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

Donald A. Moratz[†‡] Jeremy Springman[†‡] Erik Wibbels[†‡]
Serkant Adiguzel[§] Mateo Villamizar-Chaparro[¶] Zung-Ru Lin[‡]
Diego Romero[‖] Mahda Soltani[‡] Hanling Su[‡] Jitender Swami[‡]

October 31, 2024

Abstract

We live in a world where political regimes are under contestation every day. Despite the importance of day-to-day contestation over political 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 62 developing countries from 2012 to 2024. Using a corpus of over 100 million articles from more than 300 high-quality domestic media outlets and additional international and regional sources, 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 developing country media outlets. We use this data to identify major political events and track their importance for domestic politics, with applications for research on autocratization, media behavior, and more. This project builds on a flexible research infrastructure that updates the entire dataset every 90 days and can be quickly adapted to study new events at different levels of granularity.

*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. Check out our data dashboards at mlpeace.org.

[†]These authors contributed equally to this work.
[‡]University of Pennsylvania
[§]Sabanci University, Turkiye
[¶]Universidad Católica de Uruguay, Uruguay
[‖]Utah State University

1 Introduction

More than 5.4 billion people lived under autocracy in 2021, up from 3.5 billion in 2016 (Boese-Schlosser et al., 2022). Concentrated in the global south, this "third wave of autocratization" is narrowing civic space, limiting the ability of citizens to advocate for better governance, and weakening political accountability (Lührmann and Lindberg, 2019; Waldner and Lust, 2018). However, citizens in every country are contesting these attacks on fundamental rights by pushing for improvements in local conditions, such as access to public services, as well as progress on global issues, including climate change. Despite the importance of this day-to-day push-and-pull over the extent of citizen liberties and state control, there is surprisingly little data allowing us to study the events and processes that constitute this struggle.

This article introduces *Civic Space in Aid-Receiving Countries (CS-ARC)*, a new dataset tracking levels of activity across 20 critical civic space events for 62 developing countries from 2012 through 2024. We measure civic space activity by capturing monthly variation in levels of media attention across event types for each country. In most contexts, media attention is the best means available to track a wide range of major events. Historically, traditional media has been the main source of information about major events. 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 news is often distributed through radio or social media, the underlying news circulated on these platforms often comes from articles originally published by traditional news outlets (Quartey et al., 2023; Reuters Institute, 2019). Citizens across a wide range of countries also consistently express higher trust in news from traditional outlets (Fotopoulos, 2023; Bridges, 2019), and recent evidence suggests that traditional media contains more information about political events than other sources (Lee et al., 2022; Schäfer and Schemer, 2024).

CS-ARC is constructed from articles collected for the *High-Quality Media from Aid Receiving Countries (HQMARC)* corpus, an original repository of online news from 16 international and 12 regional sources, as well as a curated sample of nearly 350 of the most prominent domestic media outlets based across our sample of countries. Rather than using web-crawlers or pre-canned scraping tools, HQMARC uses custom web scraping and parsing to accurately capture each outlet's complete publication history. As we show in Section 3, this 'medium frequency' approach allows us to scrape each source with much greater accuracy and completeness than other popular big data media aggregators, such as GDELT, Common Crawl, and Internet Archive. Critically, our human-supervised scraping results in a corpus with a stable, well-understood composition, allowing us to understand changes in the proportion of media attention devoted to specific event within countries over time.

HQMARC includes more than 100 million articles published in more than 40 languages, which precludes the use of human coders to extract information from article text. To produce structured data for CS-ARC tracking civic space from this large volume of multi-lingual text, we use open-source computational tools to translate and extract information from each article, identifying the country in which events occur and the main event being reported on. Importantly, the large research infrastructure behind HQMARC and CS-ARC updates the entire corpus every 90 days and processes the data in a highly flexible manner that can be adapted to integrate more advanced tools as they become available and extract new types of information from text as needed.

In Section 2, we discuss how HQMARC is constructed and how CS-ARC uses this unique corpus to both detect major political events and assess their importance within countries' broader social contexts. We demonstrate the ability of open-source translation models and Large Language Models (LLMs) to accurately extract information from a highly diverse sample of media text at a large scale. This section also illustrates how the underlying corpus is also being used to produce similar data tracking a variety of other social phenomenon, including foreign influence and climate adaptation behaviors, and why it can be easily adapted to generate data at finer temporal and geographic granularity.

Section 3 discusses results from several data validation exercises. We show that HQMARC exhibits a much more accurate and stable than popular big-data repositories of media data and provides a much better foundation for event data based on media coverage. We then use CS-ARC to present original evidence on systematic differences between international and regional outlets compared to domestic outlets in the amount of focus devoted to different types of political events. Our findings have important implications for event data that is generated from predominantly international or regional media outlets, a common practice in the social sciences. We then show that CS-ARC reliably detects major political events in a sample of six developing countries.

Finally, we use CS-ARC to describe civic space across countries and show interesting variation over time. We also show how our measures of civic space activity respond to major political events around the world. In the Conclusion, we discuss why CS-ARC represents a valuable new resource and recommend applications for research on autocratization and democratic backsliding, political accountability and contentious politics, media behavior, crisis response, and program evaluation.

2 Constructing CS-ARC

Social science research relies heavily on data tracking the occurrence and timing of important events. Researchers typically turn to online media reports as the best available record of these events. However, the use of media reports to produce data tracking events has faced two perennial challenges.

First, producing structured data from the unstructured text of media reports is costly. Historically, human beings were needed to extract relevant information from text. This constrained the amount of text that could be processed and required significant lags between the occurrence of events and the production of structured data. Furthermore, the expense of human coders made it extremely costly to change or adapt coding rules in response to new information or changes in source material. Such changes would require humans to go back and re-read every article that was previously classified to implement updated instructions. Thankfully, recent advances in the computational tools available to researchers for processing text has alleviated these constraints.

Second, researchers have often relied on corpora composed primarily of international media or with poorly understood composition. With machine-coding, researchers are able to code vast amounts of text. However, accessing high-quality media covering events in a broad swathe of countries is difficult. Many sources of event data, such as ICEWS and POLECAT, rely primarily on international and English-language sources. However, reliance on these sources produces significant biases, especially for less-covered countries in the developing world.

Others, such as GDELT, Common Crawl, and Internet Archive, have taken a 'big data' approach, relying on web crawlers and automated scrapers to consume vast amounts of text from across the web. While these approaches do a better job of collecting text from less-covered countries, they suffer from inconsistent composition. Specifically, they typically capture only a fraction of news being published by the sources they target and often include inaccurate metadata on critical fields. This makes it difficult to account for the importance of events within a given media ecosystem. Because the true publication volume of articles published by constituent sources is not measured, users cannot assess the share of media attention devoted to events, omitting a critical piece of information that can indicate the salience of events within a specific national context. Importantly, this also affects news sourced from expensive private media aggregators, including Lexis Nexis, who exhibit erratic entry and exit by media outlets (and even specific sections of newspaper websites) due to complicated licensing agreements.

In this section, we describe a 'medium frequency' approach that applies recent advances in machine coding to a corpus of media reports sourced from high-quality outlets from a broad range of countries with stable and well-understood publishing practices. This innovative research infrastructure is highly-flexible, allowing us to detect a broad range of major political events and assess their importance within countries' social contexts while updating the data on a regular basis.

Building the HQMARC Corpus

CS-ARC is constructed by processing articles from the HQMARC corpus. HQMARC is an original repository of online news capturing the traditional news media ecosystem in a broad swathe of developing countries over more than a decade. Importantly, HQMARC captures the publication history of critical domestic media outlets with unprecedented accuracy and granularity.

To overcome the composition challenges discussed above, we developed a data collection infrastructure that captures the full publication history of local sources while ensuring accurate metadata. This process involves three distinct steps. First, we develop a list of high-quality local newspapers with , machine-scrapable websites. We begin this process by consulting publicly available information about each country's media market (e.g., lists maintained by university library guides and Reporters without Borders), as well was our partner organizations to identify high-quality local newspapers that publish online. We define a local newspaper as "high-quality" if it considered a reputable source of news within the country and it produces and publishes original news content, preferably daily.This definition excludes most state-owned newspapers, and in some cases, includes newspapers that report about a country from outside its borders. An example of the latter is El Faro, a Salvadorean independent online newspaper which moved its headquarters to Costa Rica as a result of persecution by the government.

From our initial list, we then select newspapers with a machine-scrapable archive that goes as far back as possible, preferably at least to 2012. We aim to have at least 3-5 local sources per country with a combined volume of at least several thousand articles per month. When publication volume decrease dramatically or ceases entirely, we have established procedures for replacing the source (available in Appendix XXX). Additionally, we supplement the output of local sources with articles published by XX reputable international and regional

sources sources.

Second, we develop custom scrapers and parsers tailored to the unique architecture and publication practices of each website. For many sources, this includes methods to bypass robot blockers (e.g., Cloudflare). Third, we carefully evaluate the performance of these scrapers and parsers for each source every 90 days. This allows us to adapt to changing website architecture over time and identify when sources reduce their publication volume or cease operations entirely. Currently, we are scraping an average of 4.6 local sources for each country with a combined publication volume of XXX articles per month. 81% of the close to 100 million articles in our database are published by local sources. In Section 3, we provide analysis of how our coverage of domestic sources compares to that of big data media aggregators and demonstrate how the reporting from these domestic sources differs from that of international and regional sources.

Processing Text Data from HQMARC

Drawing on news articles contained in the HQMARC corpus, we use a host of computational tools to extract information about civic space events from the raw text of articles. To accomplish this, we test and apply translation models to translate non-English publications into English. We also use open source geoparsing tools to extract all locations mentioned in the text to ensure that events are being attributed to the proper country. Finally, we use a fine-tuned large language model (LLM) to identify articles reporting on civic space or foreign influence events.

Specifically, we fine-tuned a RoBERTa model to detect these events by training and testing it on two corpora of human-coded newspaper articles hand built for the project. The first training data for the civic space event counts covered 6,475 (1,493 non-events and 4,982 events) articles over 20 event types. Since the project's inception, LLM use has exploded in popularity. Despite this, recent research has shown that costly, closed-source LLMs only perform moderately better at even complicated tasks than first-generation performers like RoBERTa, and usually only after costly fine-tuning (Andrade et al., 2024). In developing and deploying these models, we have highlighted the ability for researchers to apply rigorous analytical evaluation of text data from a large number of countries and languages using free, open-source LLMs. By adapting this approach for civic space, we have also demonstrated that these tools can be trained to extract information about a wide range of events from text, with the potential to monitor new types of events and media characteristics (ex. use of polarizing language) from the underlying data repository. By incorporating these data processing techniques into a robust and highly flexible data processing pipeline, we have provided a research infrastructure that will continue to produce high-quality data tracking important civic space and foreign influence events and which can be quickly expanded to provide new data on additional events as needed.

These first-generation models do not come without limitations. First, edge cases can pose serious challenges. Andrade et al. (2024) argue for the use of more complex models like GPT 4o for complicated edge cases. However, we've found that for our use case, we can gain significant improvements in accuracy by applying several basic keyword filters on classified text. As a result, several civic space event types use a keyword corpus to increase

accuracy of classification[1]. Second, these first-generation classifiers perform optimally at one task at a time. News salience is not always related to civic importance. In order to facilitate focus on events related to civic activity, we developed a second LLM to detect events that are directly related to civic space; for categories like arrest, this model helps to ensure that articles covering arrests that have no relevance to civic space are filtered out.

Translating Non-English Text

Given the well-documented biases in English-language news sources (Baum and Zhukov, 2015) even in relatively uncontroversial topics like natural disaster coverage (Brimicombe, 2022), we collect articles published in local languages. Currently, HQMARC features articles in over 40 languages in addition to English, ranging from high prevalence languages like Spanish and French, to less commonly spoken languages like Georgian, Amharic, and Kinyarwanda. To use these articles, we first translate them to English[2]. We test the efficacy of translation models by extracting text from a small sample of articles published in a given language and running the text through all available translation models on the Hugging Face open database[3]. We then assess whether the translations are sufficiently comprehensible that they produce classifiable results[4]. If they are not, we compare the performances with those of APIs available through the deep-translator package in Python[5] and choose one that yields the optimal sentence-to-sentence translations with sufficient human readability.

Identifying Locations

We perform location extraction from the text to determine where the event occurred. Articles in international and regional sources cover a wide range of countries, but local sources also frequently feature news coverage of important international events. In order to ensure that the events we report reflect activity in the country of interest, we use location extraction. We implement Named Entity Recognition (NER) to identify the locations of events using CLIFF,[6] which relies on GeoNames (for state/city/town names) and extracts the location information[7]. GeoNames is one of the most comprehensive and well-maintained sources of geographic data available, containing over 12 million unique location names across 250 countries (D'Ignazio et al., 2014). CLIFF API has detailed information of the locations detected, and we retrieve and convert the country codes of each location to assign the article to a specific location(s). If no country is found in the text, we assign the article to the country of origin of the news source for domestic sources and for international sources, the article is unused.

[1] See Appendix 2 for a full list of keywords and the reason for their inclusion.

[2] Although it is possible to classify events with the multi-language transformer models and the location extraction tools described below, training on data in one language facilitates the process of improving the model's overall accuracy.

[3] Hugging Face

[4] We assess translation efficiency by examining the translated outputs of articles to see if the translations are reasonable and interpretable to human readers.

[5] deep-translator

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

[7] GeoNames is a free, online database containing the names and location of populated places and geographic features all over the world.

Detecting Civic Space Events

Perhaps the most important part of event extraction is event classification. We fine-tune the RoBERTa model to detect reporting on civic space events using a double human-coded training dataset of hand-coded newspaper articles. The training data for the civic space event counts includes 6,475 (1,493 non-events and 4,982 events) articles covering our 20 event types. Importantly, our training data includes more than 4,000 articles originally published in another language and translated into English. The classifier produces a classification report that includes overall accuracy and a heatmap that was useful for identifying problem-areas in the model and event categories that required additional training data to improve accuracy.

In Table 1, we report out-of-sample model performance. A fine-tuned model with the default MLP settings produced classifications with overall out-of-sample accuracy close to 0.82 (civic space) on human-coded event data, with most misses coming from presence of multiple events in a single entry or from partially overlapping event categories. The precision, recall and F1 scores for each event category can be found below.

Table 1: RoBERTa Classifier Performance

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
-999	0.81	0.79	0.80

In order to improve overall classification accuracy, we allow some double classification of event types. In our validation, we discovered that certain events can overlap with other events of interest, primarily articles reporting on both corruption and subsequent arrests or legal actions. This prompted us to apply a keyword filter tailored to the specific dynamics

of corruption [8]. By searching for keywords in the main text and title, we can code articles with dual interpretations, enhancing our analysis by capturing both event scenarios and their associated narratives.

In addition to our event classification model, we have developed a model to categorize events into two distinct groups: civic-related and non-civic events. The classification of events into civic and non-civic categories is tailored to align with our specific criteria and the dimensions of civic space. Oftentimes news events receive extensive coverage despite being unrelated to civic space. The arrest or murder of a celebrity, for instance, can receive extensive news coverage even as it has little or no bearing on civic space more broadly.

The determination of civic relevance is established through the utilization of a specialized civic/non-civic classifier. This classifier is constructed using transfer learning techniques derived from our pre-trained RoBERTA model. In order to distinguish these types of non-civic events, we deploy a civic/non-civic classifier that takes articles already coded as events and identifies whether they are related to broader civic space. This classifier uses a subset of the broader training data with 2,938 human-coded articles and achieved an overall accuracy of .87. The classification model provides a 0/1 indication if an article qualifies as a civic space event. We apply this classifier to the subset of event categories for which we have identified high instances of non-civic relevant events with high levels of reporting[9]. We do retain these non-civic events in order to understand underlying news trends and they are reported in the data. To analyze this data, we sum these event counts across sources for each country by month.

Civic Salience

Discussed further in our technical validation section, we have found that many national news sources have inconsistent digital presence over our period of study. This can occur for many reasons. First, many sources produced less digital news further back in time. This might result from the gradual shift from paper to online news over the study period to the deletion or poor maintenance of web archives. Second, some sources seem to purge periods of their web archive for unknown reasons. Third and finally, some cases show discontinuous increases in the volume of news they report.

Although the reasons are idiosyncratic, this volatility in news volume is a challenge to consistently measuring civic space. Since changes in reporting on protests or legal changes can be the result of true shifts in events in the world or changes in the volume of availability of news. To address this, we normalize each month-event count by the total number of articles to derive a signal-to-noise ratio. This ratio tells us how frequently a given civic space category occurs relative to total scraped articles. We rely on this signal-to-noise ratio to analyze how civic space reporting changes relative to all reporting. This makes our data more robust to arbitrary increases or decreases in overall publication rates.

Counting the raw volume of news articles or the discrete number of events (via deduplication) for each event category provides useful information. However, we want capture the significance of these events to domestic politics. For example, the number of legal changes passed can't tell you much about their significance. However, the amount of attention does.

[8] See Appendix B for more details.

[9] The events we apply this to are: arrests, cooperation, corruption, defamation cases, legal actions, legal changes, purges, raids, threats, lethal violence and non-lethal violence

Furthermore, we need a method that works across multiple types of events (legal actions, but also arrests and protests). So we use media attention as a measure of importance

Event Detection

By measuring media attention, we capture the relative importance of different types of civic space events over time. However, this continuous measure can also be used to isolate months in which major events are occurring. To do so, we developed an ensemble algorithm to detect major 'shocks' to the share of reporting dedicated to each of our civic space event categories. We begin by applying winsorization to our data. This process reduces the impact of extreme values in our window by replacing high and low values with the next highest and lowest values, respectively.

Our peak detection method integrates a rolling window smoothing of 25 months with a grid search to fine-tune the multipliers for weighted means and the coefficients for weighted standard deviations, along with the parameters governing binning weights and decay functions. This process ensures the precise identification of civic event shocks through normalized count spikes. We apply two distinct approaches for our left hand side (LHS) and right hand side (RHS) windows. For our LHS window, we use decay weighting, whereby the weights applied to previous months decay non-linearly. For our RHS window, we use binning weights, where the weights applied to more distant months decay linearly. We make this distinction to allow for a pronounced decrease of recent events in order to capture rapid changes, but a gradual decrease with respect to future events in order to not overemphasize peaks when they reflect a shift in the underlying structure of the data.

By mitigating outlier impacts through winsorization, and applying context-sensitive decay and binning weights, our method adeptly captures the initial onset of event shocks. Post-optimization, statistical insights guide a neural network model, bolstered by definitive rules, to robustly and accurately detect major events. We have two considerations that we apply definitive rules for. First, sequential peak detection was enabled, allowing for the occurrence of two peaks in consecutive months. This accounts for situations in which an event snowballs – yielding multiple months of significant increases – which would otherwise only occur as one peak in the data. Second, we look to achieve roughly 15 percent peak identification. We do this to make sure that peaks are not too regular if data has high variance, but also to ensure that we do not miss out on meaningful shifts in low variance event types. This neural network approach allows us to position our model for seamless integration of future upgrades, and in particular allows for flexible input for future human-validated data without altering the underlying winsorization method. Figure 1 shows an example of the shock detection algorithm at work in Indonesia in the arrests category.

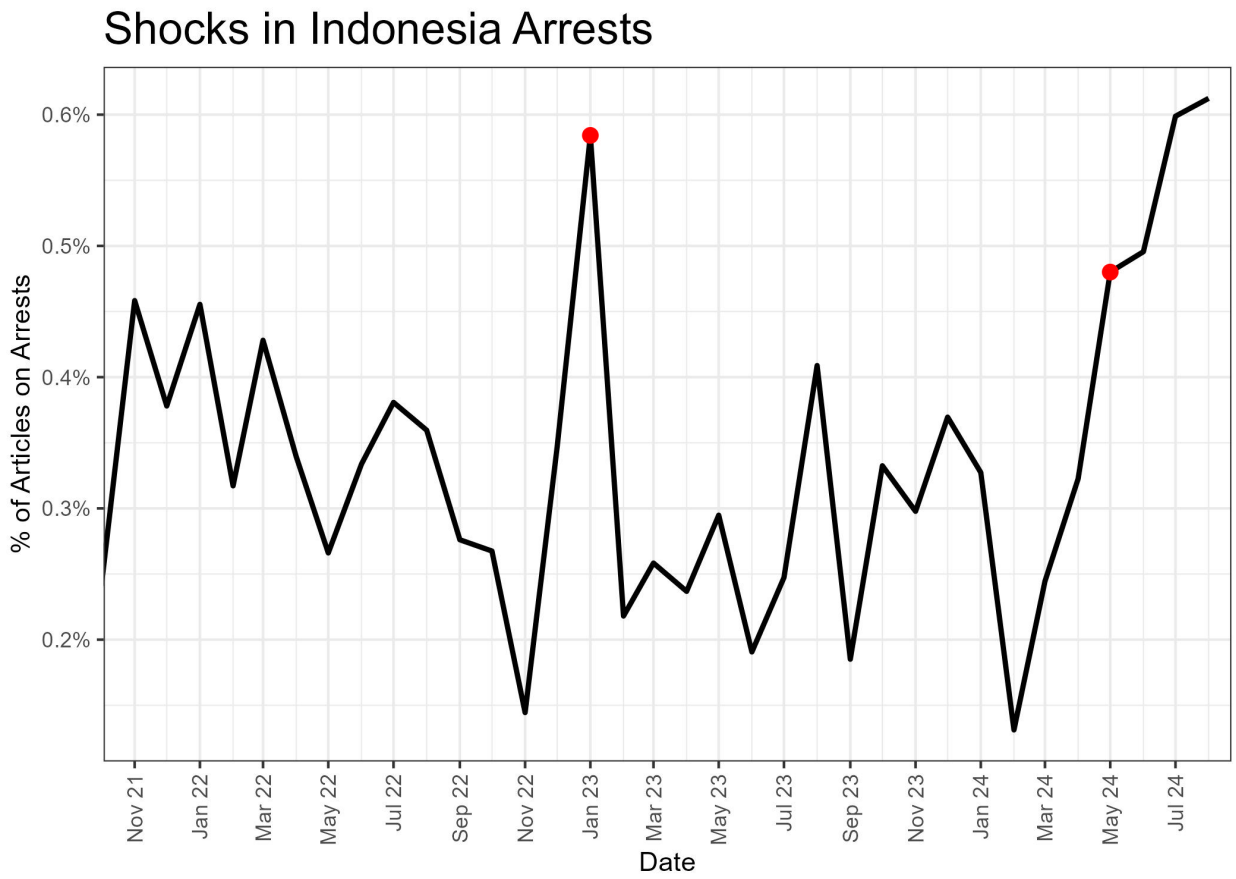

Figure 1: This figure shows the arrest category in Indonesian news. Our shock detection algorithm has detected two spikes in the past 3 years: in January 2023 and in May 2024.

Limitations

In this section, we discuss the limitations of the CS-ARC dataset. First, we discuss limitations that derive from HQMARC corpus on which CS-ARC is built. We then discuss limitations of CS-ARC itself.

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. We started data collection in 2019, suggesting that

Second, only news sources that have consistent and/or coherent internet infrastructure are included. We do this in order to ensure that movements in counts are a function of actual news rather than simply changes in the number of sources, but this comes at the cost of coverage, i.e. many sources in many countries have extremely poor web architecture. 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.

** Limitations of CS-ARC 1. Media attention is a proxy for importance 2. Normalization forces competition between measures, so really big events might make other significant events look less important than they actually are.

3 Data Validation

In this section we present results from several data validation exercises. We take two approaches to validation. First, we compare the the coverage of HQMARC with that of other big data media corpora. We show that HQMARC has consistent collects a significantly higher share of the articles published by the high-quality domestic news outlets in our sample. Second, we demonstrate the challenges of relying on prepackaged scraping tools. We show that the human validation efforts we apply solve issues with data reliability. Finally, we demonstrate the ability of CS-ARC to reliably detect major political events across a random sample of countries in the dataset.

Data Quality

In this section, we discuss what makes our data unique. First, data collection from national news sources requires careful human curation. We show that reliance on prepackaged scrapers like GDELT, Common Crawl, and the Internet Archive produce poorer coverage than our approach and introduce errors in the data. For example, we compared MLP's coverage of three Bandgladeshi news sources with that of GDELT and Internet Archive. We chose to test these alternatives on Bangladesh for 3 reasons. First, our sources for Bangladesh are relatively large; they publish a significant number of articles, which means that they are good candidates for regular scraping by large crawlers. Second, the site architecture for each of these sites is well-organized. In our scraping efforts, we have found that site architecture can vary widely in quality, with some sites having clean architecture that makes scraping accessible with little effort, while others require significant work on custom parsers to collect any data at all. While we have found that it is always necessary to create a custom parser to scrape a source, these custom parsers range from minimal to extensive. The ones we've created for our Bangladeshi sources are all relatively minimal, meaning large scale crawlers should have a fairly high level of success scraping these sources, even if there are shortcomings driven by the lack of a custom approach. Finally, while many of our sources are not in English, which introduces an additional set of scraping challenges, our biggest Bangladeshi sources are primarily published in English. This means that widely available, open-source scrapers can more easily automatically parse text, decreasing the effort required to scrape a source. In summary, Bangladesh represents a best-case scenario for large-scale data collection efforts like GDelt and the Internet Archive.

Despite these sources' relative ease of access, we find that there are significant disparities in performance between the MLP data that was collected and the data available through other collection efforts. MLP's coverage begins in 2013 for one source and in 2015 for the other two sources. By comparison, GDELT only has coverage from 2019 forward. Our process also generates much better coverage across all sources. For GDELT's best covered source, GDELT averages 2,100 articles per month compared to 2,500 per month from MLP.[10]

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

Internet Archive performs even worse. The disorganized nature of Internet Archive means that it took nearly two weeks to collect URLs from a single source from 2019-2023, and the results included numerous irrelevant, broken, and duplicate links. Although Internet Archive delivered a large number of URLs, less than half were usable.

To further demonstrate the importance of human curation, Figure 2 shows two examples of coverage changes that can occur in web-scraped media. On the left, Figure 2 shows a spike in ghanaweb.com, a large media outlet in Ghana. The dramatic increase in publications in 2020 was related to a grant the outlet received from Google that allowed it to massively expand its coverage. On the right, Figure 2 shows a spike in lusakatimes.com in Zambia. This spike was driven by one article that the website mistakenly uploaded over 1.5 million times (each with a unique URL, making this error difficult to detect). Only careful human curation can distinguish between genuine (Ghana) and artificial (Zambia) changes in publication volume, and effectively guard against such risks to data quality.

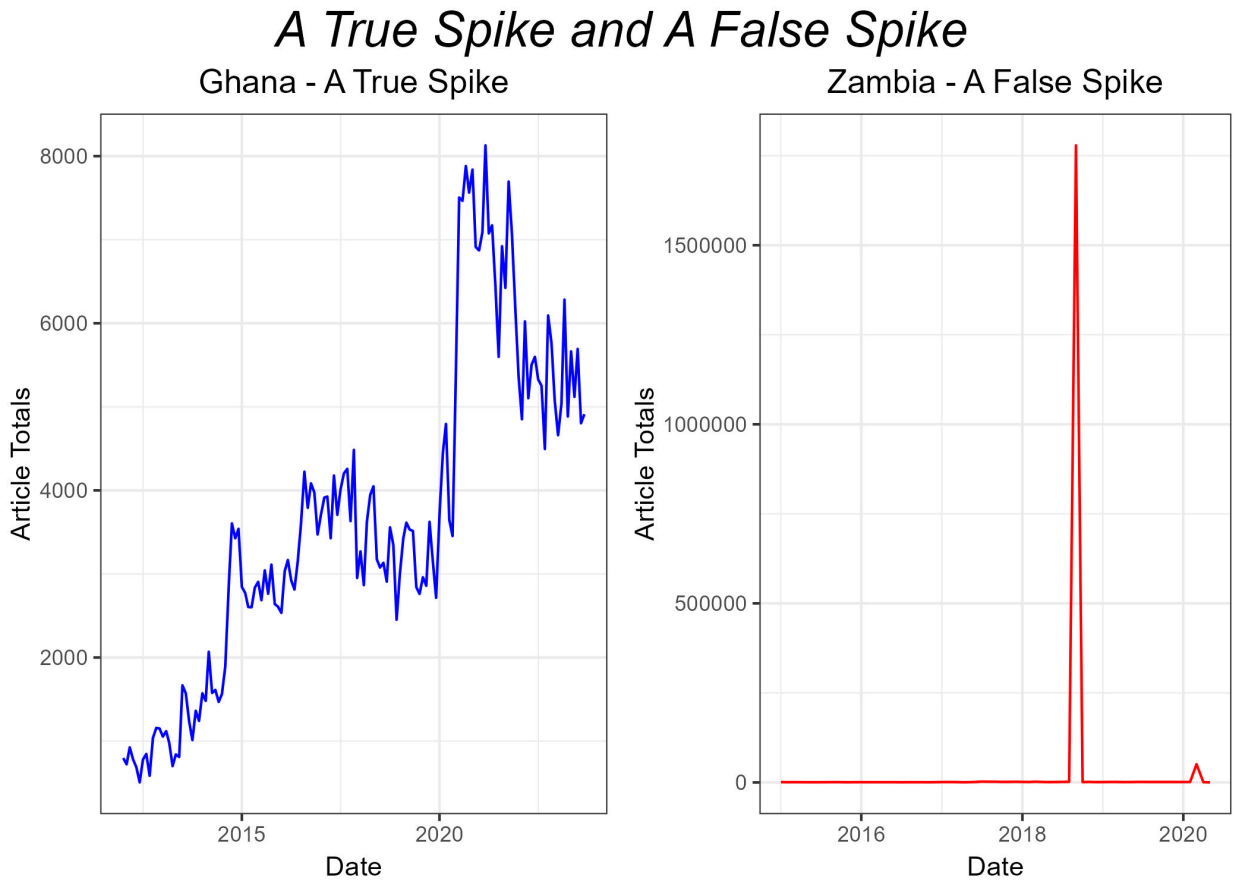

Figure 2: Changes in the volume of articles across two sources. In Ghana, the sudden in shift in volume was driven by a grant that reflected a real change in the total articles being published. In Zambia, the sudden shift in volume was driven by a single article duplicated hundreds of thousands of times.

Spike Validation

Finally, we validate the ability of our measure of media attention to detect important political events. Using the event detection algorithms discussed in our data section, we consult

the articles in our corpus to assess whether jumps in our measure of media attention correspond with real events on the ground. First, we randomly selected five countries from our database: Kosovo, Morocco, Angola, Mauritania, and Ukraine. These countries cover three regions (North Africa, Sub-Saharan Africa, and Eastern Europe) and four languages (Serbian, Albanian, French, and Ukrainian).

For the most recent three-month period of data available at the time of the exercise (April – June 2024), we use our event detection algorithms to identify country-months with anomalously high media attention across the 20 civic space event categories. In total, our method identified 40 events out of a possible 300 ($3 months x 5 countries x 20 events$). For each country-month for which an event was detected, we then pulled all articles from our corpus that were published by a domestic outlet based in the corresponding country and reporting on the corresponding event.

Due to the large volume of articles for some country-months[11], we accessed GPT 4o through the OpenAI API and asked GPT briefly describe the events being reported-on in the articles (up to five). After generating summaries of the articles in our corpus corresponding to each of the 40 country-months we detected, a research assistant as asked to read the original articles (or a sample of 50 articles for country-months with more than 50 articles) and the GPT summary of the five most important events and code each event across four dimensions:

- Of the events described in the GPT summary, how many are accurately summarized by GPT?
- Of the events described in the GPT summary, how many are accurately classified by GPT as one of the five most important events reported-on in the original articles?
- Of the events described in the GPT summary, how many happened in the assigned country?
- Of the events described in the GPT summary, how many corresponded with the assigned event category?

For 34 of 40 country-months (85%), all of the top-five events summarized by GPT occurred in the correct country, and for 38 of 40 country-months (95%), all of the top-five events summarized by GPT were coded as the correct event category. Furthermore, of the 200 total events summarized by GPT, 181 were coded as the correct event category. This provides extremely persuasive evidence that our measure of media attention is able to accurately identify months in which major political events. The prompt used to generate summaries from GPT, the instructions for human-coding, and the human-coded results are available in the 'gpt-validation' subfolder of the git repository associated with this paper.

Comparing International Media to Local Media

In this section, we provide a clear use case of our data by comparing national and international news coverage of civic space events in 62 countries, highlighting the shortcomings of

[11] Across the events that we detected, the number of articles published by local sources ranged from one to 1,002. For rare event categories, such as Defamation Case, a single article can sometimes constitute an event.

data sources that rely solely on international coverage. The overarching challenge to analyzing civic space is that the most common and careful measures, such as V-DEM, report on an annual basis. While those measures are useful for tracking overarching regime dynamics, they provide little information on the day-to-day political struggles that characterize the current era of democratic backsliding. The combination of big data and machine learning offer a solution. However, we show that any such attempt must overcome two challenges. First, international media does not provide an accurate picture of key civic space events in many countries. International sources provide sparse and inconsistent coverage of even highly salient events in many countries. Thus, big data projects and policy makers relying only on international news coverage garner a biased picture of civic space events as they unfold on the ground. Second, constructing a corpus of domestic news from countries around the world requires a great deal of human curation; improper scraping can introduce enormous error into data. Many domestic media outlets exhibit sudden shifts in publication volume and poor website architecture, requiring careful human monitoring and customized scraping and parsing to consistently capture all published articles. Distinguishing true changes to publishing rates from technical challenges in scraping and parsing articles precludes exclusive reliance on prepackaged web scrapers. As a result, our approach provides a much richer, more detailed picture of civic space in each country.

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

In Figure 3, the blue bars indicate the share of our articles in the MLP corpus that are scraped from national sources. Clearly, the vast majority of our data comes from national (rather than international) sources. The dots in the figure show the correlation between the incidence of reporting on each type of civic space event. If international and national sources are reporting on the same events, albeit in different volumes, these correlations would be high. Yet the mean correlation across event types is only .23, and is only .51 for the most similarly reported category, election activity. This suggests fundamental differences in the type of events covered by domestic and international sources.

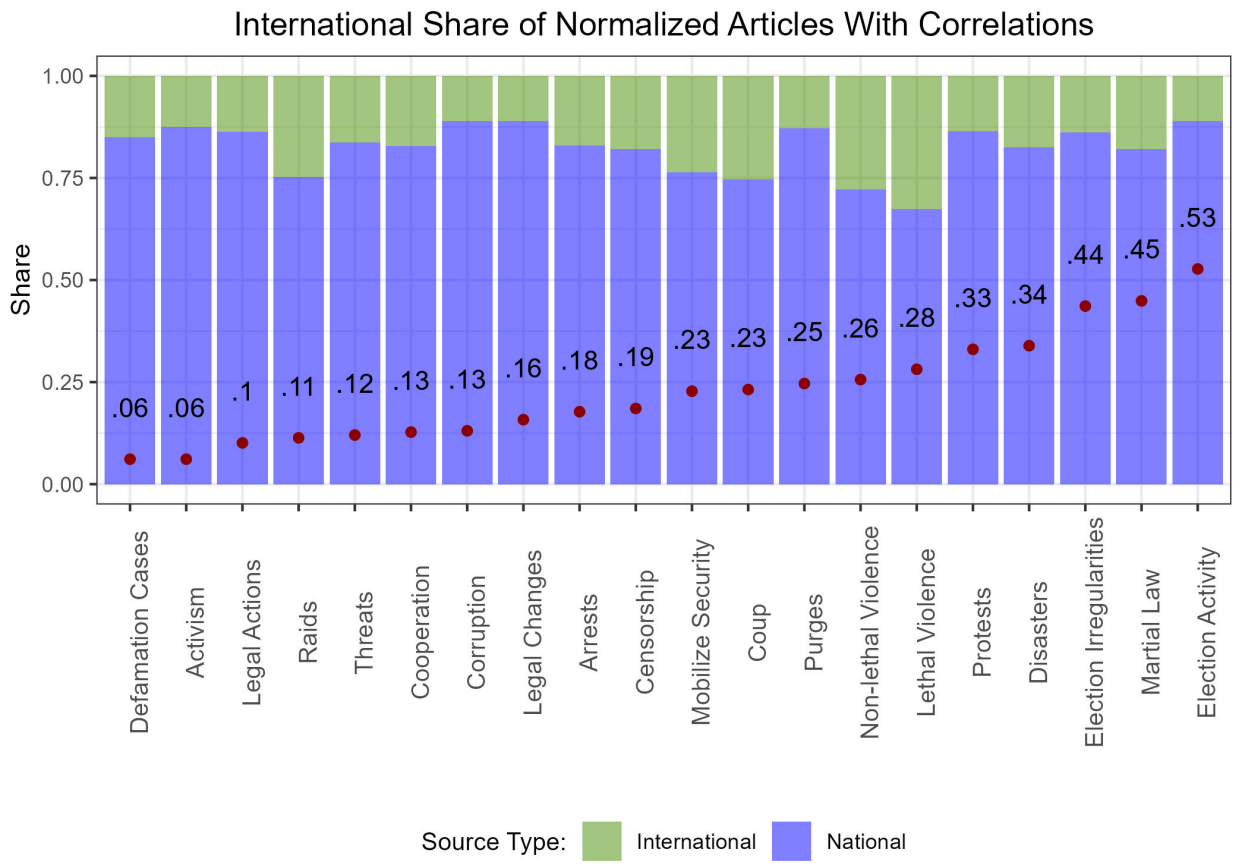

Figure 3: This figure shows the share of reporting on event types from international and national sources, and the correlation for each event type across all countries.

We also examine the rate at which different event types are covered by national and international news. Figure 4 shows event types sorted by the extent of international coverage, with national coverage stacked on top. We see that there are big differences in the types of events that are covered in international media compared to domestic media. International media is far more likely to cover acts of violence, whereas domestic media is far more likely to cover legal actions, election activity, corruption, protests, and cooperation. To the extent these are key features of civic space dynamics, reliance on international news will produce a heavily biased view of civic space. This is especially crucial in constructing a high-frequency, holistic view of civic space since oftentimes violence is preceded by events like corruption, legal actions, and protests. A system which relies only on widely available international coverage risks providing little more forewarning than more low-frequency measures such as the expert surveys conducted by organizations like V-Dem.

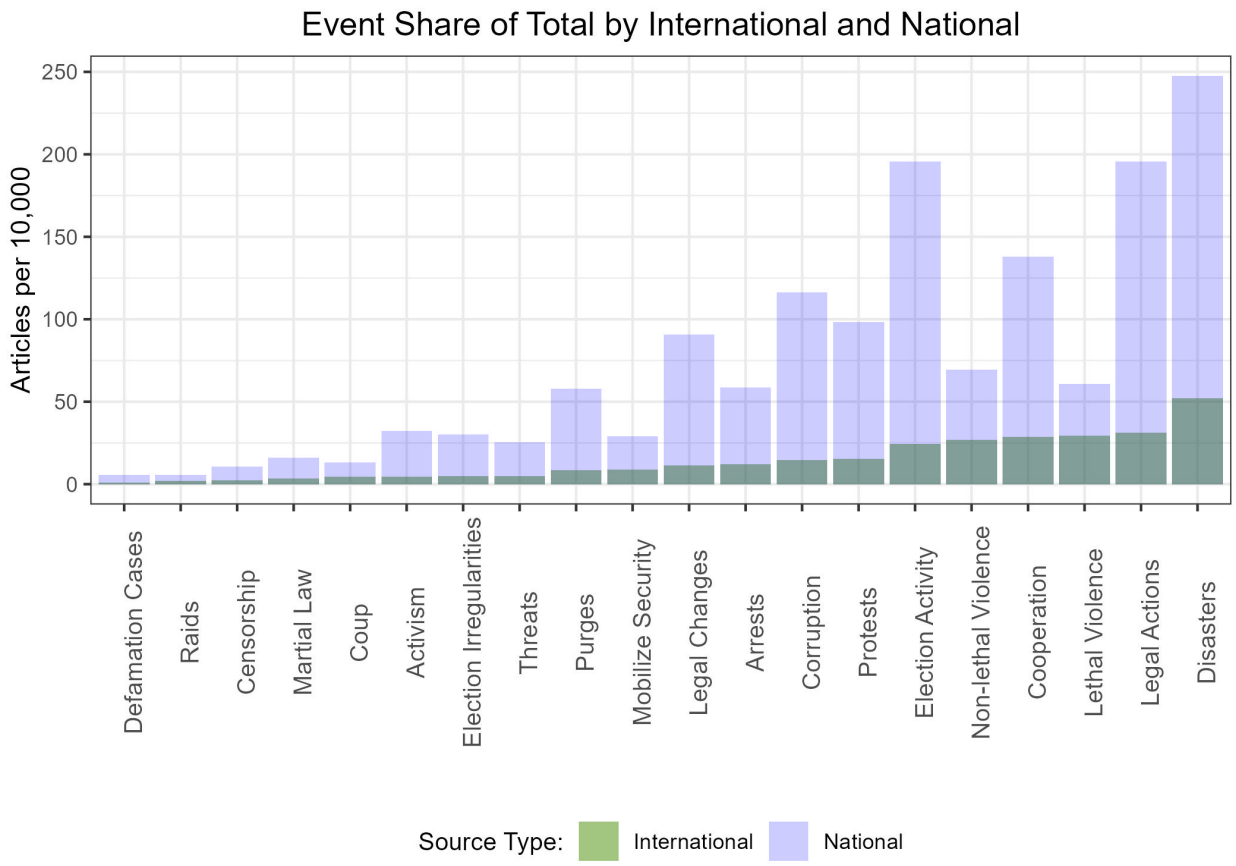

Figure 4: This figure shows the share of reporting of an event type from international and domestic coverage. It also displays the average correlation for each event type across all countries.

Finally, we examine whether these trends are driven by the volume of international coverage a country receives. It is undeniably the case that international media covers some countries more extensively than others, and it could be that international coverage of civic space is more complete in countries with more coverage. Figure 5 shows the volume of international articles published about a country (x-axis) against the correlation in civic space coverage of events between international and national sources. The figure shows that while there is a slightly positive relationship between the volume of coverage and the correlation between domestic and international reporting, the correlation is low across all countries. For instance, in Turkey and India, two countries that receive a meaningful amount of international coverage, the correlation between national and international coverage is below 0.5. There is almost no coverage of Timor Leste, and what little coverage exists has almost no correlation with what the domestic press is reporting.

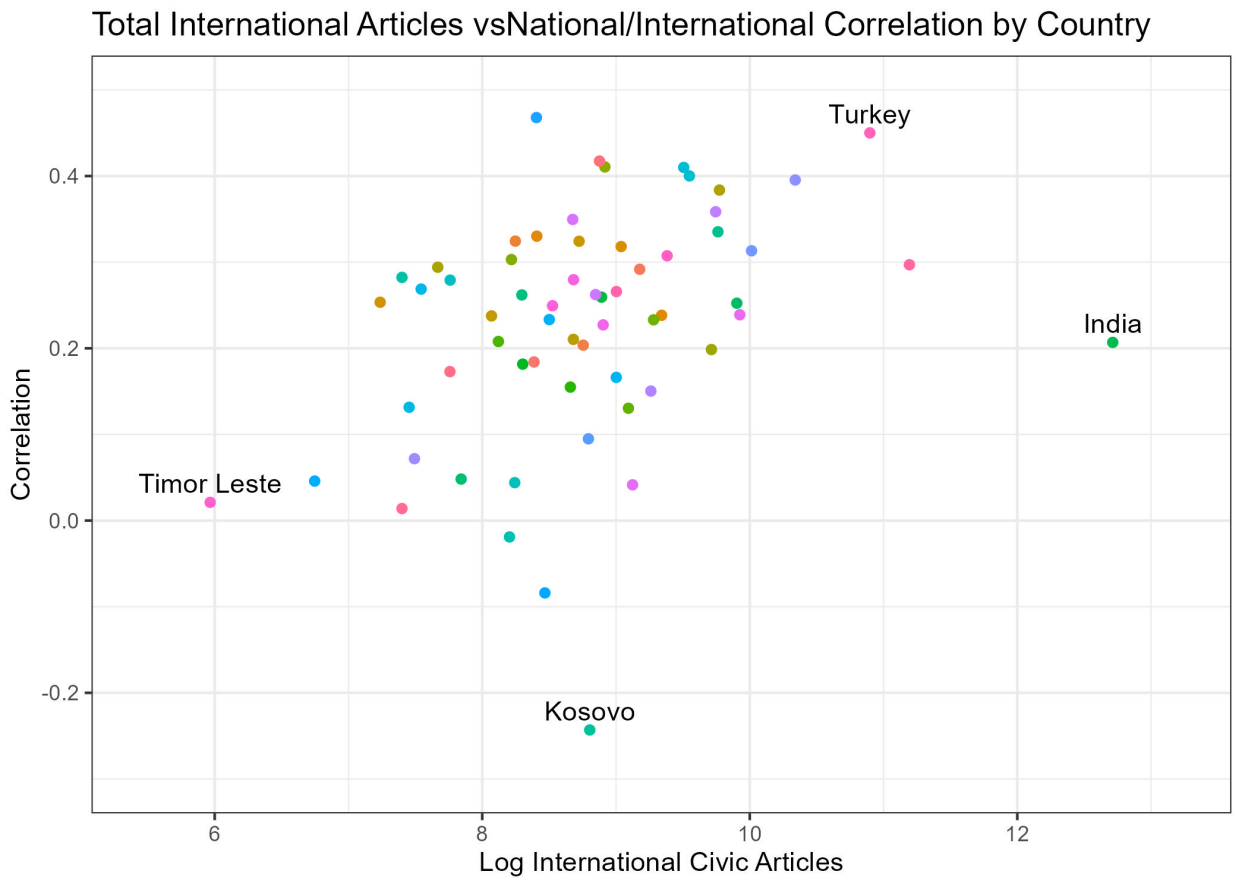

Figure 5: This figure shows the share of reporting of an event type from international and domestic coverage. It also displays the average correlation for each event type across all countries.

In Figure 6, we show a case study of an ongoing corruption crackdown in Indonesia. These articles cover several widespread corruption scandals involving commodities trade, public services and state-owned enterprises. This is particularly noteworthy, as the outgoing president was originally elected on an anti-corruption platform but has seen several ministers taken down for corruption himself (Meilasari-Sugiana et al., 2024). We chose this event because it has received significant national coverage- over 1,000 articles published in June 2024 alone related to corruption- but zero article coverage in any of our international or regional sources. Figure 6 shows coverage of corruption events and arrests. The blue line shows national news coverage, showing simultaneous spikes in both corruption and arrest coverage in May 2024 and continuing into June 2024. Meanwhile, the red and green lines are regional and international coverage respectively, neither of which cover these events in any way. This lack of coverage is despite the fact that these corruption events are not petty crimes, but embroil large national corporations and public officials. This case study highlights that not only does international data fail to give a complete picture of the salience of events on the ground, it can occasionally miss events entirely. During our preparation of this report, we found 59 events across 23 number of countries that we identified as meaningful spikes in those event categories that received zero articles of coverage in either regional or international sources.

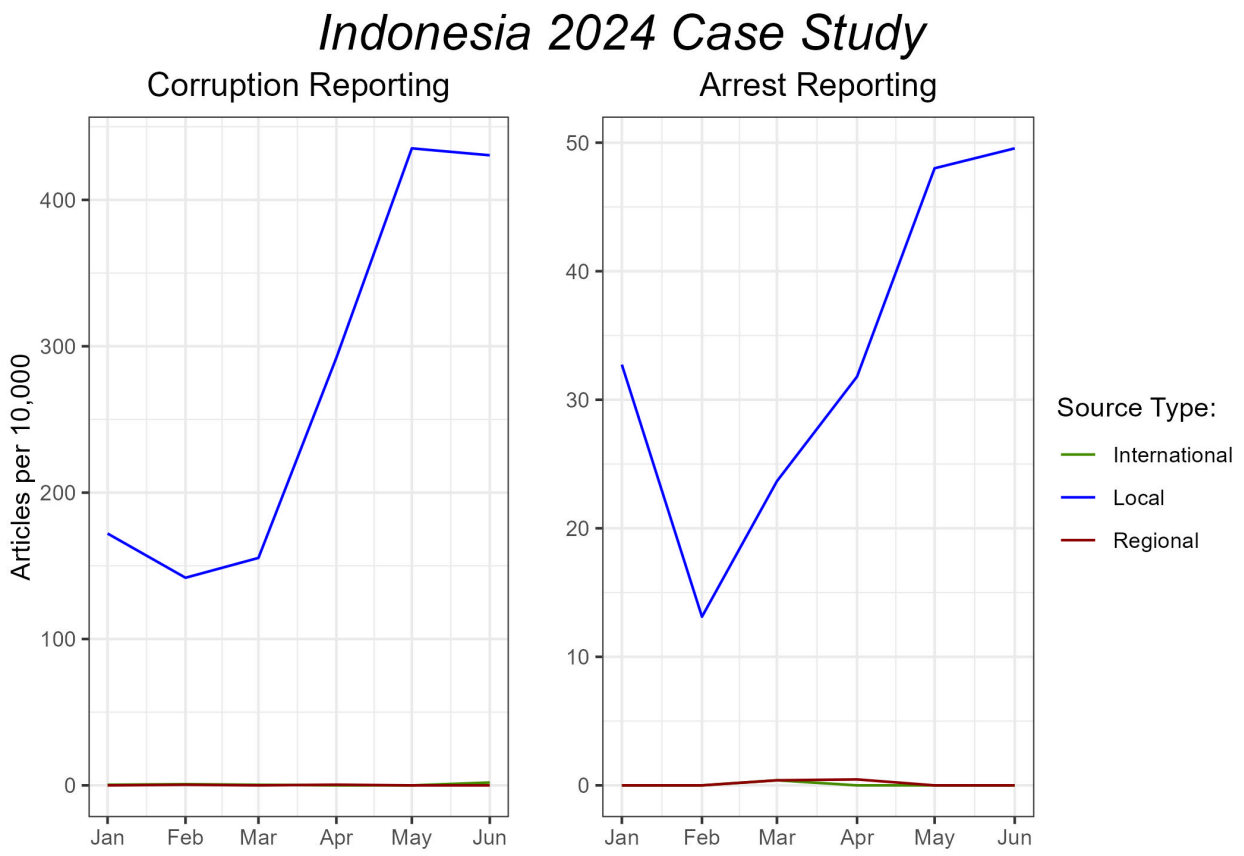

Figure 6: This figure shows the articles per 10,000 from international and domestic coverage related to corruption and arrests in Indonesia.

Conclusion

In this paper we have introduced an exciting new dataset – HQMARC. We have provided careful discussion of the methods used to create the data; custom scraping, translation, and event classification. We have demonstrated that reliance on first-generation transformers can yield high-performing classification of local media in a cost-effective and time-efficient manner. We have also discussed our efforts to validate both the underlying data itself and the analysis built upon the data. We have shown that careful, human supervised collection of data like this is necessary to ensure overall quality. We have also demonstrated that the analysis of this data, including our surge detection algorithm, reflects real, meaningful changes in local media attention. Finally, we have used our data to show that reliance on international sources has serious potential pitfalls. We have demonstrated that there is frequently little correlation between international and domestic reporting and that international coverage frequently ignores meaningful local events entirely.

This work has several important implications. First, it provides researchers with a reliable dataset to complement other, traditional measures of domestic civic activity. Second, making US foreign policy decision-making more data-driven has tremendous potential to improve outcomes. Using media data to track changing political conditions in strategically important countries is critical to this goal. However, this paper shows that policymakers should be

cautious when using data that is overly reliant on international media or collected using automated tools. HQMARC combines direct, human-supervised data collect to ensure data quality with new tools, such as large language models and machine translation, to leverage much richer coverage from domestic outlets based in developing countries. Information sourced from media is one of the primary sources of information for the US government. Future investments in media data production should incorporate these insights into their design.

References

[1] Claudio M. V. de Andrade et al. *A Strategy to Combine 1stGen Transformers and Open LLMs for Automatic Text Classification*. 2024. arXiv: 2408.09629 [cs.CL]. URL: https://arxiv.org/abs/2408.09629.

[2] Florian Arendt. "The Media and Democratization: A Long-Term Macro-Level Perspective on the Role of the Press During a Democratic Transition". In: *Political Communication* 41.1 (2024), pp. 26–44.

[3] Matthew A Baum and Yuri M Zhukov. "Filtering revolution: Reporting bias in international newspaper coverage of the Libyan civil war". In: *Journal of Peace Research* 52.3 (2015), pp. 384–400. DOI: 10.1177/0022343314554791. URL: https://doi.org/10.1177/0022343314554791.

[4] Timothy Besley and Robin Burgess. "The political economy of government responsiveness: Theory and evidence from India". In: *The quarterly journal of economics* 117.4 (2002), pp. 1415–1451.

[5] Vanessa A Boese-Schlosser et al. "Autocratization changing nature?" In: *Democracy Report* (2022).

[6] Lauren Bridges. *The Impact of Declining Trust in the Media*. Accessed: 2024-10-04. 2019. URL: https://www.ipsos.com/en-uk/impact-declining-trust-media.

[7] C. Brimicombe. "Is there a climate change reporting bias? A case study of English-language news articles, 2017–2022". In: *Geoscience Communication* 5.3 (2022), pp. 281–287. DOI: 10.5194/gc-5-281-2022. URL: https://gc.copernicus.org/articles/5/281/2022/.

[8] Catherine D'Ignazio et al. "CLIFF-CLAVIN: Determining Geographic Focus for News Articles". In: *NewsKDD: Data Science for News Publishing, at KDD 2014*. 2014. URL: https://hdl.handle.net/1721.1/123451.

[9] Stergios Fotopoulos. "Traditional media versus new media: between trust and use". In: *European View* 22.2 (2023), pp. 277–286.

[10] Reuters Institute for the Study of Journalism. *Digital News Report: India Supplementary Report*. Accessed: 2024-09-27. Reuters Institute, University of Oxford, 2019. URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-03/India_DNR_FINAL.pdf.

[11] Sangwon Lee, Trevor Diehl, and Sebastián Valenzuela. "Rethinking the virtuous circle hypothesis on social media: Subjective versus objective knowledge and political participation". In: *Human Communication Research* 48.1 (2022), pp. 57–87.

[12] Anna Lührmann and Staffan I Lindberg. "A third wave of autocratization is here: what is new about it?" In: *Democratization* (2019), pp. 1–19.

[13] Astrid Meilasari-Sugiana, Endro Gunardi, and Siwage Dharma Negara. "Corruption Eradication in Indonesia: One Step Forward, Two Steps Back". In: *ISEAS Perspective* 2024.42 (2024). URL: https://www.iseas.edu.sg/articles-commentaries/iseas-perspective/2024-42-corruption-eradication-in-indonesia-one-step-forward-two-steps-back-by-astrid-meilasari-sugiana-gunardi-endro-siwage-dharma-negara/.

[14] P. Quartey et al. *Radio and Social Media Assessment Report*. Tech. rep. Accra, Ghana: USAID Ghana MEL Platform, 2023.

[15] Svenja Schäfer and Christian Schemer. "Informed participation? An investigation of the relationship between exposure to different news channels and participation mediated through actual and perceived knowledge". In: *Frontiers in Psychology* 14 (2024), p. 1251379.

[16] David Waldner and Ellen Lust. "Unwelcome change: Coming to terms with democratic backsliding". In: *Annual Review of Political Science* 21 (2018), pp. 93–113.