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

Serkant Adiguzel, Akanksha Bhattacharyya, Spencer Dorsey, Tim McDade, Zung-Ru Lin, Donald Moratz, Diego Romero, Jeremy Springman, Hanling Su, Erik Wibbels

November 28, 2022

1. The Machine Learning for Peace (MLP) Pipeline

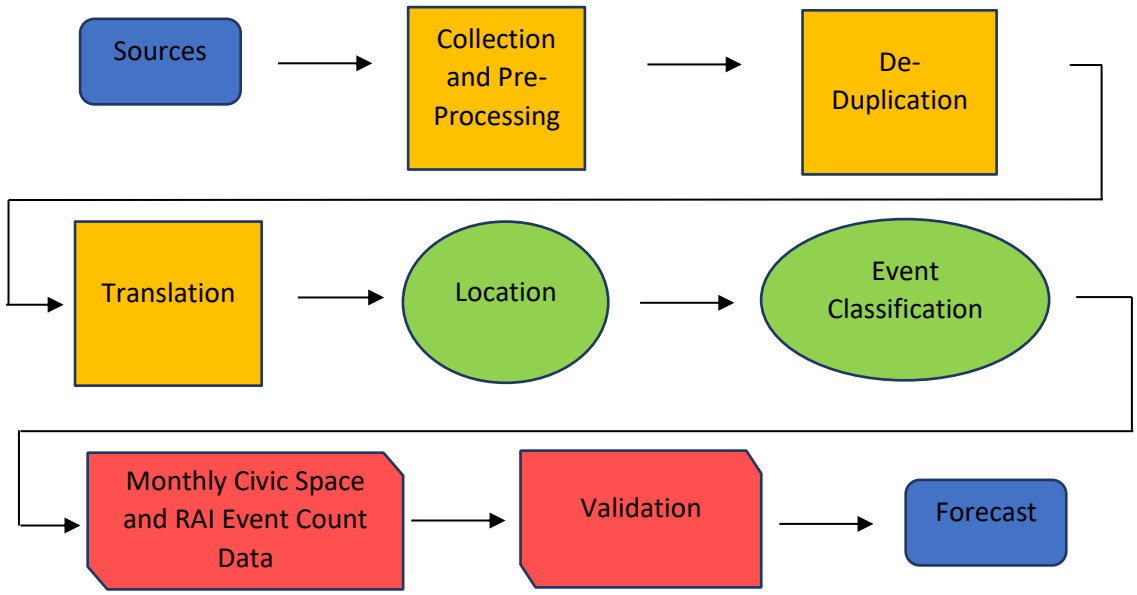
Civic space is expanding and contracting on a daily basis in countries around the world. Social movements and aspiring autocrats alike frequently seize on sudden economic, humanitarian, and political crises to contest the extent of fundamental rights and freedoms that underpin democratic accountability. 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 19 types of events relevant to civic space and 23 events capturing 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

The first step in creating an event dataset is securing news articles to parse and extract. Some event datasets—such as ICEWS and Temporally Extended, Regular, Reproducible International Event Records (TERRIER)—purchase licenses from data bundlers like Lexis Nexis. Others, like GDELT, rely on web scraping to collect stories published on news websites and collected by archivers like Common Crawl and the Internet Archive. 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 for news stories can easily run into the millions of dollars (placing them out of reach of most researchers and practitioners).

[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.

Due to the prohibitive costs of data licensing, we scrape news websites and archiving services. We begin with source domains such as nytimes.com.[2] These sources tend to have extensive archives, reliable publishing habits, and thereby eliminate the need for extensive human oversight of the scraping. To identify national sources, we aim to identify the top 3-5 newspapers by circulation for each country. We do this by consulting international or regional lists of newspapers maintained by university library guides, Wikipedia pages, and others. We also consult our partner organizations and USAID mission staff to ensure we do not overlook high-quality sources. We then check each source to ensure that it publishes original content, is machine scrapable, has a sufficiently far-reaching historical archive (with frequent publications dating back to at least 2015), and is published in a translatable language (see below on translation). We then conduct an online search to identify whether the source has a clear bias (such being state-run) and record that information. Finally, given that: a) some national news media markets are shallow and/or sources are impossible to scrape and/or politically compromised and b) international news media cover some countries very rarely, we supplement international and national sources with 2-3 regional sources.[3] In identifying these sources, we follow the same rules as above.

Using the list of international, regional and national domains, MLP relies on custom scrapers written to accommodate a wide variety of website architectures and bypass robot blockers (e.g., cloudscraper) whenever feasible. Depending on website architecture we obtain news articles by either scraping sitemaps, and newspaper archives, or simulating infinite clicking/scrolling using selenium. In some cases, we rely on GDELT, the Internet Archive, Common Crawl and/or from the websites directly to complement the output of our custom scrapers.[4] The parsed stories collected directly from the website or indirectly from GDELT and the Internet archive are then processed through a slightly modified version of the news article extraction system "news-please" to extract the publishing date, title, and story text from each article. This data is stored in MongoDB.

A common challenge is that news sources, say the AP, write an article that is syndicated across a wide range of news outlets. This results in the same story appearing in many publications. In order to avoid the error inherent in coding the same article many times as many separate civic space events, we de-duplicated on the basis of URLs and titles of articles posted on the same day. We also need to remove certain articles which contain news about events that do not contribute to civic space (like sports, celebrity/tv show news). Thus, we do pre-processing and cleaning on the articles extracted and include the ones which will go forward into the rest of the pipeline.

Another common challenge is the irregular availability of articles for a given source; sources might go from producing hundreds or thousands of articles per month to none or very few. The reasons behind this problem vary and are idiosyncratic—they range from changes or irregular maintenance of the website's archive to the lack of a proper site map to changes over time the location of the main text

[2] Our international sources are aljazeera.com, bbc.com, csmonitor.com, france24.com, nytimes.com, reuters.com, scmp.com, theguardian.com, themoscowtimes.com, washingtonpost.com, wsj.com and xinhuanet.com

[3] For example, for countries in Sub-Saharan Africa we scrape stories from The Africa Report, Africa News, and All Africa.

[4] Before 2021 we relied largely on Scrapy spiders to recursively scrape all available pages from the target domains (from sitemaps when available). Spiders allow for a customized set of rules and procedures to be defined for scraping specific websites or groups of sites. These tools 'crawl' websites by following links throughout the site and retrieving information from each different page that is detected. Our current method based on custom scrapers is much more efficient and allows us to quickly update each source at whichever interval we need.

within the code of the article. Custom scrapers allow us to get all articles published by a given source, especially for the most recent months. However, we cannot do anything in cases were newspapers themselves post fewer articles online.

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,[5] 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.[6] We then assess whether the translations are sufficiently comprehensible that they produce classifiable results.[7] 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.[8]

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,[9] which relies on GeoNames (for state/city/town names) and extracts the location information.[10] 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. Event Classification and 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

---

[5] Each country costs approximately $1,200 to translate the corpus of articles going back to 2012 via Google Translate.

[6] https://huggingface.co/

[7] We assess translation efficiency by examining the translated outputs of around 5 articles to see if the translations are reasonable.

[8] 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

[9] For technical details on CLIFF, see: https://github.com/mediacloud/cliff-annotator

[10] 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/

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.[11]

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
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

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

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.

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 and Event Data

We test the validity of the event data to ensure it properly reflects events in a country and what appears in the newspapers. To do so we rely on a team trained to conduct 3 different data quality control exercises: (i) an audit of event classification, and (ii) audit of the event counts generated, and (iii) an audit of the news stories collected.

The first task consists of manually assessing the accuracy of the event classifier for each country within a given time frame. To conduct this task, our team checks a random sample of at least 100 classified events and assesses whether they have been correctly classified. This exercise helps ensure the overall quality of our event count data. The second task consists of checking that our event count data reflects major civic space-related events, rather than simply fluctuations in the volume of news we scrape.[12] As an example, suppose that country A experiences a coup d'état. We would expect this event to increase event counts for several event types, ranging from 'change in power' and 'protest' to, potentially, lethal and nonlethal violence, depending on the nature of the coup itself. The task of our team is to check whether event counts for those event types do, in fact, increase in and after the month when the coup took place. For the third and final standard data quality control task our team: first, checks whether articles covering each one of the identified major events have been scraped, correctly classified, and stored in our database; and second, chooses two newspapers at random from a given country, manually collects all articles that cover civic space and RAI events published during a randomly chosen period of two months, and checks whether these articles have, in fact, been scraped.

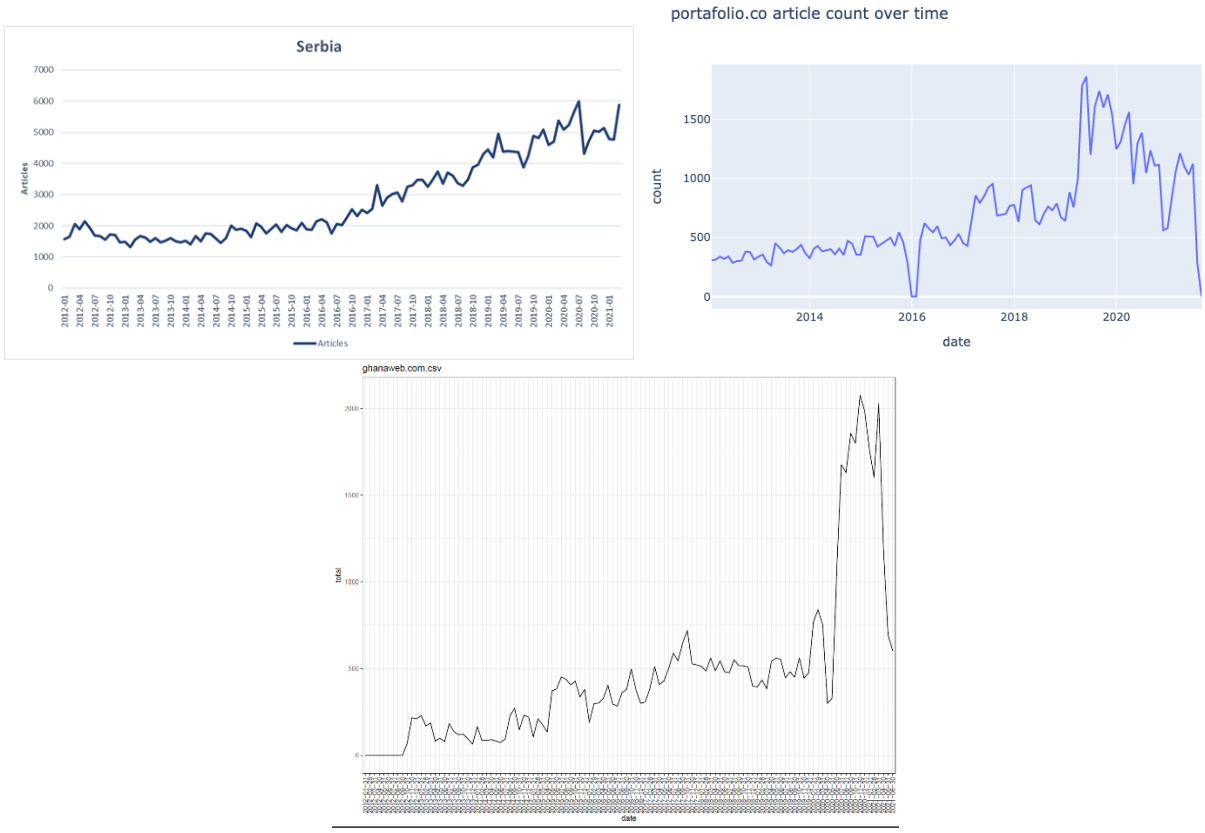
4.7 Normalization

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 1a 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.

[12] This is important for identifying news sources that might dump articles irregularly, suddenly change their web architecture, etc. any of which can impact event counts without reflecting actually goings-on in the world.

Figure 1b 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 3b 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 1a-c: 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. To be clear, this does not yield a count of the number of distinct events. If a single event, by virtue of its importance, induces a great deal of distinct articles, each of those articles are

counted. We do not attempt to distinguish the number of distinct events under each category. This approach has the advantage of providing information about the significance of events by capturing the volume of news dedicated to them.

We merge these monthly event counts 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.[13] To measure shifts in civic space, we combine 16 of our 20 total civic space event categories into a single index variable, the Civic Space Index[14], that summarizes monthly variation across the entirety of our civic space events.[15] We then estimate statistical models that use machine learning to identify patterns between past values of our 20 normalized civic space and 22 authoritarian influence event variables, as well as a large sample of economic variables (we refer to these variables as potential predictors) and current values of our Civic Space Index as well as 15 of our civic space event category variables. Specifically, these models look for correlations between the value of our Civic Space Index in a given month and values of our predictor variables between *h* and 12 months prior to that given month (*h* is set to the number of months into the future being forecasted).

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 1 to 7months into the future. We measure the accuracy of these models to predict Civic Space Index using prediction intervals.[16] These intervals are reported as the 80th percentile of expected outcomes that result from the cross-validation folds in our models. These prediction intervals can be interpreted as the stability of output of the model, where the larger the

[13] For each country, we utilize every variable available that meet two criteria: the variable is updated at least quarterly (frequency) and we observe at least as many unique values as there are years in the data (variation). Our most common economic variables are Utility CPI, Foreign Exchange Reserves, Transportation CPI, Imports, Inflation Rate Month-over-Month, Food Price Inflation, Interest Rates, Exports, Business Confidence, Crude Oil Production, and others.

[14] The Civic Space Index is constructed for each country using Principal Component Analysis (PCA). Specifically, our index is the score on the first principal component.

[15] We drop the Election Activity, Election Irregularities, Political Cooperation, and Disasters variables from this Civic Space Index. We do this for several reasons. First, the majority of Disasters in our dataset are natural disasters, which we are not interested in trying to predict. Second, Election Activity is defined as reporting on regular electoral activities. Because this variable is highly correlated with the timing of regularly scheduled elections, we do not see it as an important component of civic space that we are interested in predicting. Finally, Cooperate frequently captures a broad range of positive statements between important individuals or organizations, and rarely captures instances of genuine cooperation. For this reason, Cooperate provides a helpful measure of the frequency of positive statements in public discourse, but it does not indicate material accomplishments that we are interested in predicting. We consider all these variables to be potentially important predictors of changes in civic space, but we exclude them from the Civic Space Index because we are not interested in predicting their occurrence.

[16] To implement forecasts, we use a regularized linear regression model, known as elastic net (EN), with leave-one-out cross validation. EN is designed to guard against overfitting when using a large number of predictor variables by identifying the subset of variables that are most consistently predictive across different subsets of the data and selecting only those variables to use in a predictive model. To ensure stability and optimize lambda selection, we utilize leave-one-out cross validation. Leave one out sets k=n, where n is the number of observations. This cross-validation uses n-1 observations, ran n times to test lambda selection to select the optimal risk consistent lambda (Homrighausen and McDonald, 2012). These models are implemented using the glmnet package in R. In addition to h month lags of predictor variables, we also include 12-month lags to capture potential seasonality.

range of possible outcomes, the less confident we are in the final value (Steinberger and Leeb, 2016). In addition to the forecast of the Civic Space Index, we provide similar forecasts for the normalized values of Arrests, Censorship, Civic Action, Corruption, Defamation Cases, Legal Action, Legal Change, Protests, Purges, Raids, Lethal Violence, and Non-Lethal Violence. These forecasts are then presented to practitioners in the form of regular, non-technical charts on our website that apply this learning to the most recently available data to make predictions about civic space activity in the coming months seven months.

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.

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, lemonade.fr, liberation.fr and lefigaro.fr

Sub-Saharan Africa:

- **Africa Regional Sources:**
  africanews.com, theeastafrican.co.ke, iwpr.net
- **Angola:**
  Jornaldeangola.ao, opais.co.ao, jornalf8.net, angola24horas.com, portaldeangola.com, angola-online.net, vozdeangola.com
- **Benin:**
  lanouvelletribune.info, news.acotonou.com
- **Cameroon:**
  Journalducameroun.com, camerounweb.com, 237actu.com, 237online.com, cameroonvoice.com
- **Democratic Republic of the Congo**
  lesoftonline.net, acpcongo.com, radiookapi.net, lephareonline.net
- **Ethiopia:**
  addisstandard.com, addisfortune.news, capitalethiopia.com
- **Ghana**
  dailyguidenetwork.com, ghanaweb.com, graphic.com.gh, newsghana.com.gh
- **Kenya:**
  kbc.co.ke, citizentv.digital, theeastafrican.co.ke, nation.africa
- **Mali:**
  malijet.com, maliweb.net, news.abamako.com
- **Malawi:**
  mwnation.com, nyasatimes.com, times.mw, faceofmalawi.com, malawivoice.com
- **Mauritania**
  alwiam.info, lecalame.info, journaltahalil.com
- **Mozambique**
  Correiodabeiraserra.com, canal.co.mz, infromoz.com, mmo.co.mz, cartamz.com, verdade.co.mz, clubofmozambique.com, portalmoznews.com, jornaldomingo.co.mz, tvm.co.mz
- **Niger:**
  actuniger.com, nigerinter.com, tamtaminfo.com, lesahel.org
- **Nigeria:**
  guardian.ng, thenewsnigeria.com.ng, vanguardngr.com
- **Rwanda:**
  newtimes.co.rw, therwandan.com, kigalitoday.com, umuseke.rw
- **Senegal:**
  ferloo.com, xalimasn.com, lesoleil.sn, enqueteplus.com, lasnews.sn
- **South Africa:**
  Timeslive.co.za, news24.com, dailysun.co.za, sowetanlive.co.za, isolezwe.co.za, iol.co.za, son.co.za

- **Tanzania:**
  thecitizen.co.tz, ippmedia.com, mtanzania.co.tz, dailynews.co.tz, habarileo.co.tz
- **Uganda:**
  monitor.co.ug, observer.ug
- **Zambia**
  lusakatimes.com, mwebantu.com
- **Zimbabwe:**
  thestandard.co.zw, theindependent.co.zw, herald.co.zw, chronicle.co.zw

Middle East and North Africa

- **Morocco:**
  leconomiste.com, lematin.ma, assabah.ma
- **Tunisia:**
  lapresse.tn, assarih.com, babnet.net, jomhouria.com
- **Turkey**
  sozcu.com.tr, posta.com.tr, diken.com.tr, t24.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
- **Georgia:**
  ambebi.ge, georgiatoday.ge
- **Hungary:**
  index.hu, 24.hu, 168.hu, hvg.hu, demokrata.hu
- **Kosovo:**
  kosova-sot.info
- **Ukraine:**
  delo.ua, interfax.com.ua, kp.ua, pravda.com.ua, 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, diario.elmundo.sv, elfaro.net, elsalvador.com

- **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, ultimahora.com, lanacion.com.py

Asia

- **Asia Regional Sources:**
  timesofindia.indiatimes.com, asiatimes.com, asia.nikkei.com, iwpr.net
- **Bangladesh:**
  bd-pratidin.com, prothomalo.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, jagran.com, thehindu.com, bhaskar.com, hindustantimes.com, deccanherald.com, firstpost.com
- **Malaysia:**
  thestar.com.my, malaymail.com, 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, betafsil.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
    - Keywords
        - Keyword list one: case|lawsuit|sue|suit|trial
        - 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
    - Keywords
        - 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
    - Keywords

        - 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
    - Keywords
        - For arrest: arrest; detain*
        - For legal action: legal process; case*; investigat*; appeal; charged; prosecut*
        - For purge: resign*; fire*
    - Purpose
        - 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. 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; Russia; Russian; 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; China; Chinese