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

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

August 21, 2025

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

[†] These authors contributed equally to this work.

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

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

1. Introduction

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

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

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

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

HQMARC employs a human-supervised, source-specific scraping methodology that prioritizes data quality and comprehensiveness over the broad but shallow coverage typical of automated web crawlers. This process proves particularly valuable for domestic news sources, whose websites are less stable than international outlets. Our efforts yield significant advantages over both "big data" media repositories like GDELT, Internet Archive, and Common Crawl and expensive commercial databases like Factive and LexisNexis, delivering a stable corpus composition with superior linguistic diversity and coverage of high-quality developing-country sources. However, *HQMARC's* size and linguistic diversity makes human classification prohibitively expensive. To produce *ML4P's* structured data on civic space events, we apply free, open-source computational tools to translate and extract information from each article, identifying the main event being reported on and the country in which the event occurs.

This paper proceeds in six parts. In the next section, we discuss how *ML4P* complements existing data on regimes and opens up new avenues for research. Thereafter, we describe the data production process and the advantages of our approach. We compare the coverage of *HQMARC* to other major

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