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

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

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 66 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 yields three achievements: first, our corpus provides unprecedented coverage of reporting by developing country media outlets, addressing biases in other media event data; second, the resulting monthly event data set covers a wide range of new civic space activities; and third, 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.

[1] Sabanci University, Turkiye
[2] University of Pennsylvania
[3] Univerity of Texas at Austin
[4] Stanford University
[5] Temple University
[6] Universidad Católica del Uruguay, Uruguay

*This paper is the result of a collaborative effort, and authors appear in alphabetical order. Donald A. Moratz took the lead in drafting the manuscript, Jeremy Springman in conducting the analysis, Erik Wibbels in guiding the project as PI, and all other authors contributed to the development and writing of the paper. 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.

1. 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). 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 these day-to-day struggles over political liberties, data for studying the events and processes that constitute this push-and-pull is limited. Existing measures of civic space rely largely on annual, expert-coded indicators that classify the nature of political regimes (Coppedge et al. 2023; U.S. Agency for International Development 2022; World Justice Project 2024). While these regime indices have opened new domains of rigorous research, they are not designed to provide insight into the quotidian politics where battles over civic space actually take place.

This article introduces the Machine Learning for Peace (*ML4P*) dataset, which provides monthly data on 20 civic space events across 66 developing countries from January 2012 through December 2024. By offering a dynamic view of where and when these events occur, *ML4P* represents an important advance in understanding the civic space dynamics that underpin broader regime change. *ML4P* is constructed from the High-Quality Media from Aid Receiving Countries (*HQMARC*) corpus, an original collection of articles scraped from 354 prominent *domestic* media outlets publishing in 40 languages. We supplement these domestic outlets with content scraped from 12 regional and 15 global outlets. 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. This process proves particularly valuable for domestic news sources, whose websites are less stable than international outlets. Our efforts yield 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 corpus composition with superior linguistic diversity and coverage of high-quality developing-country sources. Of course, *HQMARC's* size and linguistic diversity makes human classification prohibitively expensive. To produce *ML4P's* structured data on civic space events, we apply open-source computational tools to translate and extract information from each article, identifying the main event being reported on and the country in which the event occurs.

«««< Updated upstream This paper proceeds as follows. Section 2 discusses how *ML4P* complements existing regime data and opens new research avenues. Section 3 details our data production process and methodological advantages, highlighting *HQMARC's* superior coverage compared to other major media repositories. Section 4 introduces the *ML4P* data and presents validation exercises confirming its quality and underscoring the importance of relying on domestic rather than international media. The final sections address *ML4P's* limitations and explain why, despite them, it remains a valuable resource for studying democratic backsliding, contentious politics, and media

[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.

behavior. ======= This paper proceeds as follows. Section 2 discusses how *ML4P* complements existing regime data and opens new research avenues. Section 3 details our data production process and methodological advantages, highlighting *HQMARC's* superior coverage compared to other major media repositories. Section 4 introduces the *ML4P* data and presents validation exercises confirming its quality and underscoring the importance of relying on domestic rather than international media. The final sections address *ML4P's* limitations and explain why, despite them, it remains a valuable resource for studying democratic backsliding, contentious politics, media behavior, and crisis response. »»»> Stashed changes

2. 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 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 provide information about levels of democracy over time and space and capture distinct features of regimes, ranging from freedom of the press, rule of law, the ease of civic organizing, and beyond.

While V-Dem has improved the rigor of annual indices, such measures are ill-suited to the everyday politics where civic-space battles unfold. Ultimately, annual changes in the nature of regimes are the result of specific actions and events occurring at specific moments in time. Existing measures capture the cumulative impact of civic events over 12-month periods. Our project complements them by tracing the shorter-term events—often unfolding over days or weeks—that drive the broader shifts recorded in annual indices. Hungary's democratic erosion since 2010 illustrates the point: the 2010 media law centralized outlets, the 2011 constitutional reforms packed the Court, and the 2012 electoral law gerrymandered districts. Each mattered on its own, yet annual indices smooth over the mechanisms and timing of backsliding. *ML4P* is designed to shift analytical focus to these fast-paced civic events that underlie regime change.

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; Sundberg and Melander (2013)), the Political Event Classification, Attributes, and Types (POLECAT; Halterman et al. (2023)) dataset, 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 only a modest, contentious slice of struggles over civic space. Neither covers the legal changes, civic activism, press restrictions, corruption or election irregularities that regularly rock civil society. Alternatively, GDELT relies on the Conflict and Mediation Event Observations (CAMEO) coding ontology and covers a broad range of events, but it focuses on inter-state disputes and strategic interactions (P. A. Schrodt, Gerner, and Yilmaz 2012), classifies events using a complex, rigid system, and relies on limited and often dated actor dictionaries. POLECAT relies on the powerful, flexible PLOVER ontology, but

it too is designed largely to capture inter-state and strategic international interactions (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 civic events. Indeed, we define 19 civic space event types, ranging from political arrests and censorship to corruption, legal actions, and legal changes (see Appendix B for a complete list of event categories). We also code disasters and election activity—episodes aspiring autocrats often invoke to justify curbs on civil society. Although protests and lethal violence are tracked elsewhere, *ML4P* provides the first systematic coverage of most other event types.

Together, these events provide a rich monthly portrait of civic-space contestation and offer the potential for new research on regime dynamics. By capturing 19 distinct civic space events at monthly frequency across 66 countries, *ML4P* enables analysis of both the outcomes of democratic backsliding in new detail and the specific mechanisms and temporal dynamics through which regimes change. Our approach also allows researchers to measure the salience of different kinds of events in domestic media and is flexible enough to quickly apply future changes to coding criteria and/or add entirely new event types to the entire corpus.

3. Constructing ML4P

Social scientists rely heavily on media to produce event data (P. Schrodt and Yonamine 2013). While the shortcomings of this approach are well documented (Daphi et al. 2025; Earl et al. 2004), monitoring media reports remains the best means available to track the occurrence of many events across a wide range of contexts. Evidence suggests that access to traditional media effectively increases citizen knowledge of 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).

Efforts to create event data from media have long faced two obstacles: reliance on human coders to extract information from unstructured text, which limited scale and created lags. 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, a model that makes recoding costly and slow (ACLED 2023). Recent advances in machine learning now allow accurate automated coding, overcoming these constraints (Tarr, Hwang, and Imai 2023; Brandt et al. 2024; Halterman and Keith 2024; Halterman et al. 2023; Mueller and Rauh 2018).

Second, reliable repositories of high-quality, domestic media corpora are difficult and costly to build. As a result, many prominent event datasets rely heavily on international rather than country-specific sources, as evidenced by their limited linguistic diversity (Raleigh, Kishi, and Linke 2023). Reporting by international and regional media outlets on political events in developing countries contains significant biases (Baum and Zhukov 2015), even in coverage of natural disasters (Brimicombe

2022). *HQMARC* addresses this by scraping from a curated list 354 prominent *domestic* media outlets publishing in 40 languages; over 95% of its articles come from local media outlets.

Furthermore, most other projects rely on private aggregators like Factiva or LexisNexis, which source in a limited number of languages and provide inconsistent coverage due to erratic changes in licensing agreements that researchers can rarely account for.[2] For instance, our analysis of all sources available from the LexisNexis University archive shows that for six *ML4P* countries it includes no domestic outlets, 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 Section 6 shows, domestic and international outlets differ substantially in what they report, with many domestically important events missing from international coverage.

Alternatively, "big data" 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 as we show below, they fail to achieve comprehensive or consistent capture from many domestically important news sources. For example, while GDELT crawls a massive number of sources publishing in more than 100 languages, the lack of human oversight means that the sources they pull from changes constantly (Raleigh, Kishi, and Linke 2023). Below we show that the large-scale crawlers capture a small share of the total articles published by most sources. We further document that their use of automated parsers to extract metadata produces inaccuracies in critical fields, such as the date on which articles are published.

To address these issues, *ML4P* combines recent advances in automated text analysis with *HQMARC's* curated corpus of news. The core of *HQMARC's* approach is to identify a curated list of critical domestic sources for each country and process them with a customized workflow to achieve comprehensive capture of everything published by those sources. This approach allows researchers to calculate the share of all articles published by a given source (or sources) that covered a specific type of event. As a result, this corpus can be used to measure the salience of events over time.

The result is a 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 step in the pipeline.

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 an infrastructure designed to (1) comprehensively capture sources' full publication history and (2) maintain accurate metadata. This process involves three main steps:

[2]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 salience of events. Similarly, 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).

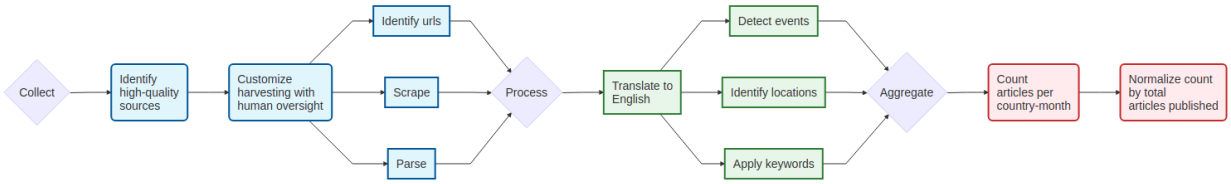

Figure 1: ML4P data production pipeline. Blue nodes capture steps in the construction of the *HQMARC* corpus. Green and Red notes capture the data processing and aggregation steps, respectively, in the construction of the *ML4P* event data.

1. **Identify High-Quality Domestic Sources:** We compile a list of local news sources with machine-scrapable websites by consulting directories of each country's media market (e.g., university library guides, Reporters Without Borders) as well as partners working in international NGOs and local civil society organizations. We conduct a desk review of each source's partisan affiliation by consulting reports on media ownership in the outlet's country (see Appendix A Section 4). In very repressive countries, we occasionally include sources based outside their home countries. For example, we include *El Faro*, a leading Salvadoran independent outlet that relocated its headquarters to Costa Rica due to government persecution.

   From this initial list, we select sources whose online archives extend as far back as possible, preferably to 2012. We aim for at least 3–5 local sources per country, yielding several thousand articles per month.[3] We 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.). When these methods fail we use more intensive tools, such as simulated infinite clicking 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.

3. **Develop Custom Scrapers and Parsers:** We then deploy scrapers and parsers tailored to each website's structure and publishing practices. These tools can bypass common barriers such as robot blockers, which affect roughly 15% of our sources. These source-specific scrapers minimize data loss and ensure accurate capture of critical metadata (e.g., publication date).

4. **Monitor and Update Quarterly:** Finally, we evaluate scraper and parser performance every 90 days, adapting to changes in website architecture. This monitoring helps detect when a source reduces its publication frequency, alters its website architecture, or shuts down entirely.

Appendix A provides comprehensive documentation of the *HQMARC* corpus. It provides the geographic distribution of domestic and regional media outlets across our sample, the linguistic diversity of the corpus, and the inventory of news sources by country. We demonstrate the importance

[3] In cases where a source's publication volume declines drastically or ceases entirely, we follow standardized replacement procedures.

of this custom workflow with a case study comparing *HQMARC's* coverage with that of several "big data" media corpora. We compare *HQMARC's* coverage of three prominent Bangladeshi news outlets to that of GDELT and Internet Archive. We focus on Bangladesh because: first, these outlets publish a high volume of articles, making them likely to attract automated crawlers; and second, their website architecture is straightforward, maximizing the likelihood that crawlers and automated parsers will accurately retrieve articles. As a result, we regard these outlets as a "best-case scenario" for large-scale media repositories.

HQMARC's coverage begins in 2013 for one source and in 2015 for the other two. GDELT does not have any articles published before 2019 for any of the three sources and thereafter captures many fewer articles. For the source with the smallest disparity, GDELT retrieves an average of 2,100 articles per month, compared to 2,500 in *HQMARC*. GDELT also includes numerous broken links, duplicate articles, etc. that *HQMARC's* human review removed. Moreover, GDELT's five-second delay per query makes it time-consuming to scrape a full historical archive of this size. Internet Archive achieved coverage similar to that of *HQMARC*, but more than *half* of the urls were broken. Furthermore, collecting URLs from Internet Archive for 2019–2023 required roughly two weeks for a single source.

The advantages of *HQMARC* extend beyond coverage. Big data media repositories rely on generalized scraping and parsing tools without human oversight. Figure 2 highlights one of the many ways that this can introduce errors. The figure shows a large spike in articles published by major outlets in Ghana and Zambia. On the left, this spike captures a genuine increase in articles published by `ghanaweb.com`, which resulted from a Google grant that enabled the outlet to expand its reporting. 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 on more than 1.5 million *unique* urls on the source's website. In both cases, we noticed the spike in publication volume and investigated the cause. Our human-in-the-loop approach effectively guards against such errors, enhancing the overall reliability of *HQMARC*. Importantly, such errors can be caused by a wide range of scraping and parsing failures, including dates that are incorrectly formatted 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 computational tools to classify the text of articles stored in the *HQMARC* corpus. We describe each step in this process, including translating article text, identifying the primary locations, classifying articles into relevant events, ensuring the political relevance of events, measuring the salience of events at the country-month level, and detecting months with high levels of civic activity.

Translating Non-English Text

After scraping articles, we translate the first 600 characters of all non-English articles into English using neural machine translations (NMT) through Hugging Face or OpenNMT.[4] To select translation models, we sample articles in each language, run them through all available models hosted

[4]While multilingual transformer models capable of classifying events directly in multiple languages exist, these models currently do not support the diversity of languages 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.

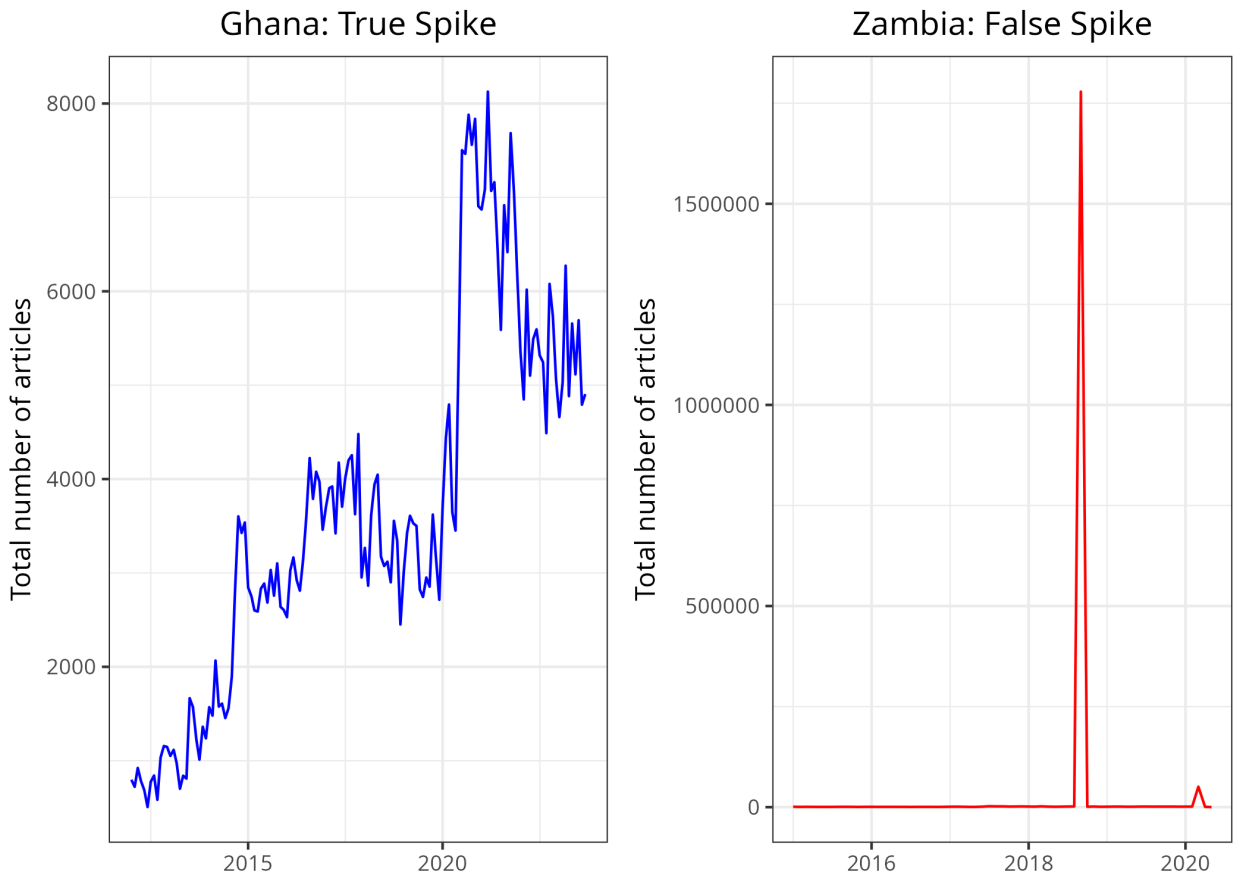

Figure 2: Changes in the volume of articles across two sources.

by Hugging Face, and assess whether the translations are sufficiently clear to identify the main event being reported on. When multiple models produce satisfactory results, we select the one with the clearest sentence-level translations, checking against Google Translate to detect any loss of contextual detail. For every language in our corpus, we were able to find translation models that generate translations clear enough for humans and our event classifier to identify the key event being reported.

Identifying Locations

Both international and domestic outlets report on events taking place across many countries. 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. For domestic sources, we use the CLIFF-CLAVIN geoparser with the GeoNames ontological gazetteer to identify geographic entities (e.g., states, cities, towns) mentioned in the text.[5] For each location, we use the CLIFF API to retrieve the location's country code and assign the article to that country.

Classifying Civic Space Events

To classify articles into 19 civic space event categories, we fine-tuned an open-source, transformer-based RoBERTa large language model (Liu et al. 2019).[6] To fine-tune the model, we built a double human-coded training dataset of 6,475 newspaper articles originally published in both English and non-English languages. As reported in Appendix B, the model achieves 82% out-of-sample accuracy—comparable to intercoder reliability—with most misses coming from articles reporting on multiple events. To reduce noise associated with background and contextual content, we classify only the title and first 600 characters.[7] In a few cases, we allow dual classifications for overlapping events (e.g., corruption with arrests). For several categories, we further apply a targeted keyword filter to eliminate common false-positives. See Appendix B Section 1 for a definition and examples for each event category and Appendix B Section 2 for a description of keyword filtering.

We then deploy a second classifier to exclude articles that fit an event category but lack political relevance. For instance, the first model cannot distinguish politically salient arrests from routine criminal cases. To address this, we fine-tuned a 'political relevance' classifier using transfer learning from our RoBERTa model and a double human-coded dataset of 2,938 articles. The model, which outputs a binary 0/1 for relevance, achieves 0.87 accuracy.

[5] CLIFF-CLAVIN integrates the GeoNames database (D'Ignazio et al. 2014). For technical details on CLIFF, see: CLIFF Annotator. 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.

[6] Recent research shows that closed-source LLMs perform only moderately better, at far higher cost (Andrade et al. 2024), while RoBERTa performs well for most political science applications (Timoneda and Vera 2025) (e.g., the protest-classification pipeline developed by Haunss et al. (2025) with fine-tuned XLM-RoBERTa), and consistently outperform dictionary-based approaches (Wang 2024).

[7] This corresponds to the title and first 2–3 sentences. Extensive testing suggests that adding more text reduces performance by introducing context that obscures the main event.

Measuring Event Salience

Finally, we aggregate the data to the country-month level, normalizing the count of articles reporting on each event type by the total number of articles published in that country-month. The resulting ML4P measures capture the monthly share of news devoted to each event category, made possible by *HQMARC's* capture of outlets' full publication history. This ratio reflects the *relative importance* of each event type, rather than coding individual events, and allows us to assess trends over time while accounting for shifts in overall reporting volume as sources enter or leave the database (ACLED 2023).

Detecting Major Event Shocks

We supplement those measures with an indicator variable identifying months when major civic space events occurred. To do so, we developed an ensemble algorithm to detect large increases in the share of reporting dedicated to event categories. We refer to these increases as *shocks*. We label a month as having a shock when either the statistical or neural network model detects one. Full details on model design, parameter tuning, and validation are provided in Appendix E.

4. Data Description and Validation

In this section, we present the data, show the importance of *HQMARC's* reliance on domestic news sources for a rich portrayal of civic activity in countries, and results from two validation exercises. Figure 3 shows cross-national variation in the most frequently reported-on civic space event type across four annual snapshots. Several temporal shifts are evident: in 2020, Natural Disasters dominated coverage in most coverage, which reflects coverage of COVID-19, while in 2024, several countries with high-profile national elections see Election Activity coverage dominate. See Appendix C for maps showing annual averages for all years in the data.

Nevertheless, many of the most compelling applications of the data involve dynamic, within-country analysis. To illustrate this, we focus on the 2023 Guatemalan general elections, where opposition candidate Bernardo Arévalo won a surprise victory despite institutional efforts to undermine him. The electoral period ran from January to August, concluding with a run-off election. As expected, the upper-left panel of Figure 4 shows a shock in Election Activity during this period. The upper-right panel captures attempts to disqualify Arévalo between the June 25 first round and the August 20 run-off, detecting shocks in Election Irregularities both before and after the election. The lower-left panel shows increased reporting on Legal Actions in September, corresponding to the Public Prosecutor's efforts to nullify the election results. The lower-right panel identifies a shock in reporting on the massive October protests in favor of a peaceful transition of power—mobilization which was widely seen as crucial to securing the transfer of power (Schwartz and Isaacs 2023; Meléndez-Sánchez and Gamboa 2023; Romero 2024). This case demonstrates *ML4P's* ability to capture fast-paced civic events underlying an attempted—and failed—autocratic turn.[8]

[8]Beyond Guatemala, we use the data to study discontinuities in press freedom in Tanzania (Adiguzel, Romero, and Wibbels 2025) and how mainstream vs. propaganda outlets respond to corruption scandals in El Salvador (Romero and Wibbels 203AD).

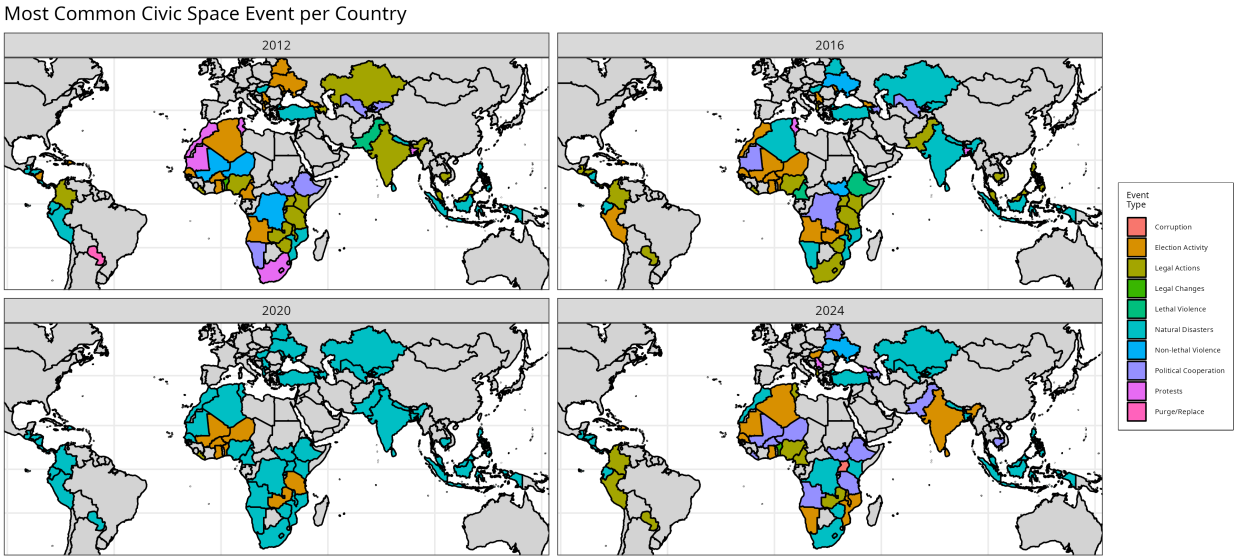

Figure 3: Most frequently reported civic event type per country-year for 2012, 2016, 2020, 2024. Countries are colored according to the civic event category with the highest reporting frequency in each year, showing the diversity of civic space concerns across different national contexts and revealing temporal shifts in civic space priorities.

Comparing International and Local Media

We now show that the *ML4P* data relies heavily on the domestic sources targeted by *HQMARC*, and that relying solely on international media would yield a very different view of civic space across our countries.

Figure 5 plots the share of articles in each event category from domestic versus international outlets. The stacked bars show the percentage of total articles about each event type from domestic (blue) versus international (green) sources; for most event types, domestic sources provide over 80% of the data, with even higher shares in some cases. The red points show correlations between domestic and international coverage by category. If both kinds of outlets covered the same events at the same time, correlations would be high. Instead, they are consistently weak to moderate, even for categories with greater international attention, such as Lethal Violence and Security Raids. For example, 83% of Arrest articles come from domestic outlets and just 17% from international ones, with a correlation of only 0.17.

The correlation in coverage between international and domestic sources does increase as countries receive more international media attention. Figure 6 shows the correlation within countries and across event categories. Importantly, although the correlation increases as countries receive more international coverage, the correlation remains low in nearly all cases. For example, India receives substantial international attention yet exhibits domestic-international correlations below 0.5.

An example from Indonesia illustrates the point. Between March and November 2024, the country saw major corruption scandals—including the PT Timah state enterprise corruption case in March, the Tom Lembong sugar import and Sahbirin Noor cases in October, and the Rohidin Mersyah case

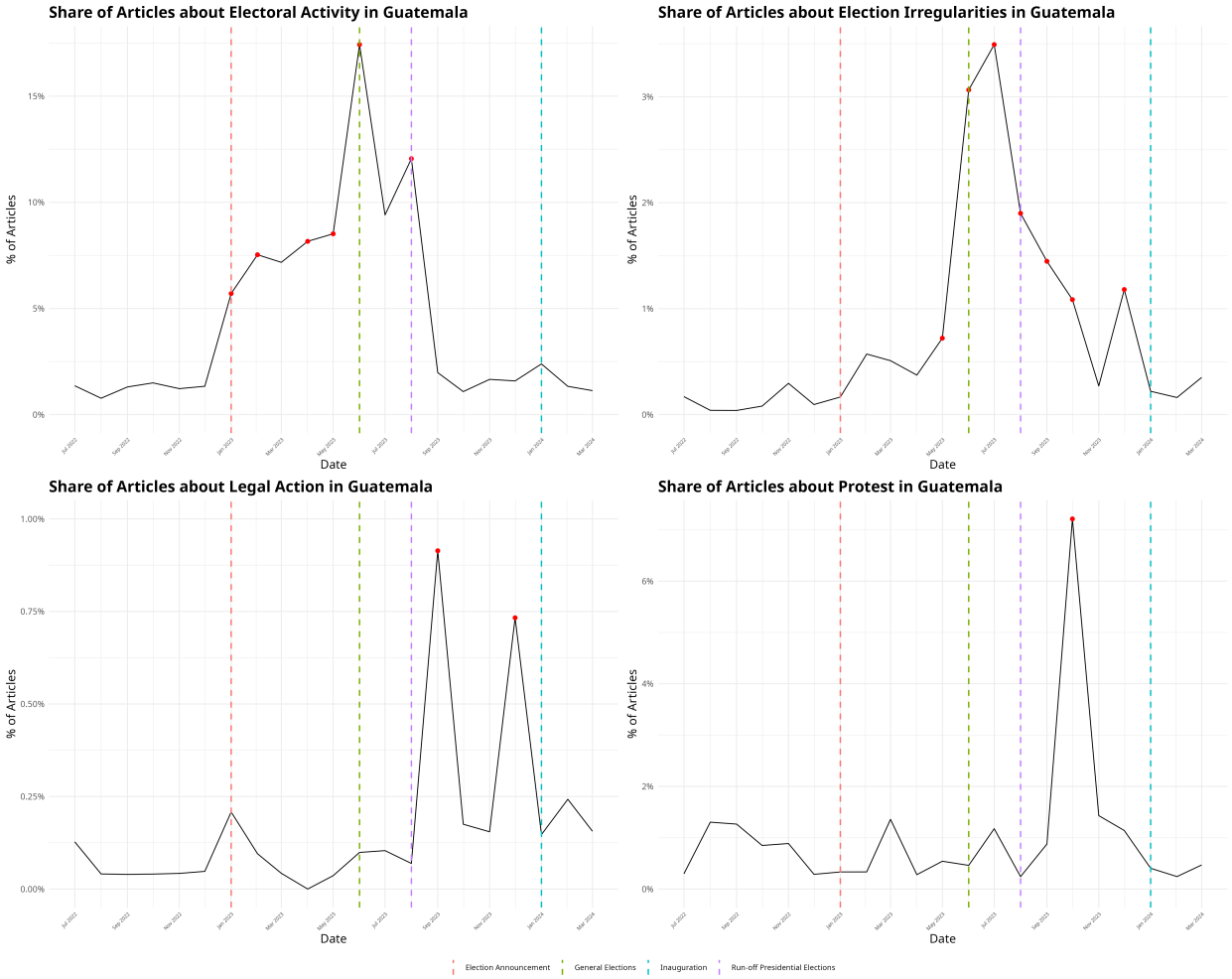

Figure 4: Elections in Guatemala 2023. The vertical lines in each panel represent key milestones in Guatemala's 2023 electoral cycle, beginning with the official election announcement in January 2023 and concluding with the presidential inauguration in January 2024. The general elections were held in June 2023, followed by the presidential runoff in August 2023.

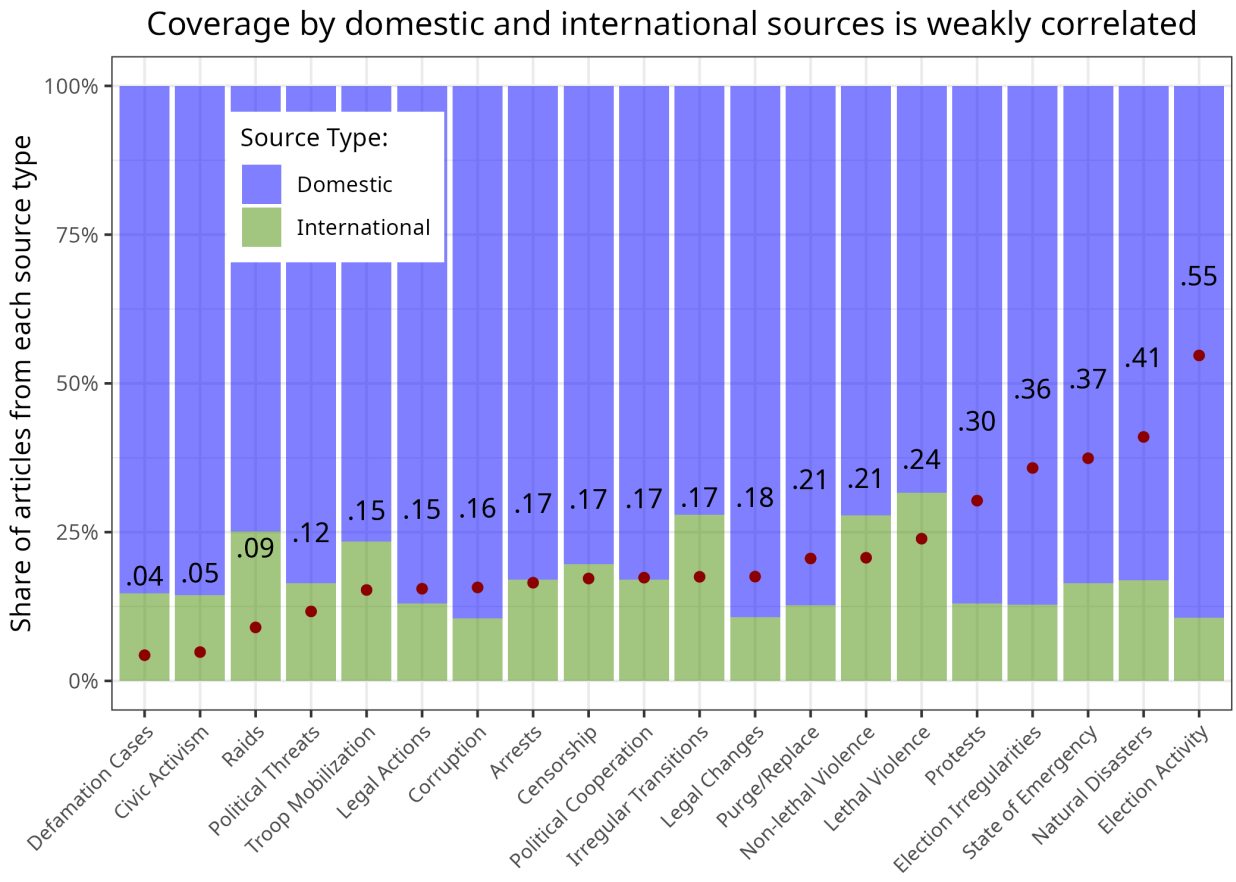

Figure 5: Proportion of civic event coverage between domestic and international sources. Stacked bars show the percentage of total articles about each civic event type that come from domestic (blue) versus international (green) sources. Events are ordered by correlation strength between domestic and international coverage (red points labeled with correlation coefficient).

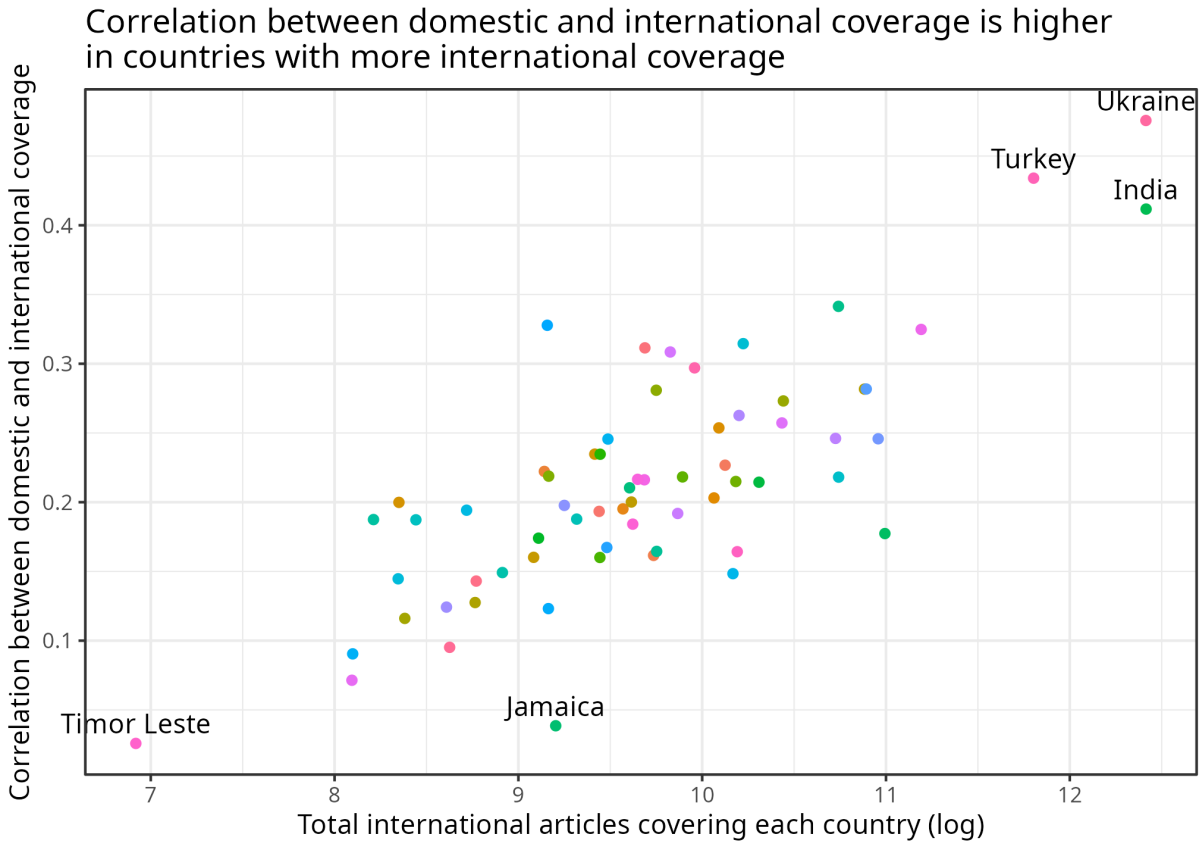

Figure 6: Relationship between international media attention and domestic-international coverage correlation. Each point represents one country. The x-axis (log scale) shows the total volume of international articles; the y-axis shows the correlation between domestic and international civic event coverage. Labeled cases highlight contrasts: Ukraine, Turkey, and India (high international attention, strong positive correlation) versus Timor Leste and Jamaica (low international attention, weak correlation).

in November. In May alone, the *HQMARC* domestic outlets published over 970 articles on corruption, while regional and international sources carried *none*. This discrepancy shows how relying on international media can obscure the salience of civic events and even miss them entirely. Across our dataset, we find 59 significant civic shocks in 23 countries with zero international coverage.

Major Event Validation

To validate our data's ability to detect major political events, we adopt two approaches. First, we assess whether months flagged as shocks align with real-world developments (true positives). We randomly selected five countries—Kosovo, Morocco, Angola, Mauritania, and Ukraine—spanning three regions and four languages (Serbian, Albanian, French, and Ukrainian). Across the three most recent months (300 possible event-months), our model flagged 40 as shocks. Because such months often include multiple events, we reviewed the top five per event-month. In 38 of 40 cases (95%), all top-five events matched the correct category (see Appendix F).

Our second approach identifies major world events expected to generate shocks in *ML4P* data and tests whether they appear (avoiding false negatives) by: examining a single historical event likely to generate media attention on one *ML4P* event category across countries, and assessing whether frequent and rare political events produce detectable spikes in relevant event categories. We present three examples of this validation approach below.

We first examine government responses to COVID-19 pandemic onset, particularly widespread implementation of school closures, curfews, and non-essential movement restrictions (Cheng et al. 2020). These lockdown measures should associate with State of Emergency event category shocks. As Appendix D shows, we detect State of Emergency count shocks across all dataset countries starting 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. We identify the most recent election (general, parliamentary, or presidential) for each Latin American country in our dataset. Figure 7 our shock detection models detect heightened electoral activity in months preceding these elections in all countries, including Nicaragua's electoral autocracy (Thaler and Mosinger 2022).

We also 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 was attempted or succeeded in the past decade. As Figure 8 illustrates, all 14 successful, attempted, or self-coups are associated with a shock in the Irregular Transition measure.

5. Use Case: Forecasting Travel Advisory Onsets with ML4P

In this section, we demonstrate *ML4P's* ability to capture civic space dynamics through an important forecasting application. Specifically, we show that high-frequency civic space indicators help predict the issuance of independent security assessments—U.S. Department of State (DOS) high-level travel advisories (HLTAs)—providing further validation that *ML4P* captures meaningful underlying political conditions.

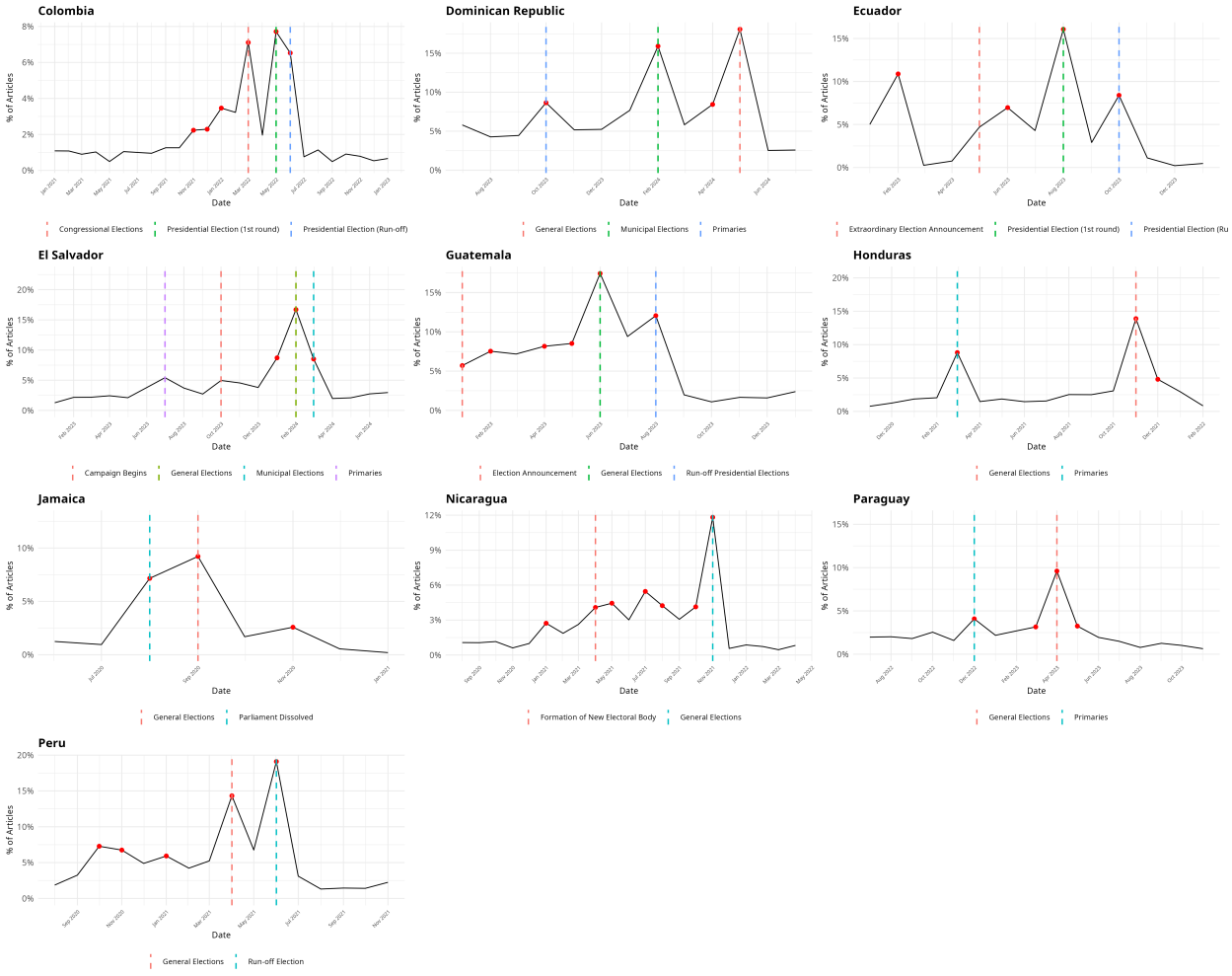

Figure 7: Elections and Electoral Activity (Latin America and the Caribbean). The vertical lines mark key milestones in each country's electoral cycle, such as primaries, congressional, or presidential elections. The red dots mark detected shocks.

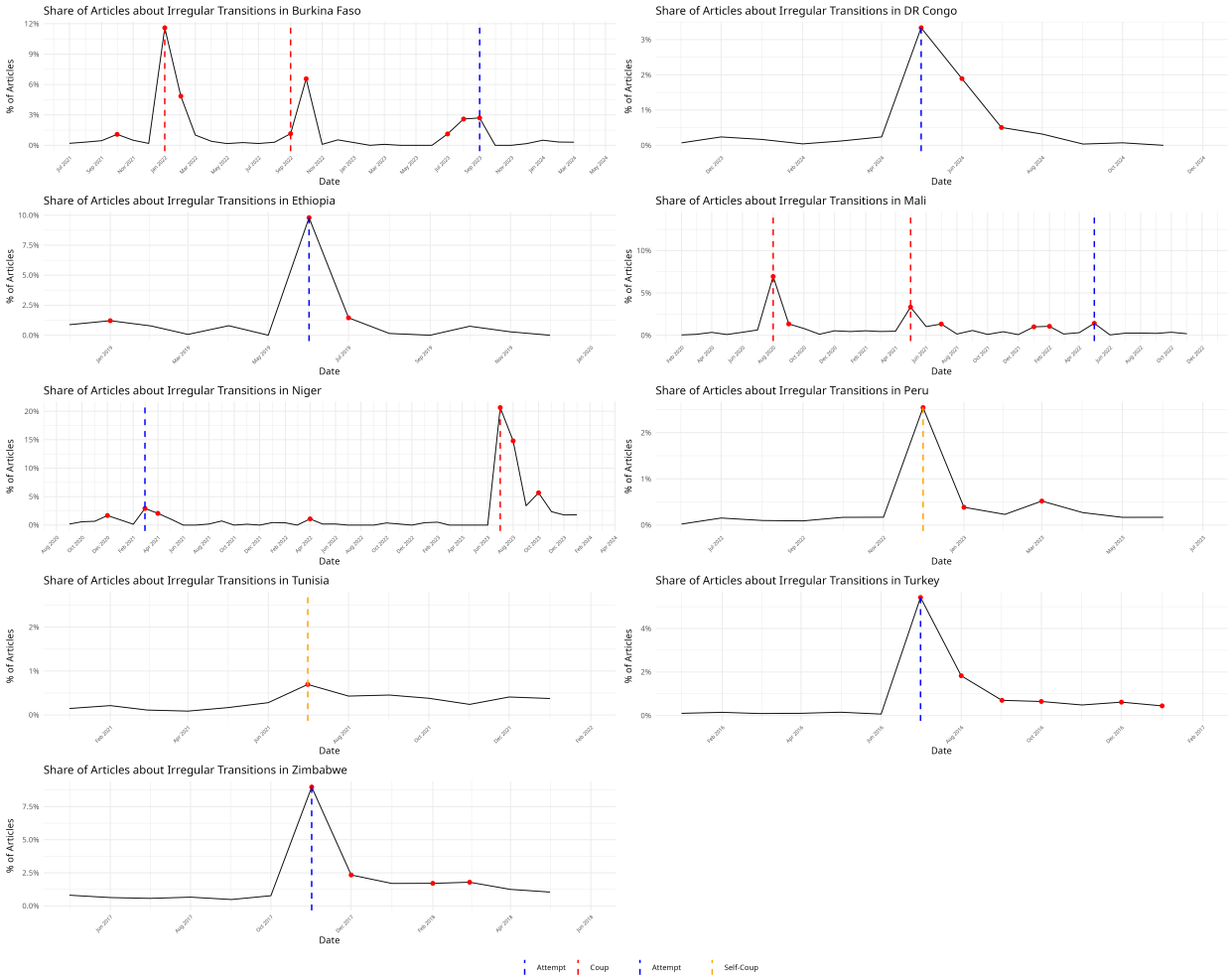

Figure 8: Coups and Self-Coups (Attempted and Successful). The vertical lines in each panel indicate the month in which a coup (successful or attempted) or a self-coup occurred in each country.

We construct a monthly panel covering 60 developing countries from 2012 through 2023 and model the onset of Level 3 ("Reconsider Travel") and Level 4 ("Do Not Travel") travel advisories by scraping DOS pages and filling gaps with the Wayback Machine..[9] These advisories serve as official DOS risk assessments and capture a diverse array of threats, including civil unrest, political repression, armed conflict, health emergencies, and crime waves. These advisories drive critical operational responses, including embassy closures, staff relocations, and security escalations.

We define a binary onset indicator equal to 1 in the first country–month an advisory appears and 0 thereafter, predicting travel advisory onsets. To forecast HLTA onsets, we use monthly event counts for all 20 of *ML4P's* event categories and a series of economic and country-specific controls (see Appendix G). HLTA onsets occur in only 1.42% of country months, making this a challenging forecasting target. We train a LightGBM gradient boosted tree model, which handles class imbalance and captures nonlinear interactions among variables, providing forecasts for 3- and 6-month horizons. To generalize to unseen periods, we use temporal cross-validation, restricting training data to past observations and preventing information leakage (Rodolfa, Lamba, and Ghani 2021).[10] We evaluate model performance using ROC-AUC (ranking accuracy), AUPRC (precision-recall tradeoffs for rare events), and the Brier score (probability calibration).

Our models strongly predict HLTA issuance, confirming that *ML4P's* event data captures meaningful geopolitical risk indicators. For the 3-month prediction horizon, our model achieves a ROC-AUC of 0.87, AUC-PR of 0.26, and a Brier score of 0.14, compared to a dummy AUC-PR of 0.01. When we count predictions as correct if an onset occurs within ± 1 month of the forecast, the performance improves to a ROC-AUC of 0.89, AUC-PR of 0.54, and Brier score of 0.12, with a slightly higher dummy AUC-PR of 0.02. We achieve similar performance on the 6-month horizon (see Appendix G, which also includes discussion of OOS forecasts for 2024).

Feature importance analysis shows that for 3-month forecasts, martial law declarations in the past two months are the strongest predictors of HLTA onsets. Other short-horizon predictors include protest activity and election irregularities. For 6-month forecasts, the top predictor is election-related activity 11 months before an onset, followed by martial law declarations and censorship. These patterns suggest that, while immediate security concerns drive short-term risk, longer-term forecasts are influenced by political processes, governance changes, and escalating repression.

6. Limitations

Although *HQMARC* gives a much more reliable representation of domestic media markets, it has several limitations. First, courtesy of uneven archiving practices by media outlets, stories from more recent years are easier to collect than older stories. Thus, the total number of stories tends to increase over time. Focusing on event salience, rather than the raw number of articles reporting on each event type, mitigates 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 sources in low-resource countries have extremely poor website architecture, making scraping extremely difficult. Third, news organizations have biases. For example, their coverage is much

[9] For consistency over time, we include all advisories before 2018 and Levels 3-4 after the introduction for the four-level system in 2018. Consultation with DOS consular staff suggests that pre-2018 advisories align closely with post-2018 Levels 3-4 in severity.

[10] See Soltani, Springman, and Wibbels (2025) for further details on the data and models.

stronger in cities than in more rural areas. Despite these limitations, *HQMARC's* focus on domestic media paints a much richer picture than that provided by international media coverage of these countries.

Two cautions arise from *ML4P's* reliance on media attention as event importance proxy. First, media coverage is influenced by editorial priorities, political pressures, and audience interests rather than just objective event significance. Critical events may receive limited coverage due to competing news cycles, censorship, or media ownership structures. Second, our normalization approach forces competition between event categories within any month. As coverage of some event types increases, others necessarily decrease. When exceptionally large events dominate media coverage other significant but less dramatic events appear relatively less important in our data.

7. Conclusion

As democratic institutions confront unprecedented challenges in an era of autocratization, understanding the everyday dynamics of civic space has become critical for both research and policy. Existing annual regime indices remain indispensable, but they inevitably smooth over the fast-moving political events that drive institutional change. *ML4P* addresses this gap by introducing the first comprehensive, high-frequency measure of civic space events across developing countries, built from more than 120 million articles published by 354carefully selected domestic outlets in 36 languages.

Our contribution is both empirical and methodological. Empirically, *ML4P* provides monthly measures of 19 civic space event types, with 95% of coverage drawn from domestic rather than international sources—an unprecedented scale for civic space data. Methodologically, we show that combining human-supervised web scraping with open-source transformer models yields comprehensive coverage and high classification accuracy while remaining cost-effective. Validation exercises confirm that *ML4P* captures real-world dynamics, from COVID-19 lockdowns to coups and democratic crises. We also document systematic biases in international media: correlations with domestic reporting are weak, and major events are often entirely absent. Furthermore, we demonstrated the predictive utility of *ML4P* through our travel advisory forecasting model, which shows that the data provides early warning signals of instability rather than mere reporting noise.

By offering temporal granularity and broad coverage, *ML4P* enables new research on the mechanics of democratic backsliding, contentious politics, media behavior, corruption, and political violence. It also allows comparative analysis of which democracies resist backsliding, and how international shocks translate into domestic political behavior. Moreover, future research could expand *ML4P's* geographic coverage and link it to other high-frequency indicators to deepen our understanding of how civic space dynamics intersect with economic conditions, social movements, and international interventions. For policymakers, the findings underscore the dangers of relying solely on international sources and highlight the value of investing in human-supervised infrastructures like *HQMARC*. More than a dataset, *ML4P* provides a new lens for understanding how civic space contracts and democracy erodes in real time, offering both scholars and policymakers an essential tool for documenting these processes and informing effective responses.

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/.

Adiguzel, Serkant, Diego Romero, and Erik Wibbels. 2025. "Autocratization and Media Responses to Government Repression of Journalism: Machine Learning Evidence from Tanzania." *Working Paper*.

Andrade, Claudio M. V. de, Washington Cunha, Davi Reis, Adriana Silvina Pagano, Leonardo Rocha, and Marcos André Gonçalves. 2024. "A Strategy to Combine 1stGen Transformers and Open LLMs for Automatic Text Classification." https://arxiv.org/abs/2408.09629.

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*.

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*.

Halterman, Andrew, and Katherine A Keith. 2024. "Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts." *arXiv Preprint arXiv:2407.10747*.

Haunss, Sebastian, Priska Daphi, Jan Matti Dollbaum, Lidiya Hristova, Pál Susánszky, and Elias Steinhilper. 2025. "PAPEA: A Modular Pipeline for the Automation of Protest Event Analysis." *Political Science Research and Methods*, 1–18.

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.

Rodolfa, Kit T., Hemank Lamba, and Rayid Ghani. 2021. "Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy." *Nature Machine Intelligence* 3 (10): 896–904. https://doi.org/10.1038/s42256-021-00396-x.

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

Romero, Diego, and Erik Wibbels. 203AD. "Propaganda and Political Scandals: Evidence from El Salvador." University of Pennsylvania: PDRI-DevLab. https://web.sas.upenn.edu/mlp-devlab/files/2023/10/SLV_Report_MLP2023.pdf.

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.

Soltani, Mahda, Jeremy Springman, and Erik Wibbels. 2025. "Forecasting High-Level Travel Advisories with Machine Learning for Peace Data." University of Pennsylvania: PDRI-DevLab. https://web.sas.upenn.edu/mlp-devlab/files/2024/11/Forecasting-DOS-Travel-Advisories-with-Machine-Learning-for-Peace-Data-V2-internal.pdf.

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.

Sundberg, Ralph, and Erik Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–32.

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.

Timoneda, Joan C, and Sebastián Vallejo Vera. 2025. "BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text." *The Journal of Politics* 87 (1): 347–64.

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.

Wang, Yu. 2024. "On Finetuning Large Language Models." *Political Analysis* 32 (3): 379–83.

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/.