Technical Report on the Production of Civic Space and RAI Event Count Data*

Explore the data using our Data dashboards

Erik Wibbels[1,2] Jeremy Springman[1] Donald Moratz[1]
Zung-Ru Lin[1] Hanling Su[1]

March 16, 2024

[1] PDRI-DevLab, University of Pennsylvania
[2] Department of Political Science, University of Pennsylvania

1. The Machine Learning for Peace (MLP) Pipeline

Civic space is expanding and contracting on a daily basis in countries around the world. Social movements and aspiring autocrats alike frequently seize on sudden economic, humanitarian, and political crises to contest the extent of fundamental rights and freedoms that underpin democratic accountability. However, existing sources of quantitative data on civic space are aggregated by year, meaning that they summarize how conditions have changed over a 12-month period, masking sudden changes occurring over weeks or months, and are published on an annual basis, meaning that they only become available many months after the year under consideration has concluded.

While these data can provide valuable insights into macro-historical trends, the absence of high-frequency data on civic space limits what researchers can learn about the factors that precipitate shifts in civic space, and the absence of real-time data limits the potential for practitioners to incorporate this learning into programming decisions. To address these gaps, the MLP research infrastructure produces up-to-date data on civic space for a large sample of countries, uses these data to identify historical patterns between economic, social, and political conditions and future shifts in civic space, and generates regular, practitioner-facing reports that apply this learning to the most recent data to make predictions about civic space activity in the coming months. This memo provides an overview of the novel data production pipeline.

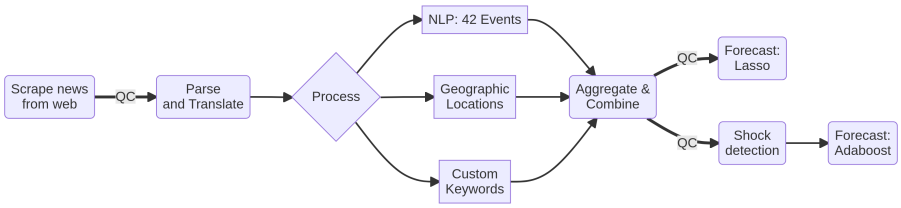
Built under the auspices of INSPIRES, the MLP research infrastructure uses recent advances in computer science to provide high-frequency data on civic space and foreign authoritarian influence "events". Event data is a common resource in social science research. An "event" in a political event dataset is a structured record of a politically relevant occurrence, such as a protest or a change in a country's laws. In collaboration with our INSPIRES Consortium partners and USAID, we have developed codebooks that define 20 types of events relevant to civic space and 22 events capturing

*This project is funded by the United States Agency for International Development (USAID) Bureau for Democracy, Human Rights, and Governance (DRG).

influence by foreign authoritarian countries (referred to as "resurgent authoritarian influence" or RAI). The consistent structure of event datasets allows researchers to track trends over time, expose relationships between events, and build predictive models.

Historically, the production of event data has required enormous amounts of human input, meaning that a very limited number of information sources can be incorporated and significant lags between the occurrence of events and the production of structured data are required. Furthermore, the need to develop precise coding rules and train human coders renders these systems inflexible to changes in source materials and the events being identified. Researchers have been left with a choice between spending large amounts of time and money on a risky attempt to build their own event data or adapting their focus to fit inside the scope of existing datasets. However, recent advances in machine-learning-based natural language processing (NLP) have radically altered what is possible for machine-generated event data systems.

Below we describe the entire pipeline that generates our innovative civic space event data. We walk readers through the key elements of Figure 1, which provides a graphic presentation of the pipeline.



2. Event Data Background

It is useful to begin with a general overview of event data in the social sciences. Categorized by their means of production, event datasets come in one of two flavors: hand-coded or machine-coded. The difference between the two, which the names betray, is that hand-coded event datasets are built by humans taking inputs like news stories and government reports and extracting the pertinent information while machine-coded systems rely on software to perform the same task. Each approach has its own advantages and disadvantages.

Prominent hand-coded event datasets include the Armed Conflict Location & Event Data Project (ACLED) and the Uppsala Conflict Data Project Georeferenced Event Dataset (UCDP GED)—two projects focused on political violence. Both projects cover dozens of countries over several decades and are also used extensively inside and outside of academia. Those projects have important advantage—most importantly, well-trained humans can be more accurate and adaptable than rules-based machine coding systems.

However, those advantages are delivered at a high price and a slow speed. The cost of human input for event data means that hand-built datasets are generally forced to limit their geographic or temporal scope. Building such an effort focused on civic space events for a large number of countries would be enormously expensive. Hand-built data also has a significant time delay built-in. In the time it takes for an efficient coder to process a single event record, a reasonably powerful

computer can process thousands of records. This means that while machine-based systems can update in near real-time, hand-built datasets generally lag by at least several months.

To address the cost and speed limitations of hand-coded event data, scholars began work on systems for machine-coding events. With machine coding, much smaller teams could process much larger quantities of data at a much lower cost. Projects like the Integrated Crisis Early Warning System (ICEWS) and the Global Database of Events, Language, and Tone (GDELT) quickly grew to hold millions of events. However, every machine-coded event dataset in the social sciences relies on the same basic—often flawed—process for generating events. Each sentence is parsed for syntax with a rule-based or statistical parser and then the components are checked against an expansive list of rules and exceptions for possible events. This approach requires exhaustive and inflexible rules, and it limits the flexibility of the system when applied to different dialects, translated speech, or novel event descriptions. As described below, many of these shortcomings have been overcome with recent innovations in NLP.

3. NLP background

A brief overview of the relevant developments in natural language processing (NLP) is useful for understanding the advantages of MLP over older systems. Sitting at the intersection of linguistics, computer science, and artificial intelligence, the goal of natural language processing is to build machines that can derive understanding or meaning from human language.

Early NLP techniques were rules-based systems that relied on dictionaries and fixed patterns to classify text or extract information. For example, early systems designed to classify the sentiment in a given sentence would look up each word in the sentence in a hand-built dictionary that would classify words as positive or negative and then average over all words in the sentence to determine its sentiment. Rules- based NLP systems grew in complexity and -to a degree- accuracy over time but consistently struggled with complexity, context, and flexibility. The encoding systems behind every current machine-coded event dataset in the social sciences are (largely) rules-based.

In recent years, innovations in NLP improved upon rules-based approaches. Most significantly, researchers began developing methods to represent words as continuous vectors known as embeddings. The goal of word embeddings is to build a mathematical representation of the meaning of a word as captured in a vector of continuous numbers. An embedding system transforms a word like "king" into a vector of numbers (i.e. [0.012, 0.131, …, 0.003]) that represent that word's meaning. Each distinct value in the vector captures a unique linguistic feature of that word, and words that are used in similar ways will have similar values. By representing multiple linguistic features of each word and the similarity of each feature of each word to each feature of every other word, embeddings form a complex representation of language.

For many years, each such system suffered from the same flaw: each word in a corpus can have only a single representation. This means that while a word's context in the training phase is used to position it in a high-dimensional vector space, each word's embedding from that space will be fixed, no matter what context it appears in later. This shortcoming left word embeddings vulnerable to several linguistic quirks including:

1. Polysemy: the capacity of words to have connected but distinct meanings

    a. Ex: "Man" can refer to the human species as a whole or to male individuals but will have the same embedding in either use.

2. Homophony: the capacity of words to have the same spelling but entirely disconnected meanings.

   a. Ex: "Date" can refer to a romantic rendezvous, the time an event occurs, the sweet fruit of a palm tree, or several other things. However, the word "date" will always have the same embedding.

3. Modifiers: words the substantially modify the meaning of surrounding words

   a. Ex: In the sentences "I am happy" and "I am not happy", "happy" will have an identical embedding.

Addressing these limitations took a number of advances. Primary amongst them is the emergence of general language models based on the transformer. Transformer models like the Bidirectional Encoder Representations from Transformers (BERT),[1] the Generative Pre-trained Transformer (GPT), and their many variants, represent the state-of-the-art for many NLP tasks such as translation, passage summarization, and text classification. These models greatly out-perform the models that are currently standard in most social scientific applications of NLP.

Transformer models excel by learning the structure of human language and the context-dependent meaning of words. This approach lets researchers train the models on enormous amounts of text data using a semi-supervised approach before fine-tuning the base model for specific tasks. This approach, generally called 'transfer learning', drastically decreases the resource demands of model creation while maintaining the high-performance of the original models.

These characteristics make the transformer the perfect tool for building custom event data. This project brings the power of transformers within the reach of a broad set of researchers. With a relatively small set of training data (approximately 100 examples per event category) and a reasonably powerful computer.

4. The Pipeline

4.1 Collection and pre-processing of articles

Data collection

The first step in the construction of an original repository of online news capturing the traditional news media ecosystem in a broad swathe of countries over more than a decade with unprecedented accuracy and granularity, resulting in a valuable new resource for research and practice with applications for media monitoring, crisis response, and program evaluation.

At the outset of INSPIRES, it was clear existing event data tracking civic space events would not be sufficient to accomplish the project's objectives. For most countries, civic space events are most frequently and reliably reported by traditional newspapers that publish their content online. While many citizens consume news through radio or social media, the underlying news circulated on these platforms often comes from articles originally published in online newspapers.

[1] A team at Google developed an architecture for a truly bi-directional embedding system and used it to train BERT. BERT was trained using the cloze task: for each sentence in its corpus, BERT would randomly mask some words and then try to predict what those words should be using the word that surrounded it. Iterating over every sentence in its (English-only) corpus multiple times, BERT learned how to use different contexts to maximize the accuracy of its predictions.

Existing event datasets – such as ACLED, GDELT, and ICEWS – collected large amounts of content published by online newspapers around the world to track politically important events. However, these sources did not produce data on critical dimensions of civic space, including political activism, government censorship, or the use of defamation to restrict civic space. For this reason, DevLab needed to develop new tools that could extract information on civic space events from text. Initially, DevLab considered applying these new tools to extract information on civic space events from the underlying collection of articles that existing event data produces had assembled. However, a careful investigation of the underlying data revealed that each existing collection had severe limitations.

ACLED relies on collecting text from online newspapers to produce data in violence and protests. However, they monitor sources by hand and only retain articles that are directly relevant to the specific events that they track. Furthermore, they do not collect data retrospectively; their repository of articles for a given country only begins from when they begin producing data for that country[^2](https://acleddata.com/acleddatanew/wp-content/uploads/dlm_uploads/2023/03/FAQs_ACLED-Sourcing-Methodology.pdf).

Other news scraping initiatives, such as GDELT, Common Crawl, and the Internet Archive rely on crawlers that 'crawl' websites by following links throughout the site and retrieving information from each different page that is detected. However, these tools often collect only a fraction of the total articles published by some sources. Furthermore, the lack of customized parsing means that crawlers often collect wildly inaccurate metadata on critical traits, such as the day, month, or even year that an article was first published.

Finally, another class of media monitoring event data sources, such as ICEWS and Temporally Extended, Regular, Reproducible International Event Records (TERRIER)—purchase licenses from data bundlers like Lexis Nexis. Purchasing news data has one main advantage: coverage. Older news stories are often unavailable on the open web and even those that are technically available are often difficult to find or otherwise unreachable through traditional scraping methods. However, licenses can easily run into the millions of dollars, placing them out of reach of most researchers and practitioners. Equally importantly, bundlers' coverage of each specific source depends on their licensing agreement with each newspaper. For many of the major local sources based in our sample of countries, these licenses only provide access to articles in recent years, i.e. well after 2012, which is the year our dataset begins.

To overcome the shortcomings of other approaches, DevLab developed a data collection system that captures the full publication history of local sources while ensuring accurate metadata. This process involves three distinct steps. First, we use publicly available information and feedback from partner organizations and USAID mission staff to identify high-quality, machine scrapable local online newspapers sources for each country. Second, we develop custom scrapers and parsers tailored to the unique architecture and publication practices of each website and bypass robot blockers (e.g., Cloudscraper). Third, we carefully evaluate performance after every update and adapt to changing website architecture over time and also replace sources in cases where publication ceases or continued scraping becomes impossible. Currently, we are scraping an average of 4.6 local sources for each country in our database and have a repository of about 100 million articles, 81% of which were published by local sources.

This repository of online news presents a completely novel source of data documenting the comprehensive publication history for multiple local news sources across each of 56 countries aid receiving countries. This dataset gives an unprecedented glimpse into the news media ecosystem in a broad swathe of countries over more than a decade and has many potential applications to event detection,

media monitoring, crisis response, and program evaluation. We also established internal processes to maintain and expand this dataset at a relatively low cost, providing an enduring source of data that can serve as a backbone for future USAID research initiatives.

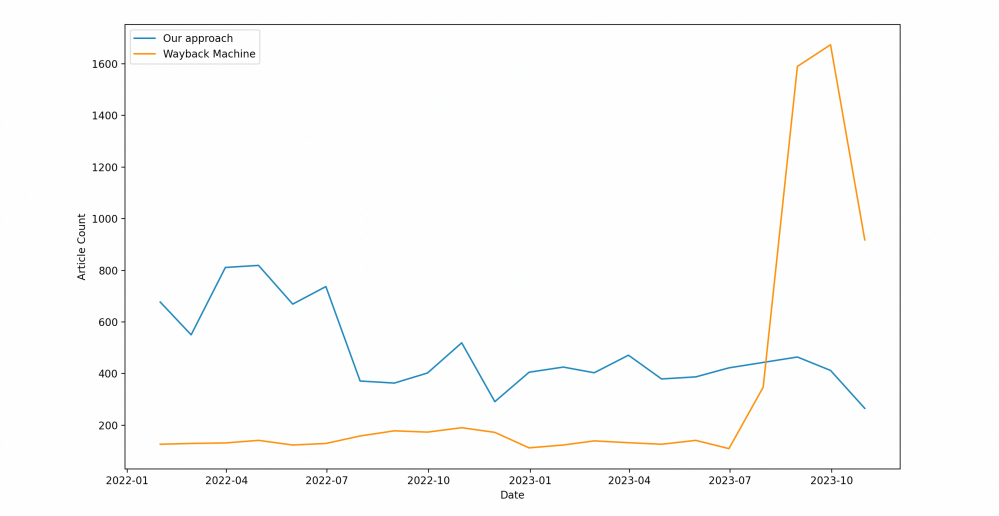


Figure 1: Check 1

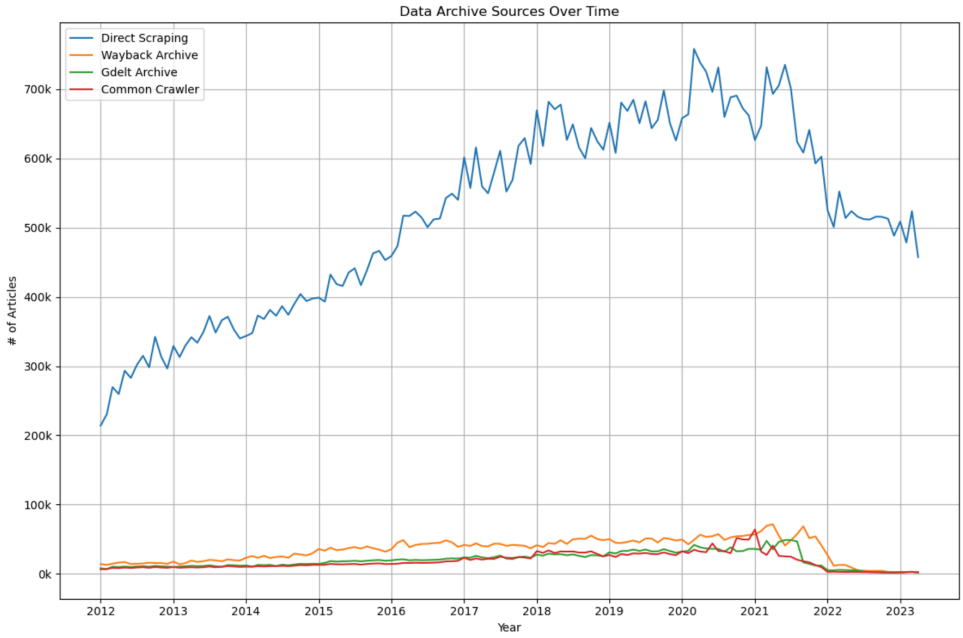


Figure 2: Check 2

Data processing

Once data is collected, the next step is extracting information about civic space and foreign influence events from the raw text data. 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 INSPIRES. The first training data for the civic space event counts covered 6475 (1,493 non-events and 4,982 events) articles over 20 event types. The second, RAI training data includes 3,400 articles over 22 event types. These models have overall out-of-sample accuracy close to 0.82 (civic space) and above 0.8 (RAI) on human-coded event data.

To further improve classification, some civic space event types use a keyword corpus to increase accuracy of classification. In addition, we trained 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.

In developing and deploying these models, DevLab demonstrated the potential for free, open-source LLMs to be fine-tuned to reliably classify reporting on civic space and foreign influence coming from a large number of countries and languages. By adapting this approach for both civic space and foreign influence events, MLP 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 MLP data repository. By incorporating these data processing techniques into a robust and highly flexible data processing pipeline, DevLab has 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.

4.2 Translation

Given the well-documented biases in English-language news sources but the prohibitive cost of translating millions of articles via Google Translate[2], we test the efficacy of translation models by extracting sample text from articles published in that 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 our other APIs and choose one that yields the optimal sentence-to-sentence translations with sufficient human readability[5].

4.3 Location

To get every piece of information needed for an event from a news story, MLP performs location extraction from the text to determine where the event occurred. MLP identifies entities using Named Entity Recognition (NER) that belong to a location or an organization using CLIFF,[6]

[2]Each country costs approximately $1,200 to translate the corpus of articles going back to 2012 via Google Translate.
[3]https://huggingface.co/
[4]We assess translation efficiency by examining the translated outputs of articles to see if the translations are reasonable.
[5]Although it is possible to classify events with multilanguage 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.
[6]For technical details on CLIFF, see: https://github.com/mediacloud/cliff-annotator

which relies on GeoNames (for state/city/town names) and extracts the location information.[7] CLIFF API has detailed information of the locations detected, and we retrieve and convert the country codes for further steps. If no country is found in the text, it will assign the article to the country of origin of the news source.

4.4. Classifier Performance

Perhaps the most important part of event extraction is event classification. In line with MLP's approach to flexibility, MLP is not packaged with a pre-built event ontology. Instead, it relies on the BERT model described above. Because the BERT project was open-sourced, others were able to take the work and use it to build even more powerful models. One of those is RoBERTa, the default model for MLP. RoBERTa uses a similar training process and slightly a different model architecture but even more data and larger batches. The team at Facebook that built RoBERTa broke most of the records that BERT set and then also released the model.

We fine-tune the RoBERTa model for our purposes by training and testing it on two corpora of human-coded newspaper articles hand built for INSPIRES. The first training data for the civic space event counts covered 6475 (1493 non-events and 4982 events) articles over 20 event types. The second, RAI training data includes 3,400 articles over 22 event types. See the Civic Space and RAI codebooks for details on the specific event types. 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. To further improve classification, some civic space event types use a keyword corpus to increase accuracy of classification. These keywords are used to ensure that articles that cover events that are similar to civic space event types, but are not directly related to civic space, are filtered out.[8]

We have previously reported on out-of-sample model performance for civic space and RAI events in the context of the training and test data that was specifically built for this project; the interested reader can refer to those earlier reports. Suffice to say that fine-tuning models with the default MLP settings produced models with overall out-of-sample accuracy close to 0.82 (civic space) and above 0.8 (RAI) 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.

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
Martial law / limits on gathering	0.92	0.90	0.91

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

[8] A list of keywords, as well as reasonings for their use, are provided in the appendix.

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
Mobilize security forces	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

To further demonstrate how accurate the system is, we trained a new classifier using ACLED data from between 2010 and 2019. Although the source-texts for the events are not publicly available, ACLED's coders provide a short summary of each event that closely resembles the first sentence of a news story. First, we trained the model with 100 examples from each of the six ACLED event categories and tested this model on an out-of-sample collection of 600 ACLED events (100 from each type), achieving overall accuracy and f1 scores (macro average) of 0.88. Using 1000 examples from each category for fine-tuning and the same 100-per-type test set, the models achieve accuracy and f1 scores of 0.94, which likely approaches or exceeds ACLED's human intercoder reliability scores. In short, our RoBERTA-based approach is very accurate.

In addition to our existing classification model, we have developed a capability to categorize events into two distinct groups: civic-related and non-civic events. 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 the pre-trained RoBERTA model. The classification of events into civic and non-civic categories is tailored to align with our specific criteria and the dimensions of civic space.

Sometimes news event 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 they have no or little bearing on civic space. In order to distinguish these types of non-civic events, we deploy a hand-trained civic/non-civic classifier that takes articles already coded as events and identifies whether they are related to broader civic space. This classifier is applied to the following event categories: arrests, cooperation, corruption, defamation cases, legal actions, legal changes, purges, raids, threats, lethal violence and non-lethal violence. We do retain the non-civic events in order to understand underlying news trends, but we do not use them for reporting on, or forecasting, civic space.

4.5 Civic Space and RAI Event Data

The classification model provides a 0/1 indication if an article qualifies as a civic space or RAI event type. To analyze this data, we sum these event counts across sources for each country by the month. This monthly event count data becomes the variables that we use to construct and forecast our civic space index (see below).

4.6 Validation of Scraping Coverage and Event Detection

To investigate the ability of our pipeline to capture the full universe of published articles from the sources identified in Appendix 1, we conducted a series of audits comparing articles published on these websites with the articles stored in our database. To do so, we first randomly selected two newspapers for each country and two months included in our time series. We then manually collected all articles published during those months and confirmed the presence of these articles in our database.

To test the ability of our system to reliably capture and identify politically important events, we conducted two further validation exercises. First, we and a team of researcher assistants went through an exhaustive validation process for 20 countries. There were three steps in this process: First and without first looking at the data, researchers gathered the dates of 'major events' for each country from 2012 onward using news aggregators, BBC timelines, Wikipedia, etc. Thereafter, we plotted the relevant event types in our data to ensure that we were properly capturing them. In some cases, news cycles follow events by a month, but our data showed spikes on and around these events in the vast majority of cases. Second, RAs were given 100 random stories per country and asked to hand code the articles and compare them against our machine coding. This process convinced us of the need to use translated text in the training phase, since our initial tests showed higher rates of misclassification for translated text, and we used this process to build the keyword lists describe above. Third and finally, researchers were assigned two national sources for each country and read and recorded every article and hyperlink on the website for a one-month sample. They then used the URLs or keywords from the titles to search for them in the database. Going in reverse order, this validation process provided a rigorous evaluation of: a) our scraping; b) our classification algorithm; and c) our capacity to identify major civic space events.

Second, we assessed the ability of MLP data to capture levels protest activity by looking at the correlation between our protest measure and that of ACLED. Specifically, we looked at whether small/large increases in our protest variable correspond with small/large events documented by ACLED. Using the full sample of countries and years common to both datasets, we aggregated ACLED to the country-month level and normalized both measures on a 0-1 scale to facilitate visual comparison. Before normalization, the ACLED variable was a discrete count while the MLP variable captured share of articles reporting on each type of event. For each country-year, we calculated the correlation between ACLED and MLP measures of protest activity. We then identified the 6 country-years with the strongest positive, lowest, and strongest negative correlation between these measures, extracted the underlying data used to generate these counts (for MLP this was the individual articles, while for ACLED this was the event-level data with event descriptions), and looked for systematic differences in (1) the type of protests being detected, (2) the way the datasets represent the timing or duration of protests, and (3) the way the datasets represent the magnitude or significance of protests.

First, we find that MLP and ACLED detect similar events. Overall, ACLED captured events that are not present in MLP more frequently than the reverse. This was especially true for very small protests that did not receive coverage in a country's national news. Second, MLP and ACLED capture the timing and duration of protests similarly. For the most significant and orchestrated protest events, MLP sometimes captures reporting on organizational planning ahead of these events, causing MLP to represent these protest events as beginning earlier than ACLED. Third, MLP better captures the importance of events by capturing the relative salience of these events as the extent to

which they dominant media coverage. This is due in-part to MLP's ability to incorporate popular commentary and dialogue on protest events as a measure of salience.

4.7 Normalization

As noted above, 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 produce less digital news as you move 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. Figure 3a provides an example from all of our sources in Serbia. The graph shows a steady climb in monthly articles from about 1,500 to over 5,000. Second, some sources seem to purge periods of their web archive for unknown reasons. Figure 3b provides an example of this from the Colombian source portafolio.co. In early 2016, the paper's archive drops to zero. Third and finally, some cases show discontinuous increases in the volume of news they report. Figure 3c shows the example of Ghana's largest online presence, ghanaweb.com, which goes from an average of about 500 articles a month to triple that in mid-2020. Again, the reasons are idiosyncratic ; in this case, the source won a large grant from Google to increase news coverage.

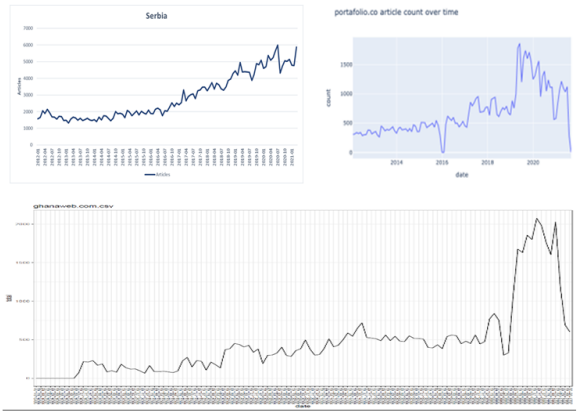


Figure 3: Examples of Source Availability through Time

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 events in the world or changes in the volume of availability of news. To address this, we divide each month-event count by the total number of articles to get a rate of signal to noise. 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.

4.8 Forecasting

Once event data is scraped, classified, aggregated, and validated, we sum the number of articles reporting on each event in every month and normalize this sum by the total number of published articles. These monthly event counts capture the share of total articles reporting on each event type on a monthly basis. This ratio tells us how frequently a given civic space category occurs relative to total scraped articles.

We also developed an algorithm that identifies major 'shocks' to civic space event categories by isolating months where the data shows a particularly large increase in the ratio of articles reporting on one of our event categories. Our peak detection method integrates a rolling window mechanism 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 indicators through normalized count spikes. By mitigating outlier impacts through winsorization, and applying context-sensitive decay and binning weights, our method adeptly captures initial event shocks. Post-optimization, statistical insights guide a neural network model, bolstered by definitive rules, to robustly and accurately detect societal event surges.

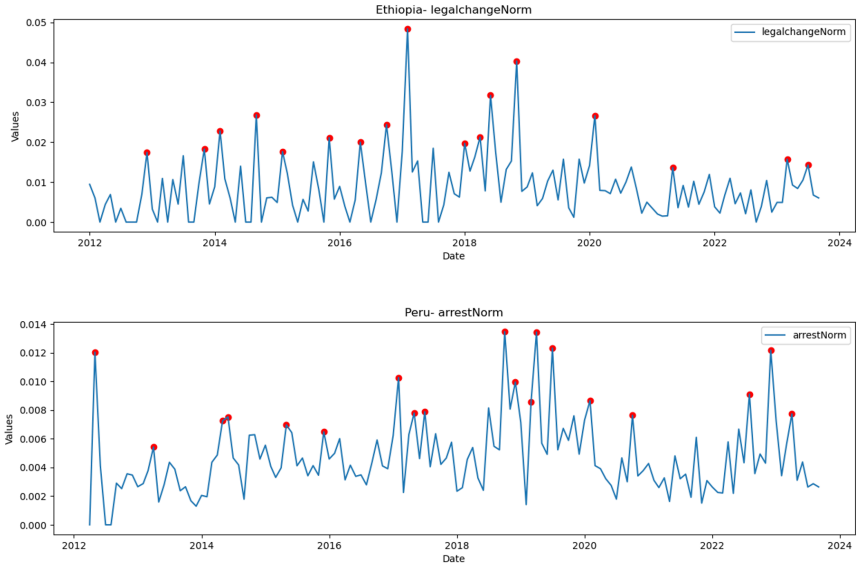


Figure 4: Peak detection

We merge these monthly event measures with high-frequency economic data from TradingEconomics and then use these data to identify historical patterns between economic, social, and political conditions and future shifts in civic space. We then train statistical models that identify patterns between past values of our 20 civic space and 22 authoritarian influence event variables, as well as a large sample of economic variables and current values of our Civic Space Index and 15 of our civic space event category variables. We train separate models to predict our 'standard' normalized measures and the 'shock' measures of civic space events.

Once these models have identified patterns that are consistent across many randomly drawn samples of our data, we feed the most recent values of our predictors to these models, which provide an

estimate of the expected value of the Civic Space Index and 15 civic space event measures 1 to 7 months into the future. Figure 5 presents the performance of these models using standard accuracy metrics. For some event categories, our models perform well across many countries and forecasts lengths, while for other categories, we see much lower performance scores. Unsurprisingly, models of civic space shocks or crises are, on average, less accurate. Yet, there are some event categories where our shock models perform well while our standard models perform relatively poorly, suggesting that taking a variety of modelling approaches has significant benefits overall.

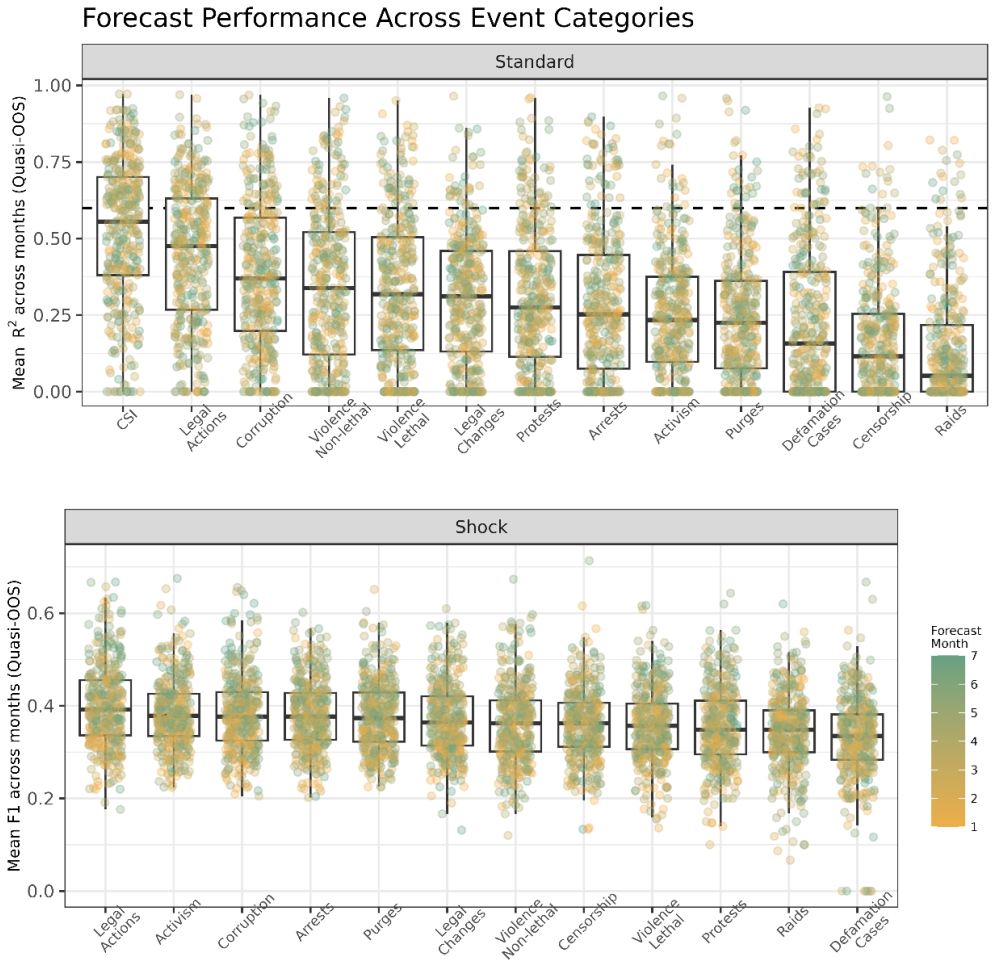


Figure 5: Each event category includes points representing seven models for each country. Performance is measured by each model's out-of-sample performance when using the optimal tuning parameter value. The dashed line in panel 1 indicates our minimum threshold to present the results of models in our regular update reports.

Figure 6 shows the variation in performance across countries. Again, there are several countries where the average performance across event categories and forecast lengths is high, while for others average performance is very low. As with variation across event categories, there are notable differences in the countries where our standard and shock models perform well, further emphasizing the importance of both measures. Figure 7 shows something similar, reporting the number of event categories where models perform about our minimum confidence thresholds. In our regular update reports, we only report predictions from models with performance scores above these thresholds.

Figure 8 shows the event category with the highest average performance for each country. For many countries, the best performing event category is different between the standard and shock models, further reinforcing the need for both models.

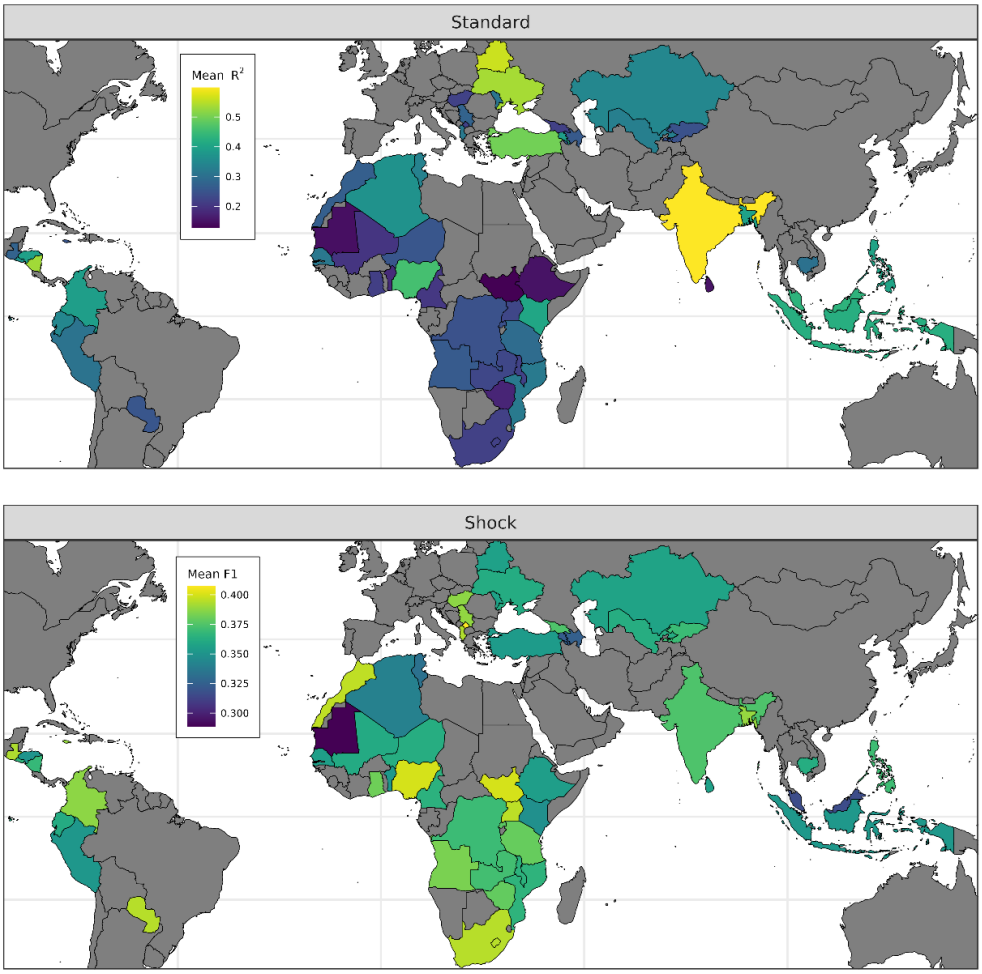


Figure 6: Average performance across event categories by country. Performance is measured by each model's out-of-sample performance when using the optimal tuning parameter value.

The results from our standard forecasting models are visualized on the MLP website via an interactive data dashboard. Users can view the historical values of each event category, the predicted future values, the average performance score across forecast lengths, and a list of the most influential predictor variables driving our model's forecasts. The results from both our standard and our shock forecasting models are also presented for non-technical audiences in the form of regular update reports that highlight any major predicted increases in activity across our event categories from our high performing models. These update reports are posted to our website every time we update the data for a batch of countries.

The figures above present standard statistical measures of predictive power for our forecasting models. However, these performance measures do not tell us about the track record of past predictions made by our models and communicated by DevLab to consumers. The DevLab team has made several decisions to guide when and how we communicate predictions to the public through

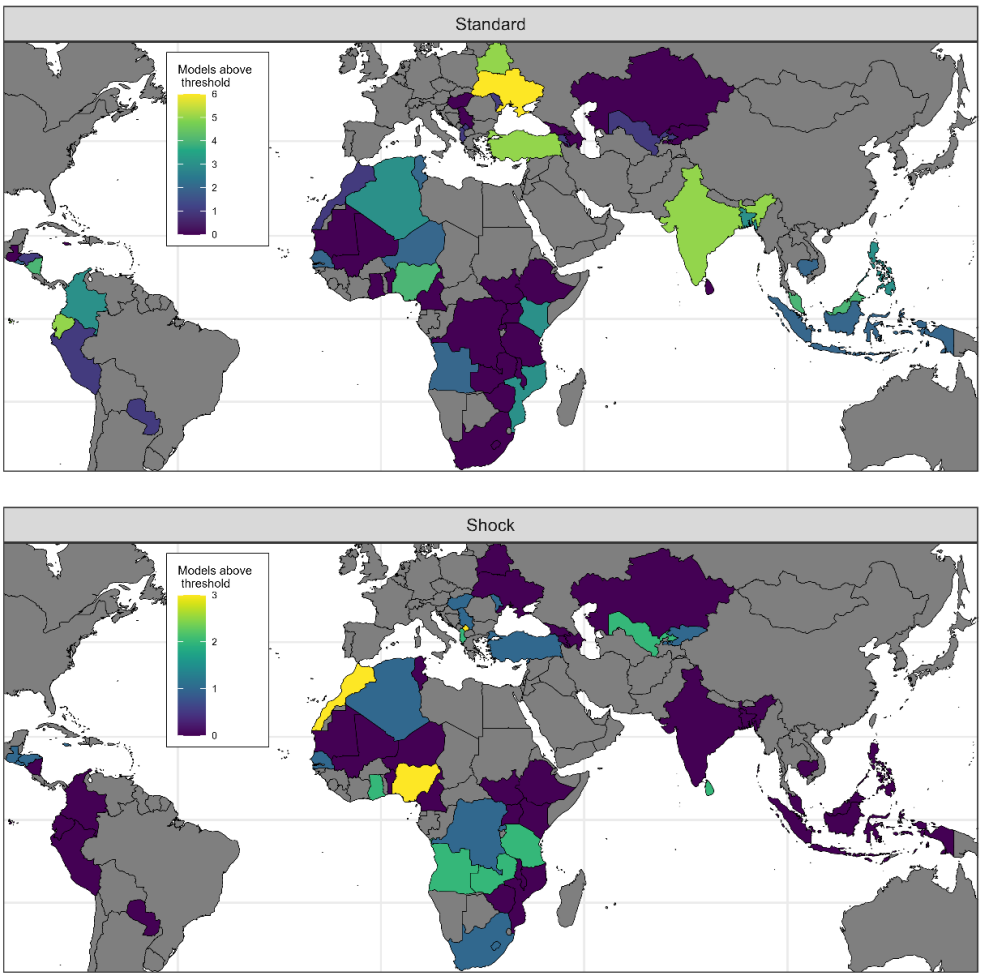


Figure 7: Count of event categories where models perform above reporting threshold by country. Performance is measured by each model's out-of-sample performance when using the optimal tuning parameter value.

our update reports. These include a minimum model performance threshold necessary to report a prediction and a minimum threshold for the size of the predicted change necessary to report a prediction. For this reason, the overall track record of our communicated predictions reflects both the performance of our models and the judgements of the research team.

To assess the overall track record of the predictions communicated in our update reports, we reviewed all predictions highlighted since we began publishing update reports in December 2022. For each prediction, we assessed whether the data from future updates aligned with the prediction. Table 2 summarizes the results of this exercise. We find that 71% of the predictions we communicated were consistent with changes in the data that we observed in future updates, while only 29% predicted changes in the data that did not happen in either the expected quarter or the quarter immediately before or after.

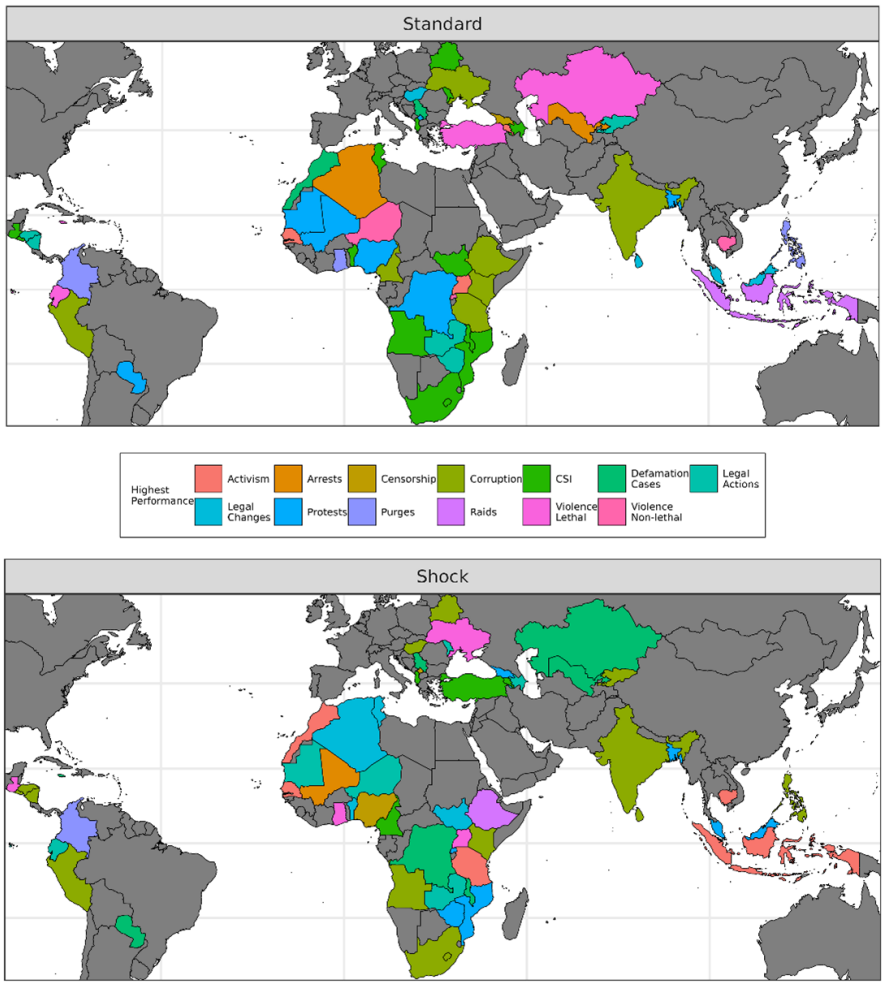


Figure 8: Event category with highest average performance by country. Performance is measured by each model's out-of-sample performance when using the optimal tuning parameter value.

Prediction occurred	Count	Share
Yes	25	48.1%
Early	9	17.3%
Late	3	5.8%
No	15	28.8%
Total	**52**	**100%**

Table 2: Prediction outcome frequency. Yes indicates that the prediction occurred within the expected quarter; Early and Late indicate that the prediction occurred in the quarter immediately before or after expected.

Figure 9 provides details on the performance of our communicated predictions across event categories. This table shows that our update reports have included accurate predictions across almost every event category that we produce forecasts for. We see that Arrests and Legal Actions were the

categories for which we communicated the most predictions. Importantly, these accurate predictions were spread across 20 different countries, with the large number of accurate forecasts coming from the Philippines, George, Indonesia, Turkey.

Event	Count: Yes	Count: No	Share: Yes
Arrests	7	3	70.0%
Legal Action	6	2	75.0%
Protest	4	0	100.0%
CSI	3	3	50.0%
Corruption	3	1	75.0%
Non-Lethal Violence	3	1	75.0%
Legal Change	2	1	66.7%
Lethal Violence	2	1	66.7%
Raids	2	1	66.7%
Censorship	1	0	100.0%
Civic Activism	1	1	50.0%
Defamation	1	0	100.0%
Legal Changes	1	0	100.0%
Purge/Replace	0	1	0.0%
Troop Mobilization	1	0	100.0%

Figure 9: Prediction outcomes across event categories. Yes indicates that the prediction occurred within the expected quarter or the quarter immediately before or after expected.

These results demonstrate that DevLab has developed and deployed forecasting models that use the MLP input data to produce accurate predictions about major civic space events. These models are incorporated into an automated pipeline that generates updated predictions for each update to the underlying data. Furthermore, the results of these models are made available to consumers through online data dashboards and summarized by the research team through regular update

reports. The regular publication of update reports describing our predictions since December 2022 has allowed us to monitor the track record of our communicated predictions, providing the basis to develop trust in our model output among practitioners and stakeholders, especially those with non-technical backgrounds.

5. Conclusion

As with any system created from news stories, MLP does have 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. 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 its limitations, MLP is a powerful and flexible tool for data generation. It demonstrates the potential for machine learning to improve the frequency and accuracy of quantitative data bearing on civic space and foreign involvement therein. It has built an article database measured in the millions and accurately classified the lion's share of them for subsequent modeling and forecasting.

We aspire to make this pipeline completely open source and thus completely adaptable to different projects, including those unrelated to civic space. Depending on researcher interests, a bespoke training dataset can be built within weeks for less than the price of an RA for a semester, there will be fewer and fewer excuses to rely on the same, slow-moving data given that the world is changing so quickly.

\newpage

Appendix 1: List of Digital News Sources Being Used by Country and 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, xinhuanet.com,

Sub-Saharan Africa:

- Africa Regional Sources:

africanews.com theeastafrican.co.ke iwpr.net

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

- Benin:
lanouvelletribune.info, news.acotonou.com,

- Cameroon:
journalducameroun.com, camerounweb.com, 237actu.com, 237online.com, cameroonvoice.com

- DR Congo:
radiookapi.net, lesoftonline.net, acpcongo.com, lephareonline.net, groupelavenir.org

- Ethiopia:
addisfortune.news, addisstandard.com, capitalethiopia.com, thereporterethiopia.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

- Mali:
maliweb.net, malijet.com, news.abamako.com,

- Malawi:
mwnation.com, nyasatimes.com, times.mw, faceofmalawi.com, malawivoice.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,

- Niger:
actuniger.com, nigerinter.com, lesahel.org, tamtaminfo.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

- South Africa:
timeslive.co.za, news24.com, dailysun.co.za, sowetanlive.co.za, isolezwe.co.za
iol.co.za, son.co.za,

- Tanzania:
ippmedia.com, dailynews.co.tz, habarileo.co.tz, thecitizen.co.tz, mtanzania.co.tz

- Uganda:
monitor.co.ug, observer.ug, newvision.co.ug, nilepost.co.ug,

- Zambia:
lusakatimes.com, mwebantu.com, diggers.news, openzambia.com, lusakavoice.com

- Zimbabwe:
thestandard.co.zw, theindependent.co.zw, herald.co.zw, chronicle.co.zw, newsday.co.zw

Middle East and North Africa

- Morocco:
leconomiste.com, lematin.ma, assabah.ma

- Tunisia:
assarih.com, babnet.net, jomhouria.com, lapresse.tn

- Turkey:
diken.com.tr, t24.com.tr, sozcu.com.tr, posta.com.tr, sabah.com.tr

Eastern Europe

- Eastern Europe Regional Sources:

euronews.com/tag/eastern-europe neweasterneurope.edu balkaninsight.com iwpr.net

- 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

- 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

- Macedonia:
koha.mk, slobodenpecat.mk, slobodenpecat.mk

- Moldova:
timpul.md, tribuna.md, unimedia.info, voceabasarabiei.md, publika.md

- Kosovo:
kosova-sot.info, balkaninsight.com, prishtinainsight.com, botasot.info

- Ukraine:
delo.ua, interfax.com.ua, kp.ua, pravda.com.ua, kyivpost.com, kyivindependent.com

- Serbia:
rs.n1info.com, juznevesti.com, insajder.net, danas.rs, balkaninsight.com

Latin America and the Caribbean:

- Latin America Regional Sources:

elpais.com cnnespanol.cnn.com iwpr.net

- Colombia:
elcolombiano.com, elespectador.com, elheraldo.co, eltiempo.com

- Ecuador:
elcomercio.com, eldiario.ec, elnorte.ec, eluniverso.com, metroecuador.com.ec

- El Salvador:
laprensagrafica.com, elfaro.net, elsalvador.com, diario.elmundo.sv

- 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

- Paraguay:
abc.com.py, lanacion.com.py, ultimahora.com

- Peru:
elcomercio.pe, gestion.pe, larepublica.pe, ojo-publico.com, idl-reporteros.pe

Asia:

- Asia Regional Sources:

asiatimes.com asia.nikkei.com iwpr.net

- Bangladesh:
prothomalo.com, bd-pratidin.com, kalerkantho.com, jugantor.com, dailyjanakantha.com

- Cambodia:
kohsantepheapdaily.com.kh, moneaksekar.com, phnompenhpost.com

- Indonesia:
thejakartapost.com, jawapos.com, kompas.com, mediaindonesia.com, sindonews.com, beritasatu.com, hariansib.com

- India:
amarujala.com, indianexpress.com, thehindu.com, hindustantimes.com, deccanherald.com, firstpost.com, indiatimes.com, timesofindia.indiatimes.com

- Kazakhstan:
caravan.kz, diapazon.kz, kaztag.kz, rus.azattyq.org

- Kyrgyzstan:
akipress.com, 24.kg, kloop.kg, super.kg, vb.kg,
kaktus.kg, kaktus.media, rus.azattyq.org

- Malaysia:
malaymail.com, nst.com.my, thestar.com.my, utusan.com.my

- Philippines:
mb.com.ph, manilastandard.net, inquirer.net, manilatimes.net

- Sri Lanka:
dailymirror.lk, island.lk, divaina.lk, adaderana.lk, lankadeepa.lk

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

Appendix 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.

Civic Space Keywords

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.

RAI Keywords

RAI keywords are used to narrow down the focus of influence to the spheres of China and Russia.

- **For Title**: China, Chinease, Russia, Russian
- **For Title and Main text:** (This list includes businesses, government agencies, and programs associated with these two countries)
- Gazprom; Lukoil; Rosneft; Sberbank; Russian Railways; Rostec; VTB; X5 Retail Group; Surgutneftegas; Magnit; Rosseti; Inter RAO; Transneft; Rosatom; Sistema; Tatneft; Gazprombank; Evraz; NLMK; Novatek; Sibur; Rusal; Norilsk Nickel; Aeroflot; Severstal; United Aircraft Corporation; Mobile TeleSystems; Magnitogorsk Iron and Steel Works; Ural Mining and Metallurgical Company; RusHydro; MegaFon; Lenta; Metalloinvest; Stroygazmontazh; T Plus; VimpelCom; Siberian Coal Energy Company; United Shipbuilding; Sakhalin Energy; Rostelecom; Alfa-Bank; Otkritie Holding; Mechel; Vnesheconombank; DIXY; Alrosa; Rosselkhozbank; Protek; OAO TMK; Russian Helicopters; United Engine Corporation; Metro Cash & Carry; Leroy Merlin Vostok; AvtoVAZ; Merlion; Avtotor; Tactical Missiles Corporation; Red&White; Mostotrest; M.video; PhosAgro; Rolf Group; PIK Group; Russian Post; Nizhnekamskneftekhim; GAZ Group; Tashir; Uralkali; Polyus; Euroset; Chelyabinsk Pipe Rolling Plant; Sodrugestvo; SOGAZ; KamAZ; Transmashholding; StroyTransNefteGaz; Zarubezhneft; Arktikgaz; UCL Holding; Credit Bank of Moscow; LSR Group; ForteInvest; Irkutsk; Uralvagonzavod; RussNeft; Putin; Sergey Lavrov; Moscow; Rusia; Rusa; Ruso; Acron; Moran Security Group; PMC Shchit; Rossotrudnichestvo; Roszarubezhneft; RSB-group; Russkiy Mir Foundation; Skolkovo Foundation; Wagner Group; Yota; Confucius institute; Asian Infrastructure Investment Bank; 361 Degrees; Agricultural Bank of China; Aigo; Air China; Alibaba; Aluminum Corporation of China Limited; Amoi; Anta Sports; Baidu; Bank of China; Bank of Communications; China Baowu Steel Group; Beijing Hualian Group; Bolisi; Bosideng; Brilliance Auto; BYD Auto; Changan Automobile; Changhe; Changhong; Changhong Technology; Chery Automobile; China Clean Energy; China Communications Construction; China Construction Bank; China COSCO Shipping; China Dongxiang; China Eastern Airlines; China Housing and Land Development; China International Marine Containers; China Life Insurance Company; China Medical Technologies; China Merchants Bank; China Merchants Energy Shipping; China Metal Recycling; China Mobile; China National Erzhong Group; China National Offshore Oil Corporation; China National Petroleum Corporation; China Natural Gas; China Nepstar; China Netcom; China Pabst Blue Ribbon; China Post; China Shipping Group; China Southern Airlines; China State Construction Engineering; China Telecom; China Three Gorges Corporation; China Tobacco; China Unicom; China Universal; China Wu Yi; China Zhongwang; Chunlan Group; CITIC Group; CNHLS; Comac; Commercial Press; COSCO; Dalian Hi-Think Computer; Dashang Group; Dayun Group; Dicos; Dongfeng Motor Corporation; DXY.cn; Eisoo; Eno; ERKE; FAW Group; Feicheng Acid Chemicals; Feiyue; Founder Group; Fushi Copperweld; GAC Group; Geely; Gome; Great Leap Brewing; Gree Electric; GreenTree Inns; Guangzhou Zhujiang Brewery Group; Gushan Environmental Energy; Hafei; Haier; Hainan Airlines; Hangzhou Wahaha Group; Harbin Brewery; Hasee; Hefei Meiling; Hisense; HiSilicon; Huawei; Huayi Brothers Media Corporation; Huiyuan Juice; Hytera; Industrial and Commercial Bank of China; Inspur; JDB Group; Jiangling Motors; Jianlibao Group; Jiuguang Department Store; Joyoung; JXD; Kingsoft; Kingway Brewery; Lenovo; Li-Ning; Little Sheep Group; Loncin Holdings; Lonking; Mailman Group; Maoye International; Meizu; Mengniu Dairy; Meters/bonwe; Midea Group; Mingyang Wind Power; Miniso; Mr. Lee; Nanjing Automobile; Neusoft; Ningbo Bird; Opple Lighting; Panda Electronics; Peak Sport Products; Pearl River Piano Group; People's Insurance Company of China; Ping An Bank; Ping An Insurance; Qihoo 360; Qinghai Huading Industrial; SAIC-GM; SAIC Motor; Sany;

Septwolves; Shaanxi Automobile Group; Shaanxi Yanchang Petroleum; Shanghai Film Group Corporation; Shanghai Pudong Development Bank; Shenyang Aircraft Corporation; Shenzhen Airlines; Shenzhen Energy; Shenzhen Media Group; Shougang; Shui On Land; Sichuan Airlines; Simcere Pharmaceutical; Sinoenergy; Sinopec; Sinopharm Group; Sinosteel; Sinovac Biotech; Skyworth; SmithStreetSolutions; State Grid Corporation of China; Suning Commerce Group; Suntech Power; Suzhou Synta Optical Technology; TCL Corporation; Telesail Technology; Tencent; Tianan Insurance; Tianjin FAW; Tongrentang; Topray Solar; TP-Link; Trands; Tsingtao Brewery; Vanke; Vinda International; Vsun; Wanda Group; WuXi PharmaTech; Xi'an Aircraft Industrial Corporation; Xiaomi; Yili Group; Yonyou; Yutong Group; Zhongjin Gold; Zhongjin Lingnan; Zoomlion; ZTE; ZX Auto; Xi Jinping; Hu Jintao; Beijing; Shanghai; BeiDou; BGI; ByteDance; China Electronics Technology Group; CloudWalk; Dahua; DJI; Hikvision; iFlytek; Megvii; Meiya Pico; SenseTime; Uniview; WuXi AppTec Group; YITU