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

Supplemental Materials

Table of contents

Appendix A: Data in HQMARC

The High-Quality Media from Aid Receiving Countries (*HQMARC*) corpus contains articles scraped from 354 prominent domestic media outlets based across our sample of 66 countries and publishing in 36 languages. We supplement these domestic outlets with content scraped from 12 regional and 15 international outlets. In sharp contrast to other sources of event data, more than 95% of the articles in *HQMARC* are scraped from domestic media outlets.

Using a customized harvesting workflow for each source domain, we achieve comprehensive or near-comprehensive capture of all relevant articles published by each source from 2012 through 2024 (or the earliest/latest publication date). Most large media data corpora run a blind, domain-wide crawler that follows every hyperlink it can find. Instead, we build a small, customized harvesting workflow for each source domain. The first step is URL discovery: we look for a structured entry-point, usually a public sitemap. If the sitemap is incomplete or missing, we switch to site-specific search strategies (pagination through section indexes, keyword queries, RSS feeds, etc.). Only when those programmatic methods fail do we fall back on heavier tools such as Selenium, and even then the goal is to retrieve clean article links, not to crawl arbitrary pages. Once a domain's URL list is assembled, we scrape those URLs directly. The scraper for each source knows exactly where the title, date, main text and other structured elements live in that site's HTML, so the resulting data land in a uniform, column-ready format. If a field doesn't parse correctly on the first pass (e.g., a date embedded in a non-standard tag, the parser not able to extract the correct text), we apply a post-scrape fixer for that column; we do not launch another crawl. Because the URLs are pre-vetted and the parsers are domain-specific, the pipeline yields cleaner, more consistent articles than a generic crawler would.

Section 1: Domestic and Regional Sources in HQMARC

We divide our countries into five geographic regions. Each region has dedicated regional sources that cover all countries within that geographic region. Figure 1 shows the number of domestic and regional sources for each country. Figure 2 shows the number of *active* domestic sources for each country-month. Not all sources yield data for every month in our sample. Some sources only start publishing articles after the beginning of our coverage period in January 2012 or cease publishing before the end of our coverage period in December 2024. Other sources hhave incomplete archives or deploy blocking, which restricts our ability to capture articles at the beginning or end of the time series, respectively. We provide data on the months for which articles are actively scraped from each source in the Source Coverage section in our github repository.

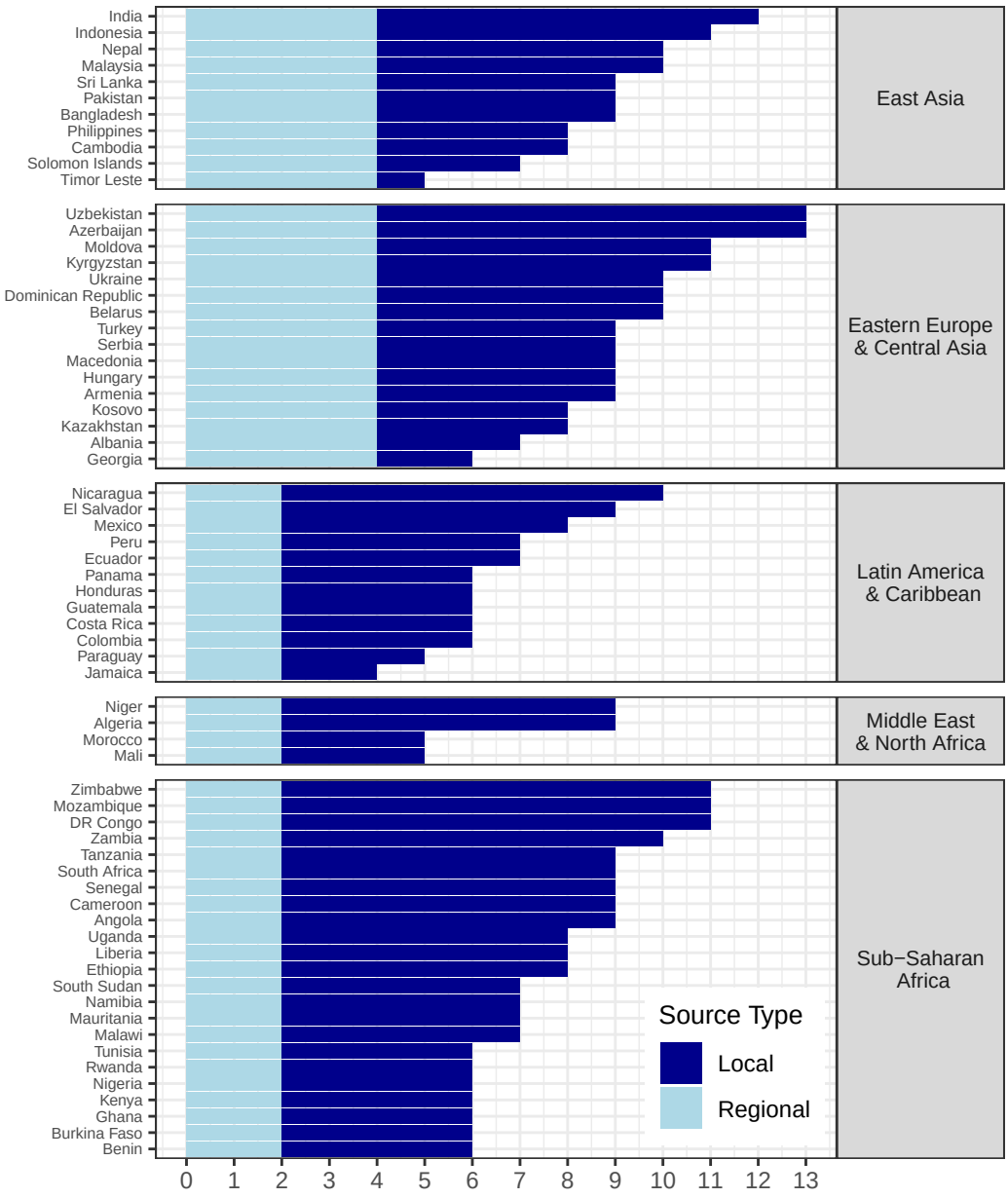


Figure 1: Number of domestic and regional media outlets in the HQMARC corpus by country

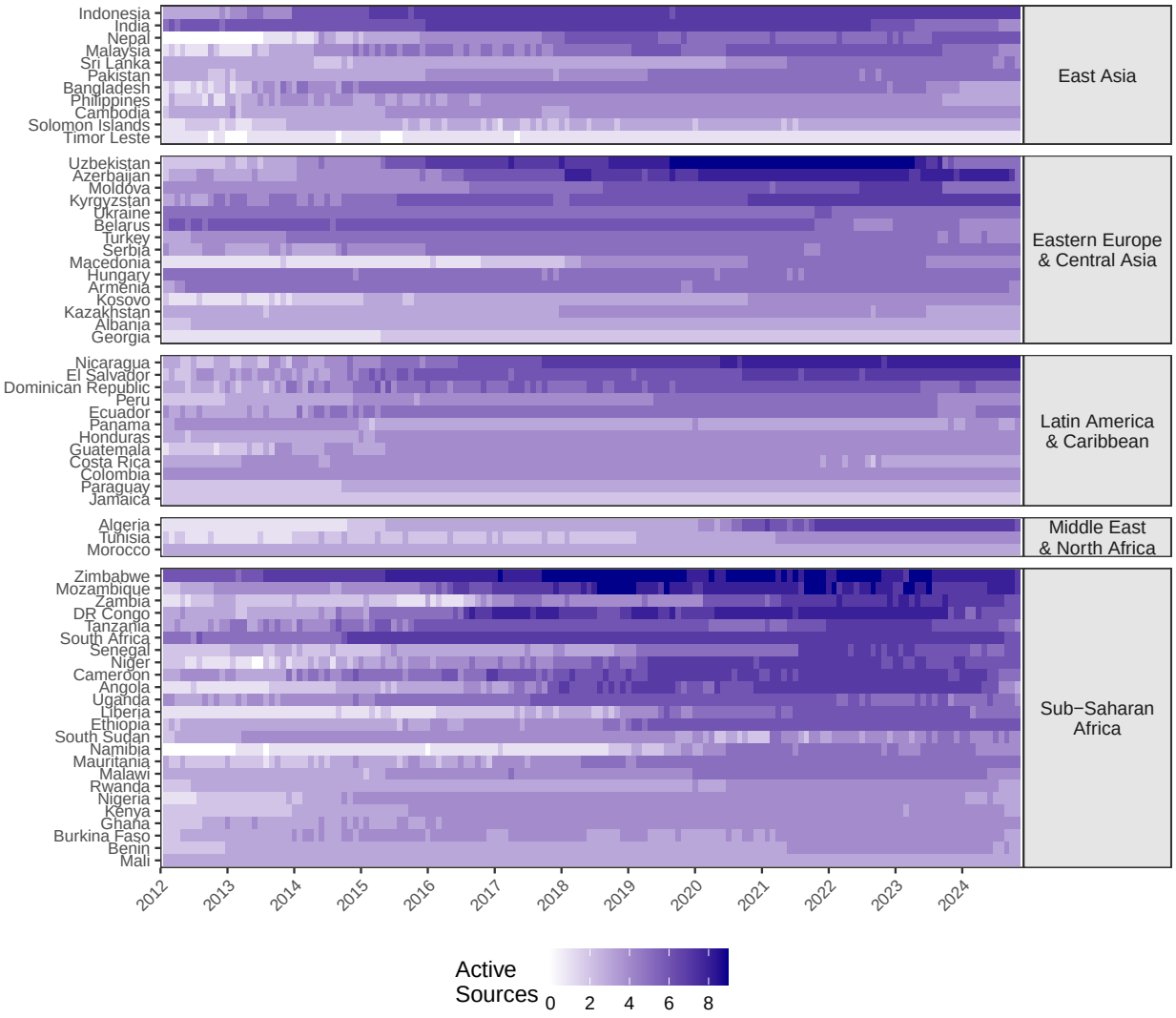


Figure 2: Number of Active Domestic Sources by Country-Month

Section 2: Languages in HQMARC

The *HQMARC* corpus includes domestic media outlets publishing in 36 languages. These languages include Afrikaans, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Belarusian, Bengali, Chinese, English, French, Georgian, Hindi, Hungarian, Indonesian, Kazakh, Central Khmer, Kinyarwanda, Kongo, Macedonian, Malay, Nepali, Portuguese, Romanian, Russian, Sinhala, Spanish, Serbian, Swahili, Tagalog, Turkish, Ukrainian, Urdu, Uzbek, Zulu. Below, we provide the languages associated with media outlets in each country using the language ISO codes.

Table 1: Eastern Europe & Central Asia

Country	Languages
Albania	sq, en
Armenia	hy, ru
Azerbaijan	en, az, ru
Belarus	be, ru
Georgia	ka, en
Hungary	hu
Kazakhstan	kk, ru, en
Kosovo	sq
Kyrgyzstan	en, ru, kg
Macedonia	sq, en, mk
Moldova	ro, ru, en
Serbia	sr
Turkey	tr
Ukraine	uk, ru, en
Uzbekistan	en, ru, uz

Table 2: Middle East & North Africa

Country	Languages
Algeria	fr, ar
Morocco	ar, fr
Tunisia	fr, ar

Table 3: Latin America & Caribbean

Country	Languages
Colombia	es
Costa Rica	es, en
Dominican Republic	es
Ecuador	es
El Salvador	es, en
Guatemala	es
Honduras	es, en
Jamaica	en
Nicaragua	es
Panama	es
Paraguay	es
Peru	es

Table 4: East Asia

Country	Languages
Bangladesh	bn
Cambodia	km, en
India	en, hi
Indonesia	en, id
Malaysia	en, ms, zh
Nepal	ne, en
Pakistan	ur, en
Philippines	en, tl
Solomon Islands	en
Sri Lanka	en, si
Timor Leste	tet, pt

Table 5: Sub-Saharan Africa

Country	Languages
Angola	pt
Benin	fr
Burkina Faso	fr
Cameroon	fr
DR Congo	fr
Ethiopia	en, am
Ghana	en
Kenya	en
Liberia	en
Malawi	en
Mali	fr
Mauritania	fr, ar
Mozambique	pt, en
Namibia	en
Niger	fr
Nigeria	en
Rwanda	en, fr, rw
Senegal	fr
South Africa	af, en, zu
South Sudan	en, ar
Tanzania	en, sw
Uganda	en
Zambia	en
Zimbabwe	en

Section 3: Digital News Sources

In this section, we list the individual sources included in *HQMARC*. First, we list our international sources and regional sources. Second, we list the sources associated with each country, organized by region.

International Sources

- aljazeera.com
- bbc.com
- csmonitor.com
- france24.com
- nytimes.com
- reuters.com
- scmp.com
- theguardian.com
- themoscowtimes.com
- washingtonpost.com
- wsj.com
- lemonde.fr
- liberation.fr
- elpais.com
- lefigaro.fr

Eastern Europe & Central Asia Regional Sources

- balkaninsight.com
- euronews.com
- iwpr.net
- neweasterneurope.eu

Middle East & North Africa Regional Sources

- africanews.com
- theeastafrican.co.ke

Latin America & Caribbean Regional Sources

- cnnespanol.cnn.com
- telemundo.com

East Asia Regional Sources

- asia.nikkei.com
- asiatimes.com
- indiatimes.com
- timesofindia.indiatimes.com

Sub-Saharan Africa Regional Sources

- africanews.com
- theeastafrican.co.ke

Eastern Europe & Central Asia Local Sources

Albania :

- gazetatema.net
- panorama.com.al
- telegraf.al

Armenia :

- azatutyun.am
- aravot.am
- 168.am
- 1in.am
- golosarmenii.am

Azerbaijan :

- azeritimes.com
- azadliq.info
- abzas.org
- turan.az
- zerkalo.az
- mikroskopmedia.com
- xalqcebhesi.az
- musavat.com
- ru.echo.az

Belarus :

- nashaniva.by
- novychas.by
- nv-online.info
- belgazeta.by
- zviazda.by
- sb.by

Georgia :

- ambebi.ge
- georgiatoday.ge

Hungary :

- index.hu
- 24.hu
- 168.hu
- hvg.hu
- demokrata.hu

Kazakhstan :

- caravan.kz
- diapazon.kz
- kaztag.kz
- rus.azattyq.org

Kosovo :

- kosova-sot.info
- balkaninsight.com
- prishtinainsight.com
- botasot.info

Kyrgyzstan :

- akipress.com
- 24.kg
- kloop.kg
- super.kg
- vb.kg
- kaktus.kg
- kaktus.media

Macedonia :

- koha.mk
- slobodenpecat.mk
- makfax.com.mk
- skopjediem.com
- novamakedonija.com.mk

Moldova :

- timpul.md
- tribuna.md
- unimedia.info
- voceabasarabiei.md

- publika.md
- ipn.md
- zdg.md

Serbia :

- rs.n1info.com
- juznevesti.com
- insajder.net
- danas.rs
- balkaninsight.com

Turkey :

- diken.com.tr
- t24.com.tr
- sozcu.com.tr
- posta.com.tr
- sabah.com.tr

Ukraine :

- delo.ua
- interfax.com.ua
- kp.ua
- pravda.com.ua
- kyivpost.com
- kyivindependent.com

Uzbekistan :

- fergana.ru
- kun.uz
- gazeta.uz
- podrobno.uz
- batafsil.uz
- sof.uz
- anhor.uz
- asiaterra.info
- daryo.uz

Middle East & North Africa Local Sources

Algeria :

- twala.info
- 24hdz.com
- echoroukonline.com
- elkhabar.com
- el-massa.com

- elwatan-dz.com
- echaab.dz

Morocco :

- leconomiste.com
- lematin.ma
- assabah.ma

Tunisia :

- assarih.com
- babnet.net
- jomhouria.com
- lapresse.tn

Latin America & Caribbean Local Sources

Colombia :

- elcolombiano.com
- elespectador.com
- elheraldo.co
- eltiempo.com

Costa Rica :

- larepublica.net
- news.co.cr
- ticotimes.net
- diarioextra.com

Dominican Republic :

- diariolibre.com
- listindiario.com
- elnacional.com.do
- hoy.com.do
- elcaribe.com.do
- elviajero.com.do

Ecuador :

- elcomercio.com
- eldiario.ec
- elnorte.ec
- eluniverso.com
- metroecuador.com.ec

El Salvador :

- laprensagrafica.com
- elfaro.net
- elsalvador.com
- diario.elmundo.sv
- diarioelsalvador.com
- revistafactum.com
- gatoencerrado.news

Guatemala :

- prensalibre.com
- republica.gt
- lahora.gt
- soy502.com

Honduras :

- elheraldo.hn
- laprensa.hn
- proceso.hn
- tiempo.hn

Jamaica :

- jamaica-gleaner.com
- jamaicaobserver.com

Nicaragua :

- confidencial.com.ni
- laprensani.com
- nuevaya.com.ni
- articulo66.com
- laverdadnica.com
- ondalocalni.com
- canal2tv.com
- lajornadanet.com

Panama :

- elsiglo.com.pa
- critica.com.pa
- panamaamerica.com.pa
- newsroompanama.com

Paraguay :

- abc.com.py
- lanacion.com.py
- ultimahora.com

Peru :

- elcomercio.pe
- gestion.pe
- larepublica.pe
- ojo-publico.com
- idl-reporteros.pe

East Asia Local Sources

Bangladesh :

- prothomalo.com
- bd-pratidin.com
- kalerkantho.com
- jugantor.com
- dailyjanakantha.com

Cambodia :

- kohsantepheapdaily.com.kh
- moneaksekar.com
- phnompenhpost.com
- cambodiadaily.com

India :

- amarujala.com
- indianexpress.com
- thehindu.com
- hindustantimes.com
- deccanherald.com
- firstpost.com
- indiatimes.com
- timesofindia.indiatimes.com

Indonesia :

- thejakartapost.com
- jawapos.com
- kompas.com
- mediaindonesia.com
- sindonews.com
- beritasatu.com
- hariansib.com

Malaysia :

- malaymail.com
- nst.com.my

- thestar.com.my
- utusan.com.my
- thesun.my
- malaysiakini.com

Nepal :

- onlinekhabar.com
- english.onlinekhabar.com
- en.setopati.com
- thehimalayantimes.com
- kathmandupost.com
- nepalitimes.com

Pakistan :

- jang.com.pk
- nation.com.pk
- dailytimes.com.pk
- pakobserver.net
- tribune.com.pk

Philippines :

- mb.com.ph
- manilastandard.net
- inquirer.net
- manilatimes.net

Solomon Islands :

- solomonstarnews.com
- solomontimes.com
- sibconline.com.sb

Sri Lanka :

- dailymirror.lk
- island.lk
- divaina.lk
- adaderana.lk
- lankadeepa.lk

Timor Leste :

- thediliweekly.com

Sub-Saharan Africa Local Sources

Angola :

- opais.co.ao
- jornalf8.net
- angola24horas.com
- portaldeangola.com
- angola-online.net
- vozdeangola.com
- jornaldeangola.ao

Benin :

- lanouvelletribune.info
- news.acotonou.com
- lematinal.media
- levenementprecis.com

Burkina Faso :

- lefaso.net
- burkina24.com
- evenement-bf.net
- laborpresse.net

Cameroon :

- journalducameroun.com
- camerounweb.com
- 237actu.com
- 237online.com
- cameroonvoice.com
- lebledparle.com
- thesunnewspaper.cm

DR Congo :

- radiookapi.net
- lesoftonline.net
- acpcongo.com
- lephareonline.net
- groupelavenir.org
- matininfos.net
- cas-info.ca
- actualite.cd
- 7sur7.cd

Ethiopia :

- addisfortune.news

- addisstandard.com
- capitalethiopia.com
- thereporterethiopia.com
- ethiopianmonitor.com
- addisadmassnews.com

Ghana :

- dailyguidenetwork.com
- ghanaweb.com
- graphic.com.gh
- newsghana.com.gh

Kenya :

- kbc.co.ke
- citizen.digital
- nation.africa
- theeastafrican.co.ke

Liberia :

- thenewdawnliberia.com
- liberianobserver.com
- analystliberiaonline.com
- frontpageafricaonline.com
- inquirernewspaper.com
- thenewsnewspaper.online

Malawi :

- mwnation.com
- nyasatimes.com
- times.mw
- faceofmalawi.com
- malawivoice.com

Mali :

- maliweb.net
- malijet.com
- news.abamako.com

Mauritania :

- alwiam.info
- lecalame.info
- journaltahalil.com
- alakhbar.info
- saharamedias.net

Mozambique :

- correiodabeiraserra.com
- canal.co.mz
- mmo.co.mz
- cartamz.com
- verdade.co.mz
- clubofmozambique.com
- portalmoznews.com
- jornaldomingo.co.mz
- tvm.co.mz

Namibia :

- namibian.com.na
- confidentenamibia.com
- thevillager.com.na
- observer24.com.na
- informante.web.na

Niger :

- actuniger.com
- nigerinter.com
- lesahel.org
- tamtaminfo.com
- airinfoagadez.com
- nigerexpress.info
- journalduniger.com

Nigeria :

- guardian.ng
- thenewsnigeria.com.ng
- vanguardngr.com
- thenationonlineng.net

Rwanda :

- newtimes.co.rw
- therwandan.com
- kigalitoday.com
- umuseke.rw

Senegal :

- xalimasn.com
- lesoleil.sn
- enqueteplus.com
- lasnews.sn
- ferloo.com

- nouvelobs.com
- sudquotidien.sn

South Africa :

- timeslive.co.za
- news24.com
- dailysun.co.za
- sowetanlive.co.za
- isolezwe.co.za
- iol.co.za
- son.co.za

South Sudan :

- radiotamazuj.org
- sudantribune.com
- paanluelwel.com
- onecitizendaily.com
- eyeradio.org

Tanzania :

- ippmedia.com
- dailynews.co.tz
- habarileo.co.tz
- thecitizen.co.tz
- mtanzania.co.tz
- jamhurimedia.co.tz
- mzalendo.co.tz

Uganda :

- monitor.co.ug
- observer.ug
- newvision.co.ug
- nilepost.co.ug
- sunrise.ug
- eagle.co.ug

Zambia :

- lusakatimes.com
- mwebantu.com
- diggers.news
- openzambia.com
- lusakavoice.com
- dailynationzambia.com
- zambianewsnetwork.com
- zambianobserver.com

Zimbabwe :

- thestandard.co.zw
- theindependent.co.zw
- herald.co.zw
- chronicle.co.zw
- newsday.co.zw
- thezimbabwean.co
- zimbabwesituation.com
- newzimbabwevision.com
- zimlive.com

Section 4: Assessing Outlet Independence

To ensure data quality, *HQMARC* aims to capture news from the most trustworthy media outlets available for each country. During the construction of *HQMARC*, our team scraped each source every three months to capture new articles published since the last scrape. For countries ranked outside of the top 100 for media freedom by Reporters Without Borders, we consider their media environments restrictive. For all sources in countries with restrictive media environments, we investigate potential sources of bias for all our sources at a three-month frequency. These frequent bias checks are due to changes in coverage of sources and changes in ownership of sources over times.

For countries with restrictive media environments, we aim to source at least half of the articles from sources that are coded politically independent. For countries where less than half of the articles are collected from independent sources, we encourage users to apply weighting.

Human coders follow detailed coding instructions to assess each outlet according to Ideology (5-point scale from Right to Left), Independence from the regime or opposition (Pro-regime, Opposition, or Independent), and Ownership (Government owned or non-government owned). Coders begin by consulting a list of public resources and performing a subjective review of a sample of published articles. Coding instructions and source-level bias data are available on our github at mlp-data-intro/source_metadata/.

Section 5: Assessing and Comparing Coverage

To demonstrate the importance of our custom workflow, we present a case study comparing *HQMARC's* coverage with that of several "big data" media corpora, demonstrating that *HQMARC* captures a significantly more articles from high-quality domestic outlets. We then demonstrate the pitfalls of relying on automated parsing tools without human oversight.

We compare *HQMARC's* coverage of three prominent Bangladeshi news outlets to that of GDELT and Internet Archive. We focus on for Bangladesh three reasons. First, Bangladeshi outlets publish a high volume of articles relative to other countries, making them more likely to attract automated crawlers. Second, the website architecture for each outlet is straightforward, maximizing the likelihood that crawlers and automated parsers will accurately retrieve articles. Third, many Bangladeshi sources publish primarily in English, reducing the additional hurdles of multilingual

parsing. As a result, we regard these outlets as a "best-case scenario" for large-scale media repositories.

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

Appendix B: Civic Space Definitions, Keywords, and Classifier Performance

Section 1: Event Definitions

- Activism
  - Any activity by individuals, community organizations, or NGOs that involve some amount of lobbying, raising awareness, or collective action beyond just the donation of money. Importantly, this does not include protests or activities related to election campaigns.
  - **Example:** "Retired soldiers make up subsidiary of Just Cause.A group of retired soldiers met yesterday morning in a restaurant in the city of San Miguel to learn about the work of the Association Causa Justa, in order to organize a branch in the eastern part of the country.Jaime García, vice president of Causa Justa, explained that this association has been working for a couple of years to defend the interests of retired soldiers, and since then they have organized a branch in San Ana and now they intend to have another group."
- Arrests
  - An institution within the government-controlled security apparatus –i.e. the police, the military, or other– apprehends people or groups of people who are part of an opposition movement or party, a civil society organization, a media organization, or a protest.
  - **Example:** "The Police in Abia on Wednesday arrested 51 persons suspected to be members of the proscribed Indigenous People of Biafra, claiming they 'operate as members of Judaism,' in Umuahia."
- Censor

  - The government actively prevents free speech by individuals in the media, in public or online. This includes banning certain content from individual speech or news stories, dictating how certain concepts or people can be referred to in public speech, or directly dictating agenda setting for media organizations. This category also includes the government censoring internet websites, internet shutdowns, fines on independent media, limitations on foreign ownership of media outlets, and political actors gaining influence within media organizations. Magnitude is a scale (see below; try to determine the importance of the target with information given).
  - **Example:** "The Tanzanian government has suspended newspaper The Citizen for seven days after the publication ran a story on the falling value of the Tanzanian shilling."

- Cooperation

  - Domestic political, social, or business actors collaborate on one or a range of issues or demonstrate an intent to do so. This does not include intragovernmental cooperation, except for local and federal cooperation. Cooperation indicates a willingness for domestic actors to work together to resolve important issues. Cooperation involving international actors is not a civic space event, but is an RAI event.
  - **Example:** "Opposition movement NJPE has pledged to work together with President Muhammad to reduce rampant poverty in the capital's outskirts."

- Corruption

  - The abuse of entrusted power for private gain. This includes street-level corruption (for instance, bribes to police), high-level corruption (in public contracting, for instance), and legal actions around corruptions.
  - **Example:** "Ex-President Sarkozy gets Jail Sentence for Corruption in France"

- Coup

  - Changes in government or persistence of government that are not in the rules of succession or transition. Coups, refusing to cede power, or a power grab after an unfair election are examples of this event type. Peaceful government transitions are nonviolent transfers of power or legitimate continuity of government elected by democratic means and accepted by a majority of political forces. Code peaceful change as '1' in the direction column. Speech is the pledge to accept the result of an election (a threat of a coup falls under 'political threat').
  - **Example:** "Forces loyal to Turkey's president quashed a coup attempt in a night of explosions, air battles and gunfire that left at least 161 people dead and 1,440 wounded, yesterday. President Recep Tayyip Erdogan vowed that those responsible "will pay a heavy price for their treason".

- Defamation Case

  - Cases in which an individual or a group related to an opposition movement or party, current or former government officials, a civil society organization, a media organization, a business leader, a member of a minority group, or a protest are accused, usually by a government official, of: directly defaming/libeling/slandering the government or one or some of its members. Alternatively, these are: strategic lawsuits against civic actors

to intimidate public participation. This category is a subset of the 'Legal Action' event type (see below).
  - **Example:** "A Phnom Penh court on Friday found veteran opposition chief Sam Rainsy guilty of defaming Prime Minister Hun Sen and ordered him to pay damages of $1 million, the latest blow to an opposition crippled by legal cases this year."

- Disaster
  - This category includes: (1) natural disasters such as earthquakes, hurricanes, floods, famines or food crises or any catastrophic event that results or may result in serious damage. loss of life, and/or political implications; (2) infrastructural accidents that can endanger the population –a dam or pipe bursting, the sinking of a ferry, etc.) The goal is to focus on the types of events that might reflect on government accountability and civic space, not minor events like traffic accidents. Note: this includes deaths or economic strife caused by disasters, including reports on deaths/infections caused by COVID-19.
  - **Example:** ROMBO district residents in Kilimanjaro region are in danger of facing food shortage if concerted efforts to control wildlife in consuming crops are not employed.

- Election Activity
  - Reporting on regular electoral activities including the introduction of candidacies, the conduct of campaigns, the announcement of results, and the formal transfer of power for public office. This includes situations where the incumbent party/candidate retains power. Normal election results are an example of this event type. Specifically, this includes peaceful government transitions that are nonviolent transfers of power or legitimate continuity of government but does not include coups, which are accounted for separately.
  - **Example:** "The President of the Executive Board of the Serbian Progressive Party, Darko Glisic, said that the SNS won a convincing victory with 60.2 percent of the votes, based on the processed 221 out of 231 polling stations where the voting for the parliamentary elections in Serbia was repeated."

- Election Irregularities
  - The altering or attempted altering through legal or extra-legal manipulation of the conduct, rules or results of elections or electoral processes. This could include rescheduling/postponing/cancelling regularly-scheduled elections, calling irregular elections or constitutional referenda, hampering the work of independent election observers, eroding the autonomy/authority of the electoral commission, as well as any other institutional change that directly affects the electoral process. This additionally could include the cancellation of party lists and the boycott of elections by opposition, as well as calls for recounts or other forms of contesting elections. Finally, this includes vote-buying, intimidation, or vote-rigging. This does not include standard election proceedings.
  - **Example:** "Nigeria's electoral authority has delayed presidential and national assembly elections by one week amidst protests from the two main opposition parties. The government alleged logistical problems with ballot delivery to justify the delay."

- Legal Action

  - Legal action refers to the prosecution or investigation of criminal activity or disputes over government authority, as well as the trials, convictions or dismissals that result. This event is related to civic space if the targeted people are part of an opposition movement or party, current and former government officials, a civil society organization, an NGO, a media organization or member, a (legal or illegal) business leader, member of minority group, or a protest. This specifically does not include arrests, which are defined as its own event type. Defamation cases are a subset of 'Legal Action' (see above).
  - **Example:** "Kosovo President is Indicted for War Crimes for Role in War with Serbia."

- Legal Change

  - Legal change refers to any changes or proposed changes in the laws of a nation in such a way as to affect civic space. This includes legal restrictions on speech, political groups, NGOs, and the changing of constitutions as well as changing of the powers of the government. This also includes some restrictions on assembly, but does not include curfews and other martial law declarations, which are specifically covered by other event categories.
  - **Example:** "Venezuela's constituent assembly yesterday unanimously passed a law that mandates punishment including a prison sentence of up to 20 years for anyone who instigates hate or violence on the radio, television or via social media."

- Cooperation

  - Domestic political, social, or business actors collaborate on one or a range of issues or demonstrate an intent to do so. This does not include intragovernmental cooperation, except for local and federal cooperation. Cooperation indicates a willingness for domestic actors to work together to resolve important issues. Cooperation involving international actors is not a civic space event, but is an RAI event.
  - **Example:**

- Lethal Violence

  - Any action of aggression by a government entity, organized group or individual that results in the death of one or more people, excluding crimes of passion.
  - **Example:** "The Kaduna state government, Friday evening, disclosed that 33 women and children were killed by rebels in Kajuru local government of Kaduna state, less than twenty-four hours to the conduct of the presidential and parliamentary elections."

- Martial Law

  - The executive branch declares a state of emergency or suspends, temporarily or indefinitely, the ability of citizens to gather or protest against the order.
  - **Example:** "On Wednesday, President Duterte approved the extension of martial law in the country's volatile south by a year due to continuing threats by Islamic State group-linked militants and communist insurgents."

- Mobilize Security Forces

  - An event in which the government mobilizes police forces, military troops or government-affiliated militias.

  - **Example:** "More than 500 security personnel have being mobilised for Saturday's governorship election in Sokoto State, says the state commissioner of police (CP), Alhaji Adisa Bolanta."

- Non-Lethal Violence

  - Any action of aggression by a government entity, organized group or individual that physically harms one or more people or property but does NOT result in death, excluding crimes of passion.
  - **Example:** "At least four persons were injured at Oruk Anam Local Government Area of Akwa Ibom State on during yesterday's national election. Reports claim that unknown assailants attempted to snatch ballot boxes while voting was still ongoing."

- Protest

  - Planned or spontaneous public mobilization of a large group of people. Labor strikes, political rallies and riots are also included in this category. When protest events involve deaths, they are coded as lethal violence, but this does not apply for non-lethal violence incidents.
  - **Example:** "A reported two thousand people took to the streets yesterday in Nairobi to protest rising fuel prices, which have doubled since the beginning of the year."

- Purge/Replace

  - Purge/replace refers to the removal, firing or resignation of individuals from a government position or the replacement of previously removed individuals, including when the removals/firings themselves are not directly mentioned in the article. This includes the resignation of the chief executive. This description applies to purging of targets such as the bureaucracy, courts, military, police, state-owned companies, or members of political parties, among others.
  - **Example:** "Poland's government carried out a sweeping purge of the Supreme Court on Tuesday night, eroding the judiciary's independence, escalating a confrontation with the European Union over the rule of law and further dividing this nation."

- Raid

  - Individuals or organizations are assaulted or aggressively coerced. Their property may be encroached or damaged as a result. Examples include a raid on newspaper offices. Victims themselves suffer no physical harm. This category also includes the government shutting down opposition organizations, NGOs, etc.
  - **Example:** "A Vanguard newspaper office located on Bassey Duke Street was, yesterday afternoon, raided by hoodlums, who carted away large sums of money and destroyed computers and other equipment."

- Threats

  - A statement of a clear and explicit intention to inflict pain, injury, damage, or other hostile action on an individual or organization. Our targets and actors of interest are part of an opposition movement, the media, a political party, government, or a civil society organization.

- **Example 1:** "A top Kenyan newspaper published a fake death notice of a prominent opposition financier on Wednesday, a bizarre error that rights groups interpreted as another sign of an anti-democratic slide. The Daily Nation apologized by mid-morning for publishing the funeral announcement for businessman Jimi Wanjigi, whose picture, history and family details were used but whose name was slightly altered. The paper said the ad was published in error and it was working with police to uncover who placed it. After a week of arrests of opposition politicians and a crackdown on independent media, a prominent rights campaigner said the announcement amounted to a death threat to Wanjigi, who funded opposition leader Raila Odinga's election campaign last year and whose house was raided by police in October."
- **Example 2:** "Nigeria's main labor unions threatened a nationwide strike over recent increases in gas prices."

Section 2 Keywords and Their Use

This document lists the categories where keywords are currently being deployed as well as a brief explanation for their use.

Legal Action

- Keyword lists
  - Keyword list one: case | lawsuit |sue | suit | trial | court | charge | rule | sentence | judge
  - Keyword list two: defamation | defame |libel | slander | insult | disparage | lese majeste | lese-majeste | lese majesty | reputation
- Purpose of Keywords
  - The purpose of these keywords is to move articles from legal action that are actually defamation case to the appropriate category. There is no longer a defamation case category, it is now entirely a subset of legal action. The first set of keywords filter out any instances where there are accusations of defamation/libel/slander that are not actual cases, but merely statements. The second set of keywords ensures that the cases are actually related to defamation, since we found that oftentimes, non-defamation cases where being assigned to this event category. The process is two-fold, requiring a key word from both lists.
  - If keyword from both lists are not present, the article is left as legal action

Censor

- Keyword list: Freedom | assembly | association | movement | independent | independence | succession | demonstrate | demonstration | repression | repressive | crackdown | draconian | intimidate | censoring | controversial | censor | muzzle | restrictive | restrict | authoritarian | non-governmental organizations | NGOs | media | parties | civil society | opposition | critics | opponents | human rights groups | arbitrary | stifling | ban | strict | boycott | protests | dissent | demonstrators | journal.* | newspaper | media | outlet | censor | reporter | broadcast.* | correspondent | press | magazine | paper | black out | blacklist | suppress | speaking | false news | fake news | radio | commentator | blogger | opposition voice | voice of the opposition | speech | broadcast | publish | limit.* | independ.* | repress.* | journalist | newspaper | reporter | internet | telecommunications | magazine | shut down | broadcast | radio

- Purpose of Keywords
  - The purpose of these keywords is to filter out instances where restrictions are applied that are not censorship. This was created primarily in response to closings of schools, businesses, and government offices in relation to COVID-19, which were frequently classified as censorship.
  - If one of the keywords is not present in the article, it is reclassified as -999

Legal Action/Purge/Arrest

- Keyword list
  - embezzle | embezzled | embezzling | embezzlement | bribe | bribes | bribed | bribing | gift | gifts | gifted | fraud | fraudulent | corrupt | corruption | procure | procured | procurement | budget | assets | irregularities | graft | enrich | enriched | enrichment | laundering | fraudulent.
- Purpose of Keywords
  - The purpose of these keywords is to assign a second event category to those articles in arrest, purge, and legal action that also feature corruption. Articles in those categories that feature these words are double counted as both corruption and the original category.

Corruption

- Keyword lists
  - For arrest: arrest; detain; apprehend; capture; custody; imprison; jail
  - For legal action: legal process; case*; investigate*; appeal; charged; prosecute*; case; lawsuit; sue; suit; trial; court; charge; rule; sentence; judge
  - For purge: resign*; fire*; dismiss; sack; replace; quit.
- Purpose of Keywords
  - The purpose of these keywords is to assign a second event category to those articles in corruption that also feature arrests, legal action, and purges. Articles in those categories that feature these words are double counted as both corruption and the original category.

##Section 3: Classifier Performance

Table 6: Performance metrics for fine-tuned RoBERTa classification model. -999

Event Category	Precision	Recall	F1
Arrest	0.91	0.88	0.89
Protest	0.85	0.98	0.91
Legal action	0.77	0.75	0.76
Disaster	0.87	0.86	0.86
Censor	0.76	0.95	0.84
Election activity	0.78	0.84	0.81
Election irregularities	0.72	0.68	0.70
Activism	0.95	0.83	0.88
State of Emergency	0.92	0.90	0.91

Event Category	Precision	Recall	F1
Cooperate	0.50	0.67	0.57
Coup	0.68	0.83	0.75
Non-lethal violence	0.79	0.81	0.80
Lethal violence	0.90	0.82	0.86
Corruption	0.74	0.71	0.73
Legal change	0.84	0.80	0.82
Security mobilization	0.83	0.77	0.80
Purge	0.91	0.86	0.88
Threats	1.00	0.78	0.88
Raid	1.00	0.83	0.91
Irrelevant events	0.81	0.79	0.80

Appendix C: Descriptive Maps

These temporal world maps provide a comprehensive overview of civic space coverage patterns across all 65 developing countries in the HQMARC corpus from 2012 to 2024. The maps reveal significant variation in both the intensity of civic space reporting and the types of events that dominate coverage across countries and over time. Geographic clustering is evident, with certain regions showing consistently higher civic space article rates and distinct event profiles that reflect regional political and social dynamics. The temporal dimension captures important shifts corresponding to major political events, democratic transitions, and periods of civic space restriction across the developing world.

Civic Space Coverage Intensity

Dominant Civic Events by Country

Appendix D: Validation Figures

These validation figures demonstrate the effectiveness of ML4P's event detection methodology by examining State of Emergency declarations across different regions during the COVID-19 pandemic. The vertical dashed red line at March 2020 marks the onset of the global pandemic, when countries worldwide implemented emergency measures including lockdowns, curfews, and restrictions on civil liberties. These plots validate our approach by showing systematic spikes in State of Emergency reporting across all regions coinciding with known policy responses to the pandemic.

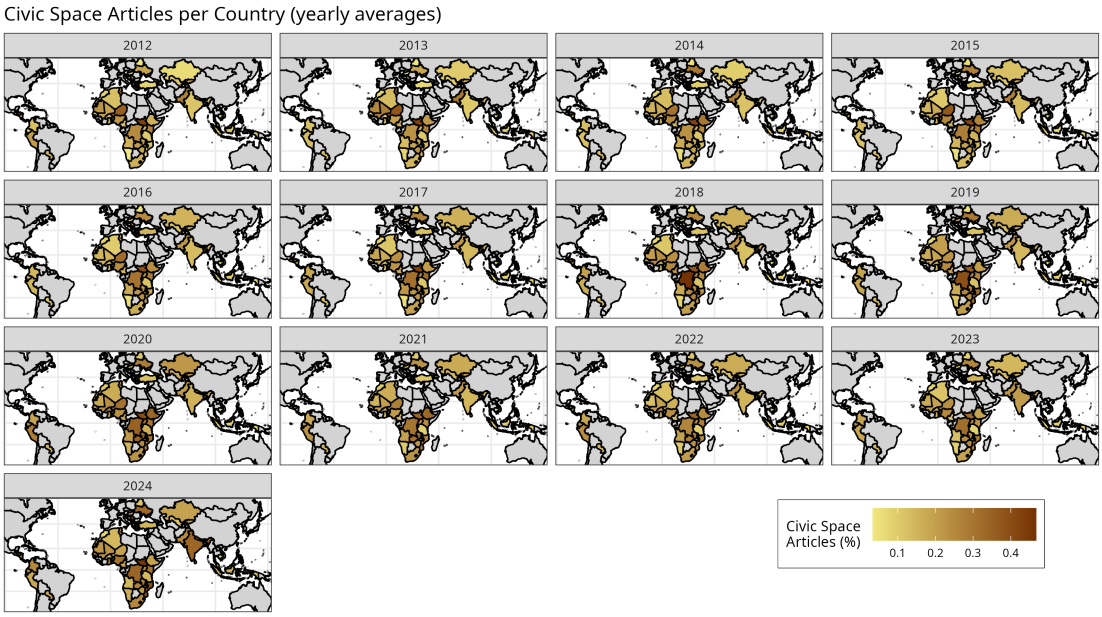


Figure 3: Civic space articles as percentage of total coverage by country-year. Countries are colored according to the proportion of their total news coverage dedicated to civic space events, providing insight into the relative salience of civic space issues across different national contexts. Light gray areas indicate countries without data coverage. The maps show temporal evolution from 2012 through 2024, revealing significant variation in civic space reporting intensity both across countries and over time.

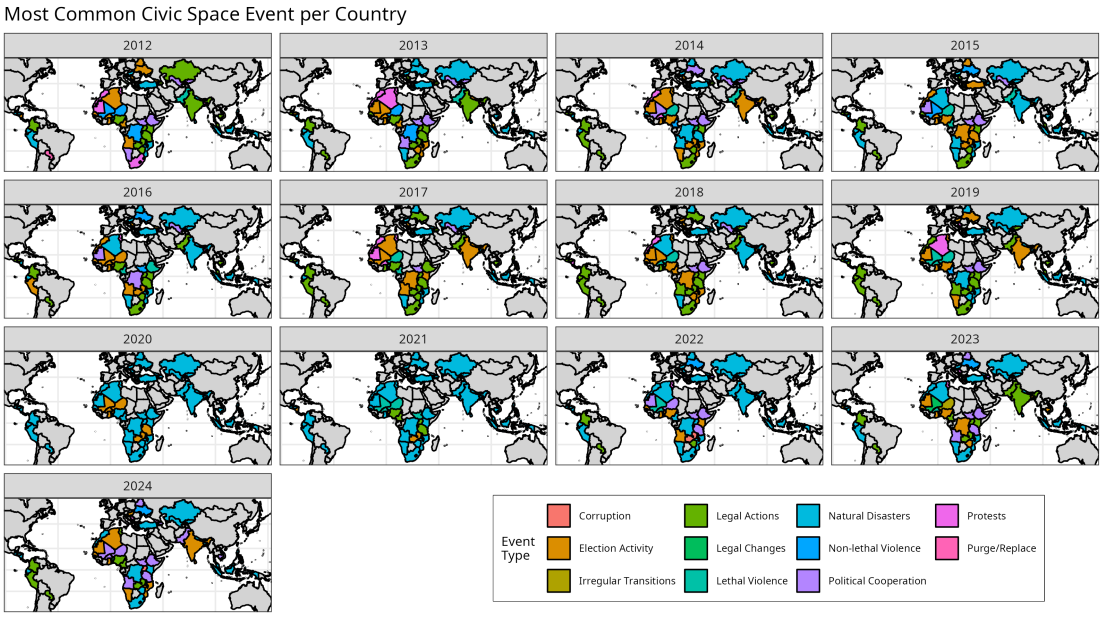


Figure 4: Most frequently reported civic event type per country-year. Countries are colored according to the civic event category with the highest reporting frequency in each year, showing the diversity of civic space concerns across different national contexts. The maps reveal regional patterns in dominant event types and temporal shifts in civic space priorities, from electoral activities to protests, corruption reporting, and other civic engagement forms.

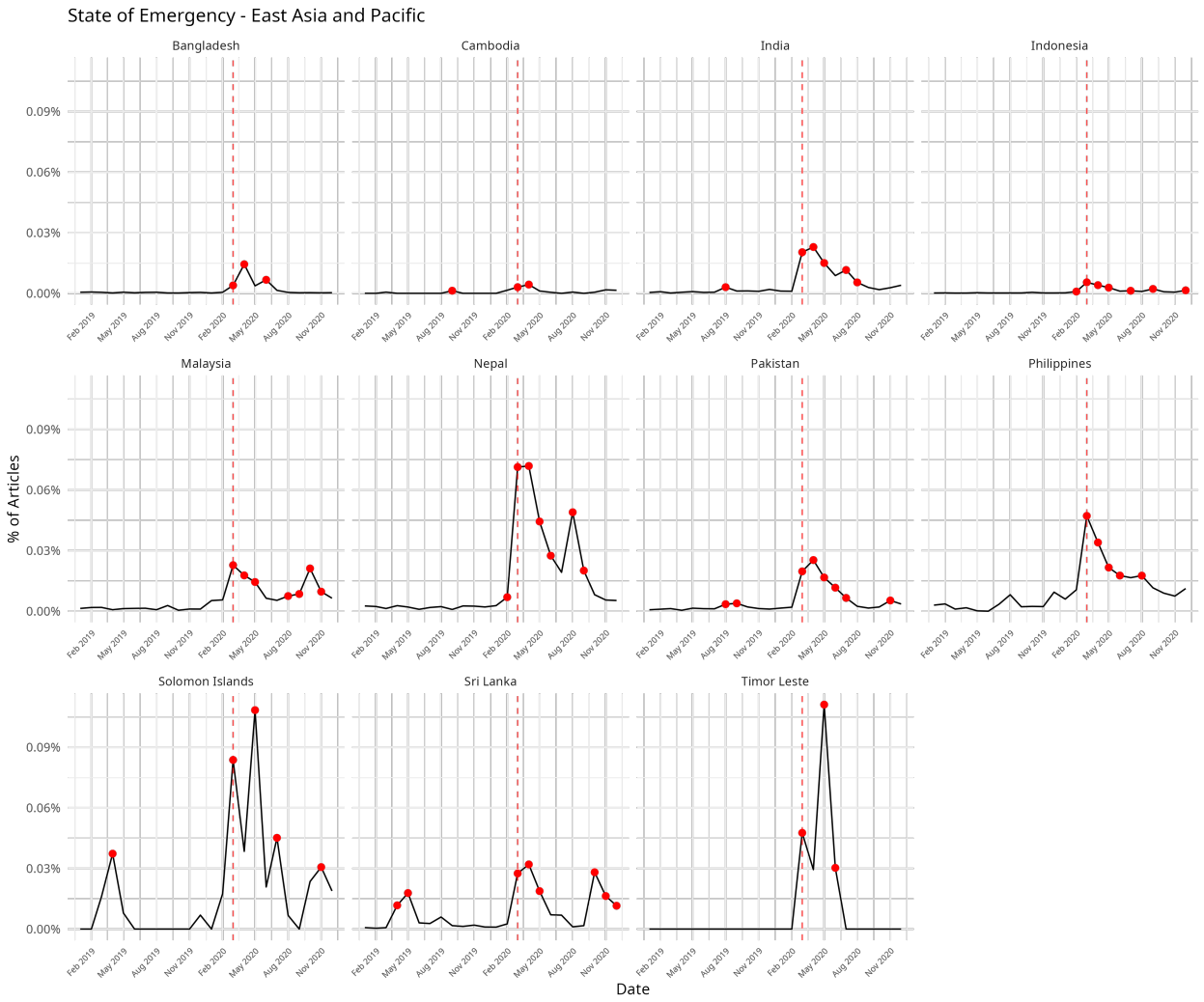


Figure 5: State of Emergency events detection in East Asia and Pacific countries from 2019-2020. The plot shows the percentage of articles devoted to State of Emergency coverage over time, with red points indicating detected shock events. The vertical dashed red line marks March 2020, the beginning of the COVID-19 pandemic. Clear spikes are visible across multiple countries starting in March 2020, corresponding to the implementation of emergency measures and lockdowns in response to the pandemic.

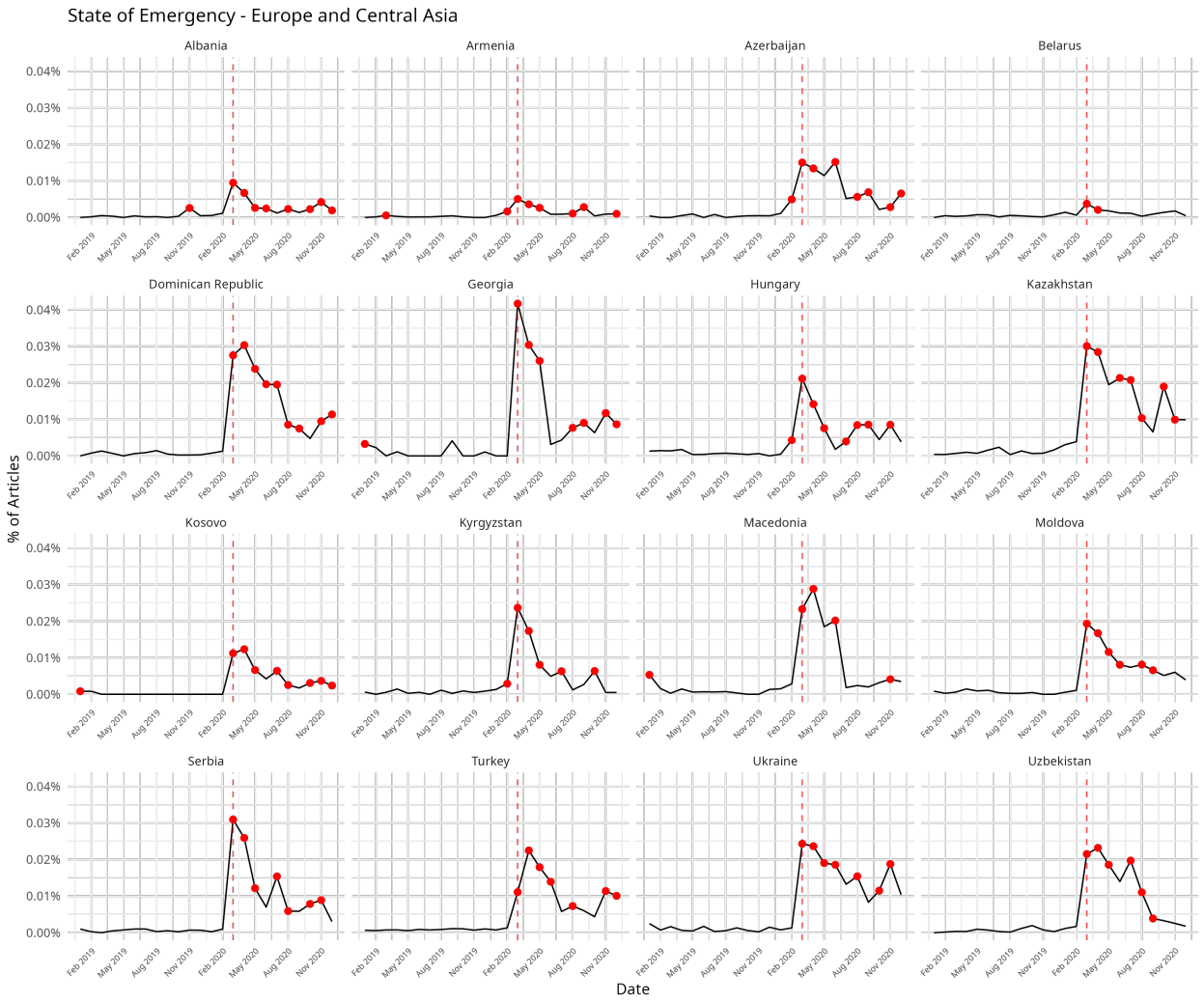


Figure 6: State of Emergency events detection in Europe and Central Asia countries from 2019-2020. The plot shows the percentage of articles devoted to State of Emergency coverage over time, with red points indicating detected shock events. The vertical dashed red line marks March 2020, the beginning of the COVID-19 pandemic. Pronounced spikes are evident across the region starting in March 2020, reflecting the widespread implementation of emergency measures in response to the pandemic.

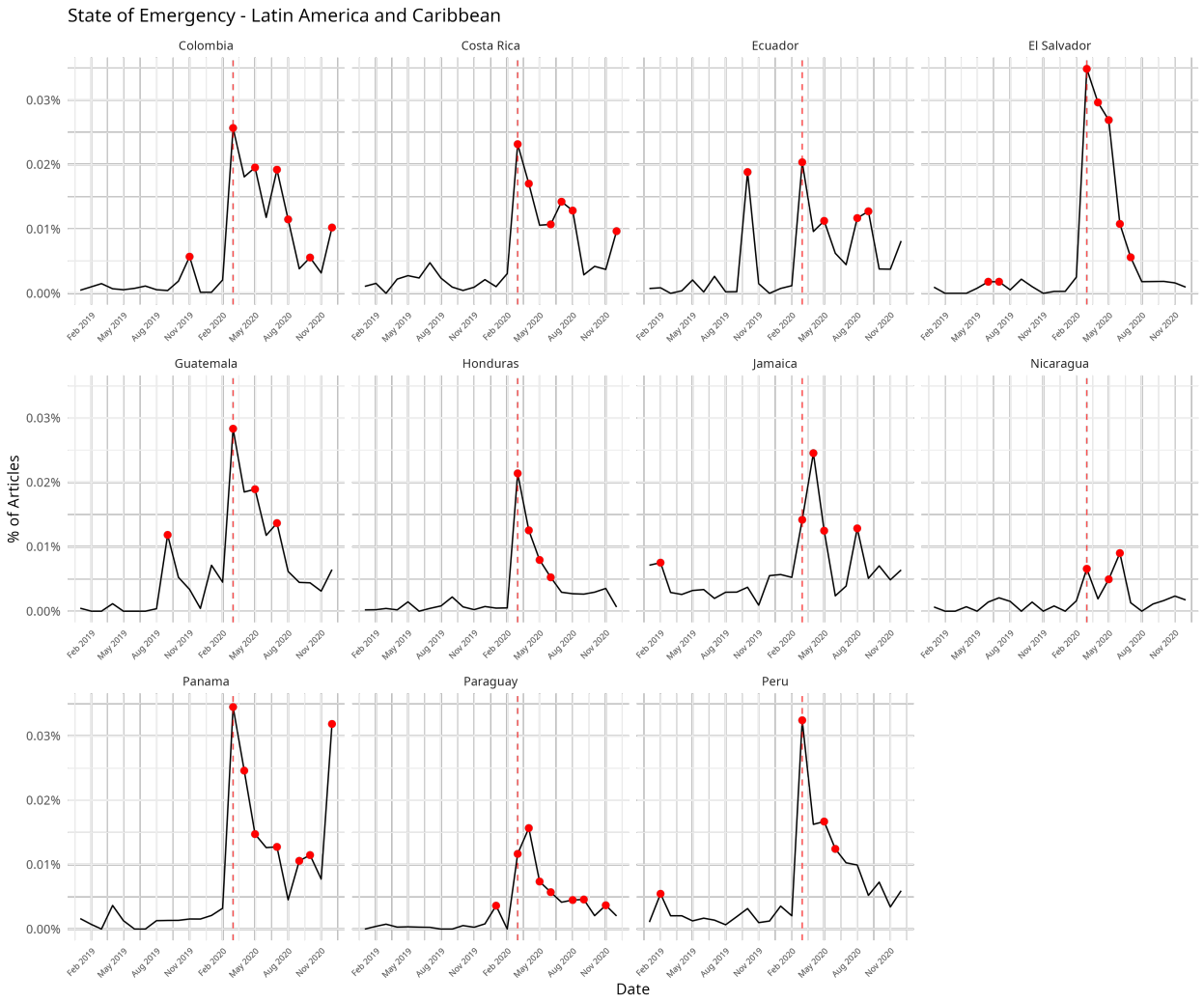


Figure 7: State of Emergency events detection in Latin America and Caribbean countries from 2019-2020. The plot shows the percentage of articles devoted to State of Emergency coverage over time, with red points indicating detected shock events. The vertical dashed red line marks March 2020, the beginning of the COVID-19 pandemic. Significant increases in State of Emergency reporting are visible across the region starting in March 2020, corresponding to the implementation of pandemic-related emergency measures.

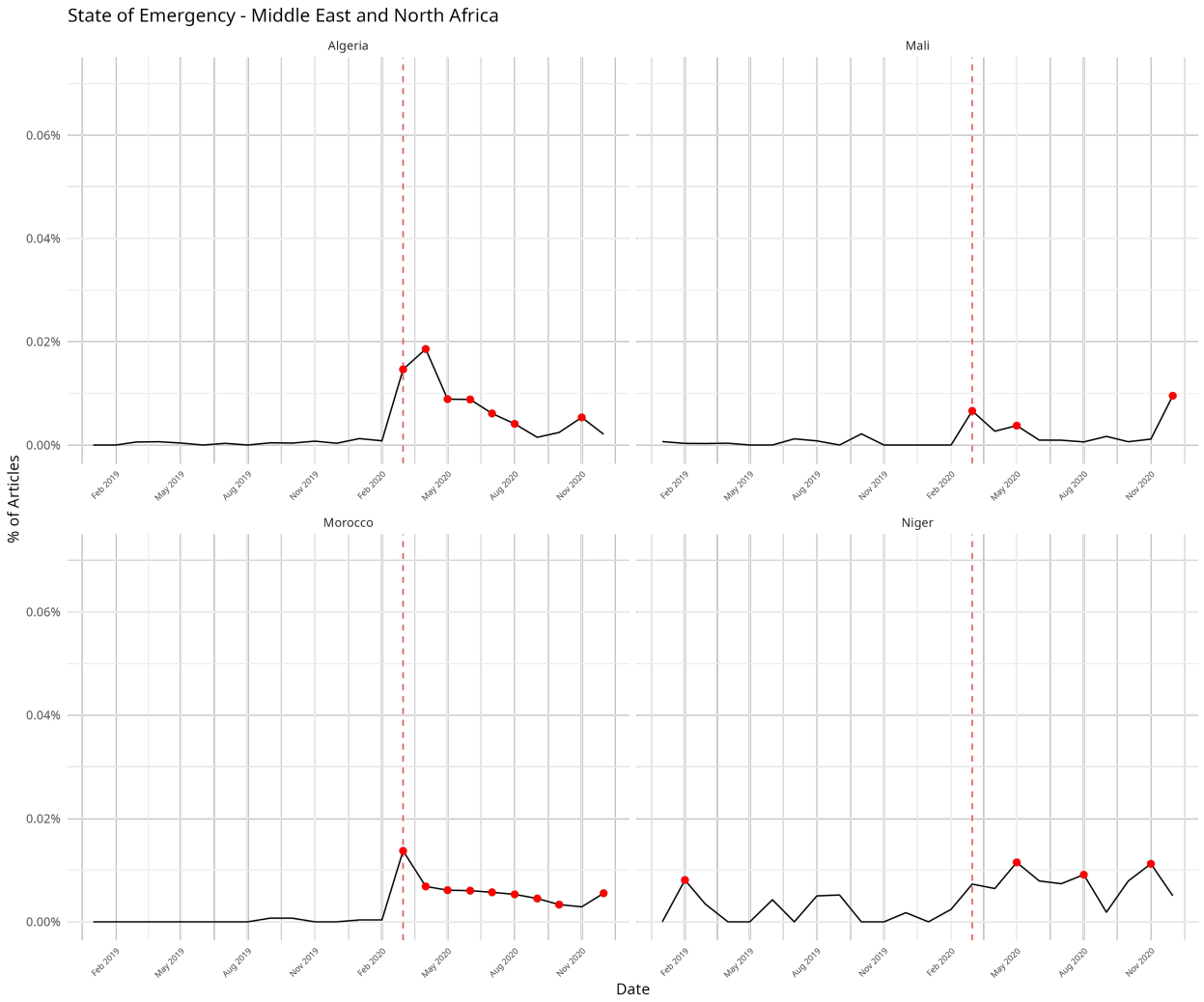
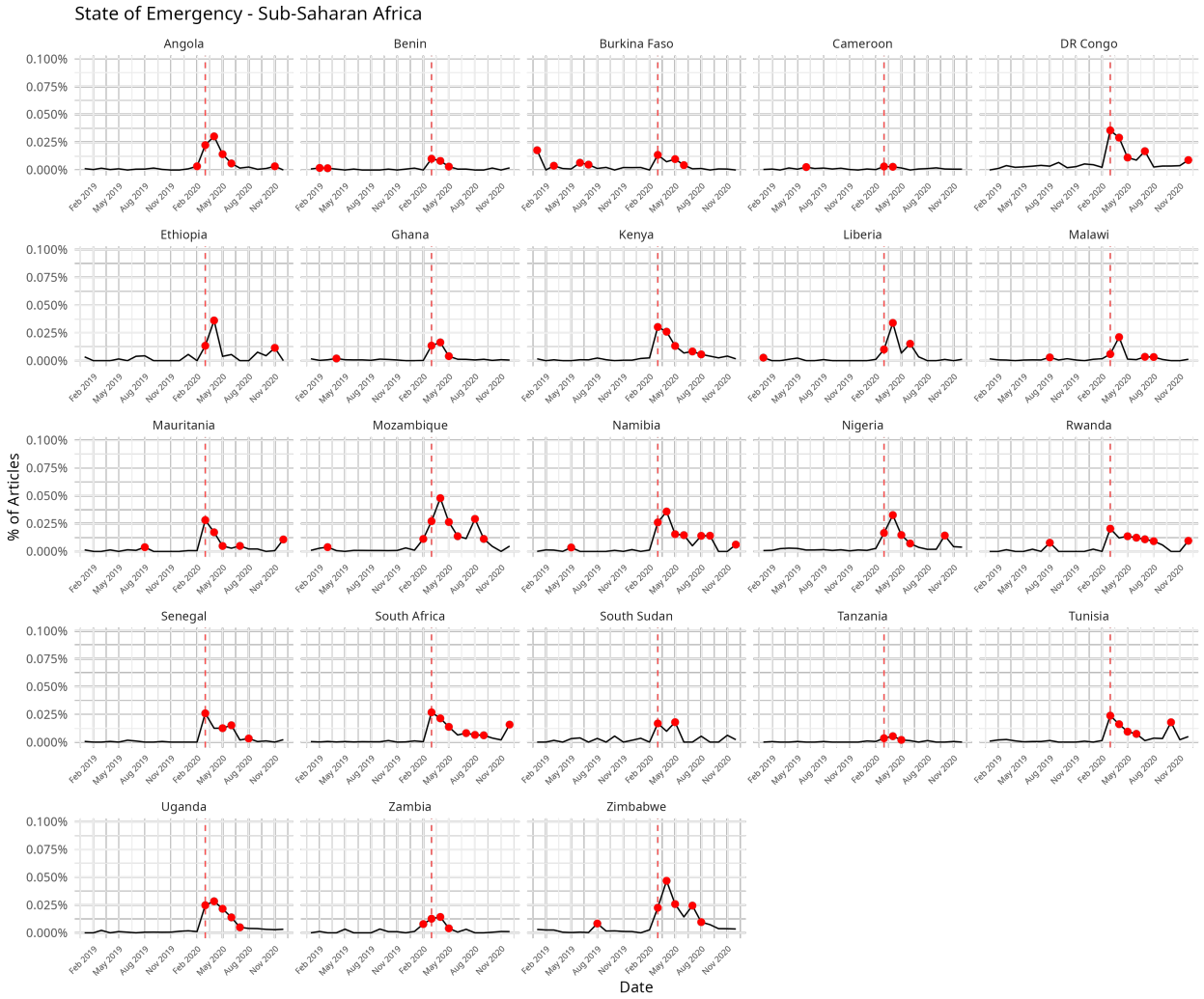


Figure 8: State of Emergency events detection in Middle East and North Africa countries from 2019-2020. The plot shows the percentage of articles devoted to State of Emergency coverage over time, with red points indicating detected shock events. The vertical dashed red line marks March 2020, the beginning of the COVID-19 pandemic. Clear spikes in State of Emergency reporting are visible across countries in the region starting in March 2020, reflecting the implementation of emergency measures in response to the pandemic.

State of Emergency Detection by Region



Appendix E: Detection of Civic Space Shocks

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

Next, we trained a neural network model to detect spikes in a human-labeled dataset covering

the full time-series for 30 country-event pairs. We conducted human-labeling by asking humans to identify months with visually distinct, sharp increases in our event measures. Human labelers were instructed to identify no more than 15% of overall months as shocks, ensuring that peaks are not overly frequent in highly variable data while still capturing meaningful shifts in lower-variance event types. When either the statistical or neural network model detect a shock, we label that month as a shock in the data.

Appendix F: Event Validation

For each of the 40 country-event-month combinations, we retrieve all relevant articles published by domestic outlets in the *HQMARC* corpus to determine whether these shocks are true or false positives. Because some of these 40 country-event-months with shocks contained hundreds of articles,[1] we used GPT-4o[2] to generate brief summaries of the five most important events reported on in the sample of articles. A research assistant then reviewed both (a) the original articles (or a random subset of 50, if more than 50 were available) and (b) the GPT-4o summaries. They evaluated each of the top five summarized events on four dimensions:

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

Appendix G: Forecasting Data and Models

Control variables include: a composite economic covariate calculated as a within-country z-score of available monthly Trading Economics indicators for each country. Given the importance of both gradual and sudden shifts in political conditions, we lag all features by up to 12 months, allowing the model to detect both long-term precursors (such as increasing government repression) and short-term triggers (such as post-election violence). We account for persistence by including an indicator distinguishing between new HLTA onsets and continued warnings, a country-specific Bayesian prior that adjusts for baseline differences in advisory issuance, and a COVID-19 indicator to control for the pandemic-driven shock in travel warnings in 2020.

Results on our 6-month model yields a ROC-AUC of 0.87, AUC-PR of 0.31, and a Brier score of 0.14 (dummy AUC-PR = 0.01), while the rolling-window variant (± 1 month) increases ROC-AUC to 0.90, AUC-PR to 0.57, and lowers the Brier score to 0.13, again with a dummy AUC-PR of 0.02.

Above and beyond these traditional measures of forecast accuracy, we also trained the model on data through December 2023 and generated predictions for advisory onsets in March and June 2024. The model correctly flagged Bangladesh for June 2024, which later received an HLTA due

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

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

to the July Revolution. It also identified Zimbabwe and Liberia as high-risk for March 2024, and while DOS did not issue an HLTA for these countries, Zimbabwe expelled USAID staff in March, prompting U.S. officials to issue multiple security statements. These cases highlight the practical value of early warnings derived from high-frequency civic space indicators, even when they do not always align with official government actions.