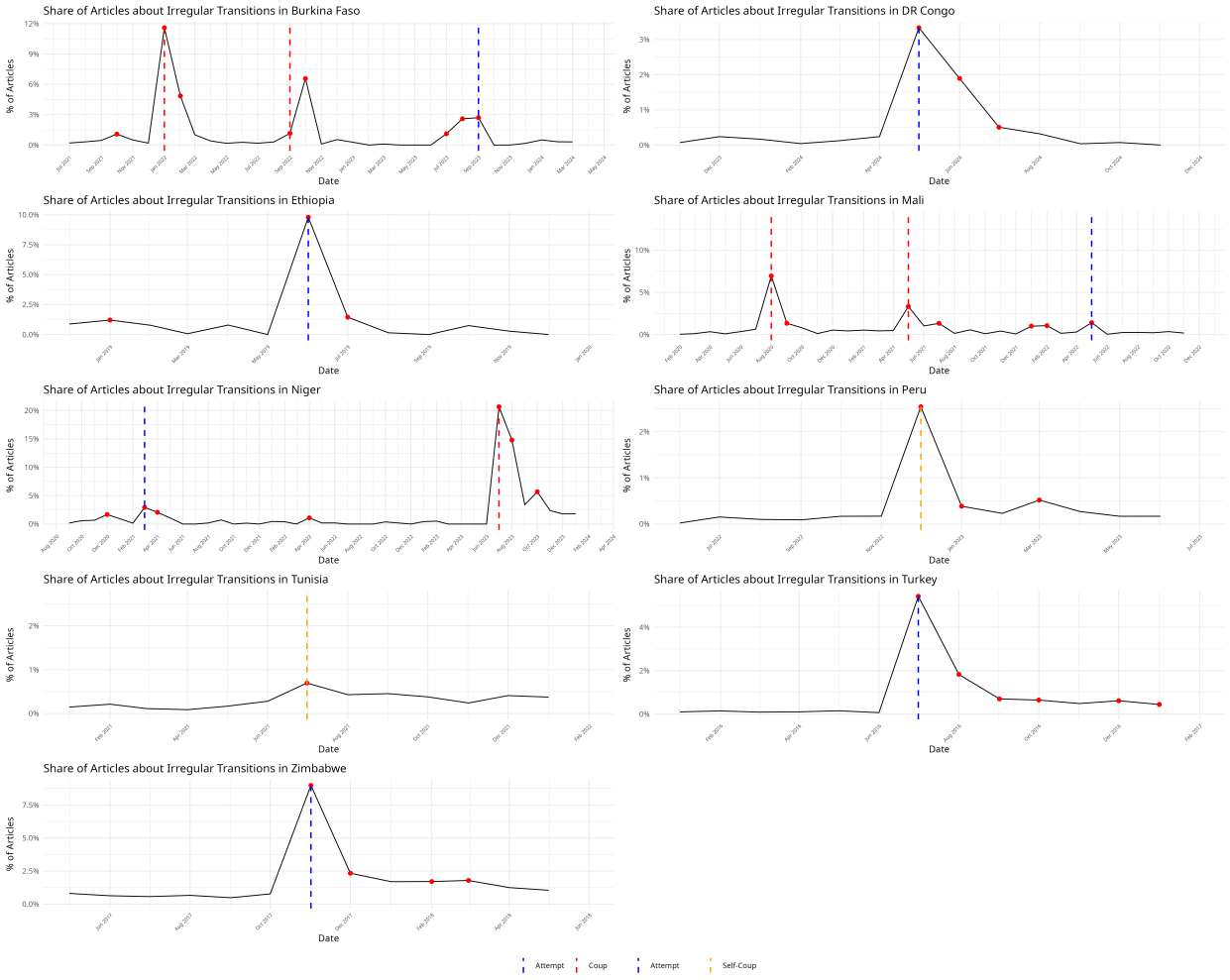

Figure 9: Coups and Self-Coups (Attempted and Successful). Notes: The vertical lines in each panel indicate the month in which a coup (successful or attempted) or a self-coup occurred in each country.

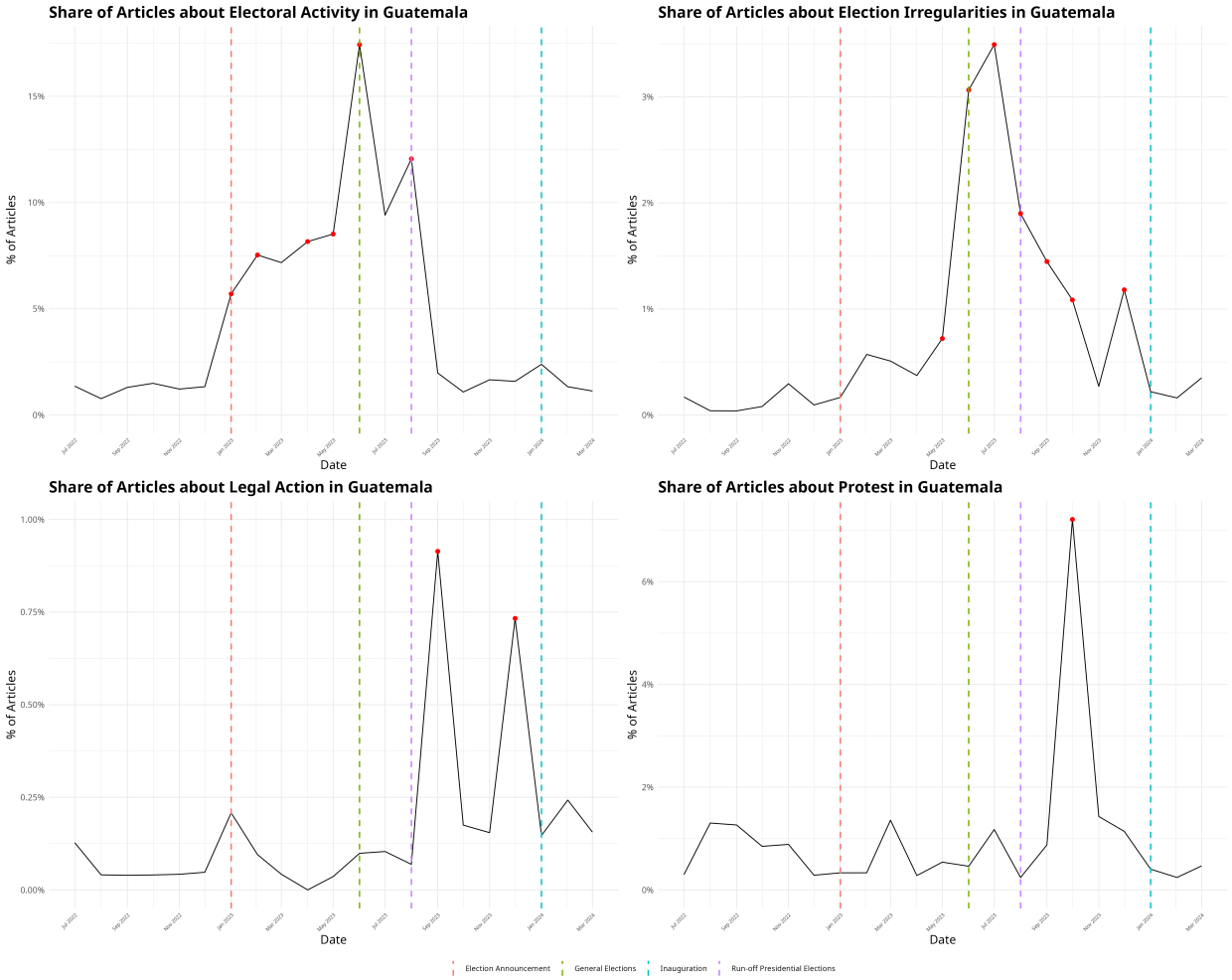

Figure 10: Elections in Guatemala 2023. Notes: The vertical lines in each panel represent key milestones in Guatemala's 2023 electoral cycle, beginning with the official election announcement in January 2023 and concluding with the presidential inauguration in January 2024. The general elections were held in June 2023, followed by the presidential runoff in August 2023.

organizations demanding the resignation of the Public Prosecutor and a peaceful transfer of power. Researchers have identified this civil society mobilization as a key factor in ensuring the eventual transfer of power (Schwartz and Isaacs 2023; Meléndez-Sánchez and Gamboa 2023; Romero 2024). Overall, this case provides compelling evidence of *ML4P*'s ability to capture the fast-paced civic space events underlying an attempted autocratic turn in Guatemala's political system.

Use Case: Forecasting Travel Advisory Onsets with ML4P

In this section, we demonstrate that *ML4P* ability to capture the fast-paced civic space events underlying broader political dynamics makes it useful in forecasting applications.

An effective forecasting model must capture real-world political instability rather than artifacts of measurement or reporting bias. By demonstrating that high-frequency civic space indicators can predict the issuance of independent security assessments—U.S. Department of State (DOS) high-level travel advisories (HLTAs)—this study provides de facto validation that MLP's data reflects meaningful underlying political conditions. If the model successfully anticipates HLTA onsets months in advance, it confirms that MLP's indicators do not merely correlate with political events but also encode predictive signals of escalating instability.

Data and Methodology

To evaluate whether MLP's data can forecast real-world security risks, we construct a monthly panel dataset covering 60 developing countries from 2012 to 2023 and model the onset of Level 3 ("Reconsider Travel") or Level 4 ("Do Not Travel") travel advisories. These advisories serve as official risk assessments, issued by the U.S. Department of State in response to civil unrest, political repression, armed conflict, health emergencies, natural disasters, and crime. Beyond their significance as public warnings, these advisories drive critical operational responses, including embassy closures, staff relocations, and security escalations. By predicting these advisories before their issuance, our model effectively demonstrates that MLP's civic space indicators capture early warning signals of deteriorating conditions on the ground.

While DOS publishes travel advisories in real time, historical archives are not systematically structured, making retrospective analysis challenging. We reconstructed this dataset by scraping the DOS website for past advisories and supplementing gaps using archived versions from the Wayback Machine (web.archive.org). Since the advisory system changed in 2018—introducing a four-level classification—we standardized earlier advisories by coding all pre-2018 warnings as "serious" (Level 3 or 4 equivalent). We then constructed a binary onset variable, flagging new advisories in a given country-month while ensuring that continued warnings were not misclassified as new events. This dataset, spanning over a decade of U.S. government security assessments, provides a unique opportunity to evaluate whether high-frequency civic space indicators contain predictive signals of advisory issuance.

To forecast HLTA onsets, we incorporate monthly event counts from MLP across 20 political event categories, including protests, arrests of activists, media censorship, election irregularities, and emergency declarations. Given the importance of both gradual and sudden shifts in political conditions, we lag all features by up to 12 months, allowing the model to detect both long-term precursors (such as increasing government repression) and short-term triggers (such as post-election

violence). Additionally, we account for persistence by including an indicator distinguishing between new HLTA onsets and continued warnings, a country-specific Bayesian prior that adjusts for baseline differences in advisory issuance, and a COVID-19 indicator to control for the pandemic-driven shock in travel warnings in 2020.

HLTA onsets occur in only 1.42% of country-months, making rare-event forecasting a key challenge. We train a LightGBM gradient-boosted tree model, which effectively handles class imbalance while capturing nonlinear interactions between political variables. To ensure that the model generalizes to unseen time periods, we implement rolling-origin temporal cross-validation, strictly limiting training data to historical observations and preventing information leakage. We evaluate model performance using ROC-AUC, which measures ranking accuracy; AUPRC, which assesses precision-recall tradeoffs for rare-event classification; and the Brier score, which quantifies probability calibration.

Results

Our models strongly predict HLTA issuance, confirming that MLP's event data captures meaningful geopolitical risk indicators. The six-month forecast achieves an ROC-AUC of 0.87 and an AUPRC of 0.31, demonstrating robust predictive performance. Expanding to a rolling three-month window, where onsets within ± 1 month of the forecast are counted as correct, further improves results, yielding an ROC-AUC of 0.90 and an AUPRC of 0.57. These results indicate that MLP's civic space indicators provide leading signals of government security assessments, months before DOS formally recognizes them in its advisories.

An analysis of feature importance reveals that election irregularities, state of emergency declarations, protest activity, and censorship surges are the most influential predictors of HLTA issuance. Election irregularities, particularly those occurring 11 months before an advisory onset, emerge as a key signal of impending instability. This suggests that government manipulation of elections—whether through fraud, voter suppression, or other irregularities—often sets off political instability that escalates over the following year. Declarations of state of emergency, especially those occurring one to three months before advisory issuance, serve as immediate precursors, reflecting rapid government responses to deteriorating security conditions. Protest activity within a six-month window is also highly predictive, indicating that sustained civil unrest frequently precedes formal security warnings. Finally, increases in government censorship, particularly those occurring six to ten months before advisory onset, signal mounting repression, which may heighten political tensions and trigger future instability.

To assess real-world forecasting utility, we trained the model on data through December 2023 and generated predictions for advisory onsets in March and June 2024. The model correctly flagged Bangladesh for June 2024, which later received an HLTA due to the July Revolution. It also identified Zimbabwe and Liberia as high-risk for March 2024, and while DOS did not issue an HLTA for these countries, Zimbabwe expelled USAID staff in March, prompting U.S. officials to issue multiple security statements. These cases highlight the practical value of early warnings derived from high-frequency civic space indicators, even when they do not always align with official government actions.

Limitations

Although *HQMARC's* "medium-data" approach gives a much more reliable representation of domestic media markets in aid-receiving countries, there are several important limitations. First, stories from more recent years are easier to collect than older stories, so the total number of stories will tend to trend up over time. Focusing on event salience, rather than the raw number of articles reporting on each event type, attempts to mitigate the influence of these trends on our measures of civic space activity.

Second, domestic sources that are more difficult to scrape are less likely to be included. Some high-quality sources in low-resource countries have extremely poor website architecture, making them extremely difficult to collect data from. Third, news organization also have their own biases. For example, their coverage is much stronger in cities than in more rural areas and many international media outlets bias their coverage towards English-speaking countries. Despite these limitations, *HQMARC* is a powerful and flexible tool for understanding how events are shifting within developing countries at high-frequency.

Another key limitation of our data is *ML4P's* reliance on media attention as a proxy for the importance of an event. This assumption introduces potential biases, as media coverage is influenced by editorial priorities, political pressures, and audience interests rather than only the objective significance of an event. Some critical events may receive limited coverage due to competing news cycles, censorship, or media ownership structures, leading to under-representation in our dataset. Conversely, sensational or high-profile stories might be disproportionately amplified, skewing the perceived relevance of events in ways that do not necessarily reflect their actual impact.

Another limitation arises from the normalization process, which forces competition between different event categories within a given timeframe (e.g., a month). When an exceptionally large event dominates media coverage—e.g., a major political crisis, a natural disaster, or a global pandemic—other significant but less dramatic events may appear relatively unimportant in our data. This effect can lead to the under-detection of meaningful events that might otherwise register as notable shocks in a less crowded news environment.

Conclusion

As democratic institutions face unprecedented challenges in the current era of autocratization, understanding the day-to-day dynamics of civic space has become critical for both academic research and policy intervention. This paper addresses a fundamental gap in our ability to study these dynamics by introducing the *ML4P* dataset—the first comprehensive, high-frequency measure of civic space events across developing countries.

Our contribution is both empirical and methodological. Empirically, *ML4P* provides monthly data on 20 civic space events across a large sample of developing countries from 2012 to 2024, constructed from over 120 million articles published by 348 hand-picked domestic media outlets publishing in 36 languages. This represents an unprecedented scale of coverage for civic space monitoring, with 95% of articles sourced from domestic rather than international outlets. Methodologically, we demonstrate that combining human-supervised web scraping with open-source transformer models can achieve both comprehensive coverage and high classification accuracy while maintaining cost-effectiveness and transparency.

Our validation exercises confirm that *ML4P* successfully captures real-world political dynamics, from COVID-19 lockdowns and electoral cycles to coups and democratic crises. Critically, we document systematic biases in international media coverage, finding that international sources exhibit weak correlation with domestic reporting and entirely miss major events. These findings have profound implications for existing event datasets that rely heavily on international sources.

The predictive utility of our data is demonstrated through our travel advisory forecasting model. This validates that *ML4P* captures meaningful early warning signals of political instability rather than mere reporting artifacts.

This work opens several important avenues for future research. First, *ML4P* enables new studies of democratic backsliding mechanisms by providing the temporal granularity necessary to trace how specific events contribute to broader regime changes. Second, the dataset facilitates research on contentious politics, media behavior under authoritarianism, and the effectiveness of democracy support programs.

For policymakers, our findings underscore the critical importance of incorporating domestic media perspectives into intelligence and assessment processes. The systematic differences we document between international and domestic coverage suggest that relying on international sources alone can lead to fundamental misunderstandings of political conditions on the ground. Investment in human-supervised data collection infrastructures, as demonstrated by *HQMARC*, represents a crucial complement to automated systems.

Future research should expand *ML4P's* geographic coverage and integrate *ML4P* with other high-frequency political indicators could enhance our understanding of how civic space dynamics interact with economic conditions, social movements, and international interventions. *ML4P* represents more than a new dataset—it provides a new lens for understanding how democracy erodes and civic space contracts in real time. As authoritarian movements continue to challenge democratic institutions worldwide, tools like *ML4P* become essential for both documenting these processes and developing effective responses to defend civic space and democratic governance.

References

ACLED. 2023. "Adding New Sources to ACLED Coverage." Knowledge Base Article. Armed Conflict Location & Event Data Project. https://acleddata.com/knowledge-base/adding-new-sources-to-acled-coverage/.

Arendt, Florian. 2024. "The Media and Democratization: A Long-Term Macro-Level Perspective on the Role of the Press During a Democratic Transition." *Political Communication* 41 (1): 26–44.

Baum, Matthew A, and Yuri M Zhukov. 2015. "Filtering Revolution: Reporting Bias in International Newspaper Coverage of the Libyan Civil War." *Journal of Peace Research* 52 (3): 384–400. https://doi.org/10.1177/0022343314554791.

Besley, Timothy, and Robin Burgess. 2002. "The Political Economy of Government Responsiveness: Theory and Evidence from India." *The Quarterly Journal of Economics* 117 (4): 1415–51.

Boese-Schlosser, Vanessa A, Nazifa Alizada, Martin Lundstedt, Kelly Morrison, Natalia Natsika, Yuko Sato, Hugo Tai, and Staffan I Lindberg. 2022. "Autocratization Changing Nature?" *Democracy Report*.

Boschee, Elizabeth, Premkumar Natarajan, and Ralph Weischedel. 2012. "Automatic Extraction of Events from Open Source Text for Predictive Forecasting." In *Handbook of Computational Approaches to Counterterrorism*, 51–67. Springer.

Brandt, Patrick T, Sultan Alsarra, Vito J DOrazio, Dagmar Heintze, Latifur Khan, Shreyas Meher, Javier Osorio, and Marcus Sianan. 2024. "ConfliBERT: A Language Model for Political Conflict." *arXiv Preprint arXiv:2412.15060*.

Brechenmacher, Saskia, and Thomas Carothers. 2019. "Civic Freedoms Are Under Attack. What Can Be Done?" https://carnegieendowment.org/posts/2019/10/civic-freedoms-are-under-attack-what-can-be-done?lang=en.

Bridges, Lauren. 2019. "The Impact of Declining Trust in the Media." Ipsos. https://www.ipsos.com/en-uk/impact-declining-trust-media.

Brimicombe, C. 2022. "Is There a Climate Change Reporting Bias? A Case Study of English-Language News Articles, 2017–2022." *Geoscience Communication* 5 (3): 281–87. https://doi.org/10.5194/gc-5-281-2022.

Cheng, Cindy, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. "COVID-19 Government Response Event Dataset (CoronaNet v. 1.0)." *Nature Human Behaviour* 4 (7): 756–68.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, et al. 2023. "V-Dem [Country-Year/Country-Date] Dataset V13." Varieties of Democracy (V-Dem) Project. https://doi.org/10.23696/vdemds23.

D'Ignazio, Catherine, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. "CLIFF-CLAVIN: Determining Geographic Focus for News Articles." In *NewsKDD: Data Science for News Publishing, at KDD 2014*. https://hdl.handle.net/1721.1/123451.

Daphi, Priska, Jan Matti Dollbaum, Sebastian Haunss, and Larissa Meier. 2025. "Local Protest Event Analysis: Providing a More Comprehensive Picture?" *West European Politics* 48 (2): 449–63.

Earl, Jennifer, Andrew Martin, John D McCarthy, and Sarah A Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annu. Rev. Sociol.* 30 (1): 65–80.

Fotopoulos, Stergios. 2023. "Traditional Media Versus New Media: Between Trust and Use." *European View* 22 (2): 277–86.

Halterman, Andrew, Benjamin E Bagozzi, Andreas Beger, Phil Schrodt, and Grace Scraborough. 2023. "PLOVER and POLECAT: A New Political Event Ontology and Dataset." In *International Studies Association Conference Paper*.

Lee, Sangwon, Trevor Diehl, and Sebastián Valenzuela. 2022. "Rethinking the Virtuous Circle Hypothesis on Social Media: Subjective Versus Objective Knowledge and Political Participation." *Human Communication Research* 48 (1): 57–87.

Leetaru, Kalev, and Philip A Schrodt. 2013. "Gdelt: Global Data on Events, Location, and Tone, 1979–2012." In *ISA Annual Convention*, 2:1–49. 4. Citeseer.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." *arXiv Preprint arXiv:1907.11692*.

Lührmann, Anna, and Staffan I Lindberg. 2019. "A Third Wave of Autocratization Is Here: What Is New about It?" *Democratization*, 1–19.

Meléndez-Sánchez, Manuel, and Laura Gamboa. 2023. "How Guatemalans Are Defending Their Democracy." *Journal of Democracy*.

Mueller, Hannes, and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review* 112 (2): 358–75.

Quartey, P., A. Y. Owusu, C. Akwei, R. Atta-Ankomah, A. O. Crentsil, G. D. Torvikey, K. Asante,

J. Springman, R. T. Gansey, and E. Wibbels. 2023. "Radio and Social Media Assessment Report." Accra, Ghana: USAID Ghana MEL Platform.

Raleigh, Clionadh, Roudabeh Kishi, and Andrew Linke. 2023. "Political Instability Patterns Are Obscured by Conflict Dataset Scope Conditions, Sources, and Coding Choices." *Humanities and Social Sciences Communications* 10 (1): 1–17.

Raleigh, Clionadh, Rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47 (5): 651–60.

Romero, Diego. 2024. "Stopping Democratic Backsliding: The 2023 Guatemalan Elections and *Movimiento Semilla*'s Rise." *Working Paper*.

Schäfer, Svenja, and Christian Schemer. 2024. "Informed Participation? An Investigation of the Relationship Between Exposure to Different News Channels and Participation Mediated Through Actual and Perceived Knowledge." *Frontiers in Psychology* 14: 1251379.

Schrodt, Philip A., Deborah J. Gerner, and Omur Yilmaz. 2012. "CAMEO Event Data Codebook." Codebook. Parus Analytical Systems. https://eventdata.parusanalytics.com/data.dir/cameo.html.

Schrodt, Philip, and Jay Yonamine. 2013. "A Guide to Event Data: Past, Present, and Future." *All Azimuth: A Journal of Foreign Policy and Peace* 2 (2): 5–22.

Schwartz, Rachel A, and Anita Isaacs. 2023. "How Guatemala Defied the Odds." *Journal of Democracy* 34 (4): 21–35.

Study of Journalism, Reuters Institute for the. 2019. "Digital News Report: India Supplementary Report." Reuters Institute, University of Oxford. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-03/India_DNR_FINAL.pdf.

Tarr, Alexander, June Hwang, and Kosuke Imai. 2023. "Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study." *Political Analysis* 31 (4): 554–74.

Thaler, Kai M, and Eric Mosinger. 2022. "Nicaragua: Doubling down on Dictatorship." *Journal of Democracy* 33 (2): 133–46.

U.S. Agency for International Development. 2022. "Civil Society Organization Sustainability Index for Europe and Eurasia 2022." Report. Washington, DC: U.S. Agency for International Development; FHI 360; International Center for Not-for-Profit Law. https://csosi.org/.

Waldner, David, and Ellen Lust. 2018. "Unwelcome Change: Coming to Terms with Democratic Backsliding." *Annual Review of Political Science* 21: 93–113.

World Justice Project. 2024. "World Justice Project Rule of Law Index 2024." Washington, D.C.: World Justice Project. https://worldjusticeproject.org/rule-of-law-index/.