Measuring Human Rights Abuse from Access to Information Requests

Sarah A. V. Ellington,[*] Benjamin E. Bagozzi,[†] Daniel Berliner,[‡]
Brian Palmer-Rubin,[§] and Aaron Erlich[¶]

October 6, 2021

Abstract

Existing measures of human rights abuses are often only available at the country-year level. Several more fine-grained measures exhibit spatio-temporal inaccuracies or reporting biases due to the primary sources upon which they rely. To address these challenges, and to increase the diversity of available human rights measures more generally, this study provides the first quantitative effort to measure human rights abuses from textual records of citizen-government interactions. Using a dataset encompassing over 1.5 million access to information (ATI) requests made to the Mexican federal government from June 2003 onward, supervised classification is used to identify the subset of these requests that pertain to human rights abuses of various types. The results from this supervised machine learning exercise are validated against (i) gold standard ATI requests pertaining to past human rights abuses in Mexico and (ii) several accepted external measures of sub-national and sub-annual human rights abuses. In doing so, we demonstrate that the measurement of human rights abuses from ATI request texts can provide measures of human rights abuse that exhibit both high validity *and* notable spatio-temporal specificity, relative to existent human rights datasets and variables.

[*] Department of Political Science & International Relations, University of Delaware. Email: saelling@udel.edu.

[†] Department of Political Science & International Relations, University of Delaware. Email: bagozzib@udel.edu.

[‡] Department of Government, London School of Economics. Email: danberliner@gmail.com.

[§] Department of Political Science, Marquette University. Email: brian.palmer-rubin@marquette.edu.

[¶] Department of Political Science, McGill University. Email: aaron.erlich@mcgill.ca.

Introduction

The *measurement* of human rights abuse is of central relevance to the study of human rights. Three of the most well-established country-year measures of human rights abuses produced this decade—the CIRI Human Rights Dataset (Cingranelli and Richards, 2010; Cingranelli, Richards and Clay, 2014), the Political Terror Scale (PTS; Wood and Gibney, 2010), and the Latent Human Rights Protection Scores (Fariss, 2014)—have now collectively received over 2,000 citations.[1] Adding additional nuance to these datasets necessitates measurements of human rights abuses at spatio-temporal scales that are more precise than the country-year unit (Cordell et al., 2019*b*). Such data would allow for more fine-grained tests of the determinants of human rights abuses, and for further synergy with theories and models of political violence—which have increasingly shifted towards sub-national and sub-annual data over the past decade (Cederman and Gleditsch, 2009; Raleigh et al., 2010; Gleditsch, Metternich and Ruggeri, 2014). These advancements would also offer scholars and advocacy groups (i) the ability to better detect and preempt human rights abuses before they spread and (ii) improved understandings of the microfoundations of human rights abuses.

In light of these evolving measurement needs, this paper considers the use of textual records of citizen-government interactions for the development of new fine-grained spatio-temporal data on human rights abuses. Such information and communications technology (ICT)-enabled platforms include citizen reporting initiatives, complaint mechanisms, official social media accounts, and our focus here: access-to-information (ATI) requests. Across the world, country-specific ICT platforms increasingly make large-scale textual records of past citizen-government interactions available online, thus offering researchers and practitioners new opportunities to measure real-world outcomes at a fine-grained level.

We specifically consider data from one ATI regime for which we have access to comprehensive records of every single individual request for government information: the case of the Mexican federal government. Following Mexico's landmark 2002 ATI law, all individual ATI

[1] Based upon Google Scholar citations for the datasets and articles cited here as of 2/25/2020.

requests filed with Mexican federal government agencies are publicly available. Additional requests made to other federal branches of government and constitutionally autonomous bodies were added to this publicly accessible system in 2016. Each individual Mexican ATI request includes the textual description of the information that a citizen or organization seeks, supplemental attachments, textual entries that contextualize the requested information, the requester's municipality, the date of the request, the government entity to which the request is directed, and information the Mexican government's response to each request.[2] We contend that measuring domestic human rights abuse concerns from the textual content of ATI requests will provide valuable and unique information on human rights abuses. For a given country of interest, such human rights abuse data are available at far more fine grained spatio-temporal scales than (NGO or government-generated) country-year reports on human rights—and the current standards-based human rights measures derived from the latter reports. At the same time, ATI-derived measures of human rights abuse are also likely to be less sensitive to many of the media reporting biases that are commonly associated with international news(wire) reporting and media-derived event data.

Measuring human rights abuses from ATI requests accordingly offers several potential broader benefits to the study of human rights. First, as alluded to above, ATI requests are likely to capture a wider range of perceived human rights abuses than those reported in NGO or media sources. Relative to the latter two sources, ATI requests are more plentiful, more spatio-temporally precise, more directly accessible to those experiencing or witnessing abuses first hand, and far less constrained by space or audience considerations. Hence, while we do not argue that ATI requests will provide a comprehensive view of domestic human rights abuses on a global or even regional scale, the data coded from such requests are likely to capture many *individual* human rights abuses that do not make it into NGO reports or daily news reports—nor into the quantitative country-year and/or event data measures derived from these latter reports. These qualities may in turn allow researchers to use ATI-derived

[2] Including the date of the government response, the government's official response—such as, for example, "information provided," "information does not exist," or "the information is (partially) classified"—and the actual information provided in the government's response.

measures of human rights abuse to improve (or validate) existing quantitative human rights measures at the measurement (Fariss, 2014) and/or analysis stage (Bagozzi et al., 2019).

Second, while our focus is on producing *subnational* and *subannual* indicators of human rights abuse, an ATI-based human rights measure could also be leveraged and analyzed in a fully disaggregated manner. That is, such a measure could be evaluated at the *individual request* level, allowing one to examine (e.g.,) the characteristics of individual human rights-based ATI requests or of government responsiveness to individual human rights requests. As such, the measurement of human rights abuse from ATI request texts will facilitate even more fine-grained analyses of human rights processes than are currently available—albeit without the global coverage offered by existing country-year measures. Third, at least in the context of Mexico (alongside select other Latin American countries), some human rights abuses—such as large scale disappearances and massacres—is of direct policy, public, and civil society interest (Innes de Neufville, 1986; Saenz, 2017; Wilkinson, 2019). To the extent that our approach identifies relevant ATI requests in this vein, our framework and data stand to have direct real world impact by providing advocacy groups with a means to rapidly and reliably identify relevant ATI requests (and the information provided in response to these requests).

We implement our proposed approach by first using qualitative assessments and keyword searches to identify a subset of all Mexican ATI requests that potentially relate to human rights abuses, generating 187,145 requests. We next draw a random sample of 3,050 of these requests, and manually code each for whether or not they actually pertain to human rights abuses. We further code the degree to which that request implicated a state, nonstate, or unknown perpetrator, and whether or not the identified abuse pertained to a discrete abuse incident. After establishing the inter-coder reliability of our manual codings, we use our 3,050 manually labeled requests within an ensemble of supervised classifiers to label all keyword-identified 187,145 ATI requests for these same qualities.[3] We internally validate

[3] Our work thus builds upon extant treatments of human rights texts as a supervised learning problem (Greene, Park and Colaresi, 2019; Cordell et al., 2019*a*; Erlich et al., 2021).

these codings against a set of gold standard records of human rights-relevant ATI requests, as coded by an NGO with expertise in this area, and then aggregate our validated human rights coded requests to various (sub-annual and sub-national) temporal and geographic scales. The latter aggregations allow us *both* to highlight the rich variation that one obtains from the coding of human rights abuses from ATI requests *and* to externally validate our data against a number of established disaggregated data sources on human rights violations.

These steps produce a novel measure of human rights abuse that (i) recovers gold-standard human rights-based ATI requests with relatively high accuracy and (ii) offers substantially more spatio-temporal variation in human rights abuse than do many existing scholarly datasets. Herein, our paper also makes two additional and notable methodological contributions. First, we introduce to political scientists a recently developed synthetic minority over-sampling technique (SMOTE) that was originally created for the classification rare genomic features in imbalanced genetic datasets (Schubach et al., 2017). This approach is especially attuned to the rarity of our ATI-based human rights concerns. It accordingly outperforms several extant classification strategies for the most imbalanced classes in our human-labeled data. These findings are relevant to political violence machine learning research more generally, given that many forms of such violence exhibit far higher imbalance than do our labeled ATI data. Second, our overall measurement approach can also serve as a useful template for future researchers interested in deriving their own measures of public concern from (Mexican) ATI request data. Indeed, by following our framework for the measurement of human rights abuses from ATI requests, one could develop alternate fine-grained ATI-based measures of (e.g.,) environmental justice, government corruption, or public health.

Below, we next provide relevant background information. We then introduce our ATI request text sample and human coding approach. Afterwards, we discuss our supervised classification strategy. This is followed by an internal validation of our classified data. We then spatio-temporally aggregate our internally validated ATI requests and externally validate these aggregations against several accepted measures of human rights abuse for Mexico. Our conclusion summarizes our key findings and contributions in terms of content and methods.

Background

Sources for Human Rights Measurement

Prominent human rights datasets such as CIRI (Cingranelli and Richards, 2010; Cingranelli, Richards and Clay, 2014) and PTS (Wood and Gibney, 2010) code data from annual reports of countries' human rights practices, as produced by the U.S. State Department and/or by non-governmental organizations (NGOs) such as Amnesty International or Human Rights Watch. Importantly, these annual reports do not provide a complete record of every human rights abuse or repressive action that occurred in a particular country-year (Hill, Moore and Mukherjee, 2013; Conrad, Haglund and Moore, 2014; Cordell et al., 2019*a*). Rather, they aggregate allegations pertaining to a subset of relevant repressive acts for the time period covered (Cordell et al., 2019*a*). As a consequence, and notwithstanding the above datasets' strengths in terms of cross-national comparability, coding efforts employing these annual reports offer only a "standards based," country-year picture of human rights abuse.

While some progress has been made in extracting sub-annual and sub-national information from the annual country reports mentioned above (Cordell et al., 2019*b*), these country reports are primarily written with country-year units in mind, and exhibit a number of other potential biases (Clark and Sikkink, 2013; Hill, Moore and Mukherjee, 2013; Fariss, 2014; Potz-Nielsen, Ralston and Vargas, 2018). Many researchers have thus understandably turned to measure human rights abuses from international or national news(wire) sources, at times supplemented with NGO reports (Davenport and Ball, 2002; Raleigh et al., 2010; Sundberg and Melander, 2013). Doing so provides highly disaggregated spatio-temporal records of individual abuses. Often these records of discrete events—commonly referred to as political event data—are measured at the daily, and latitude-longitude coordinate, level.[4] However,

[4] Prominent examples include the Integrated Crisis Early Warning System (Boschee et al., 2015) and the Geolocated Event Dataset (Sundberg and Melander, 2013). In most cases an event's latitude-longitude

such data tend to exhibit reporting biases and spatio-temporal inaccuracies, especially for "less severe" events and/or for those that occur in more remote localities (Weidmann, 2015, 2016; von Borzyskowski and Wahman, 2019). To this end, scholars have identified marked divergences in the quality of reporting on state terror among news- and NGO-derived reports and eyewitness accounts, and have accordingly emphasized a need for *more diverse data sources* in such contexts (Davenport and Ball, 2002).

The latter call for more diversity in human rights data sources motivates our consideration of ICT-enabled forms of citizen-government interaction—and ATI requests more specifically—as a data source for the coding of human rights abuse. Counter to the political event data described above, ATI requests *do not* always reflect discrete human rights abuse incidents. Rather, as the example ATI requests in the Online Appendix highlight, ATI requests correspond to a *variety* of abuse-related queries, including: implied periods of heightened human rights abuse campaigns; citizen concerns, allegations, or anxieties over past, recent, or anticipated human rights abuse; and the identity of human rights abuse perpetrators—in addition to ATI requests pertaining to specific human rights abuse incidents. Much like the aforementioned Amnesty International or US State Department country-year reports, these ATI requests accordingly provide a "latent" indicator of human rights abuse intensity (Fariss, 2014), albeit at a much finer grained spatio-temporal scale than that offered by annual country reports on human rights practices. Before turning to these abuse-related ATI requests in further detail, we next provide additional background on ICT-enabled citizen government interactions. Following this, we briefly discuss Mexico's ATI-based system of citizen-government interactions on the whole.

ICT-Enabled Citizen-Government Interaction

The rapid spread of ICT-enabled (online or via SMS) platforms for citizen-government interaction yields rich data on individual users and their concerns. These platforms, sometimes called 'civic tech' (e.g. Peixoto and Sifry, 2017; Berdou and Shutt, 2017; Erlich et al.,

information is only accurate to the city- or municipality-level of geo-precision.

2018; Grossman, Platas and Rodden, 2018), include reporting platforms for local government services, complaint mechanisms, tools to communicate with representatives, and even crowd-sourcing platforms for issues like corruption. In some cases, the data such platforms yield are already being analyzed at large scale to better understand the nature of public problems (Chatfield and Reddick, 2018). In addition to purpose-built platforms, citizens also frequently communicate issues to government via various forms of social media via government's official pages or accounts. Finally, an increasing number of countries and jurisdictions have made ATI policies accessible via ICT-enabled platforms (Fumega and Scrollini, 2018), allowing citizens to more easily query government officials, and receive responses.

Because these platforms both lower the costs to citizens of communicating on issues with their government, and can yield detailed structured electronic records of such interactions, they create new opportunities for detecting and measuring citizens' reports about real-world problems. Applied to human rights abuses, such platforms offer the potential to avoid well-known biases pertaining to human rights information intermediaries such as news media, NGOs, and government reports; and to yield fine-grained measures that vary both temporally and—where platforms include geographic information—spatially.

Of course, many such ICT-enabled platforms face serious challenges of their own, including uptake by citizens (Peixoto and Sifry, 2017), responsiveness by government officials (Sjoberg, Mellon and Peixoto, 2017), and disparities in users that often replicate social and economic divides common in other forms of political participation (Pak, Chua and Moere, 2017). It thus remains an open question as to how useful the textual records from such platforms may be at detecting and measuring public concerns over real-world problems at scale. We focus here on a platform that has been *relatively* successful in terms of the volume and breadth of citizen usage, and the reliability of government responses: the case of Mexico's national ATI system.

Mexico's ATI System

Mexico's ATI system offers one specific instance of ICT-enabled citizen-government interaction.[5] Three reasons justify our choice of Mexico in this context. First, as described below, all relevant ATI request data are publicly available. Second the comprehensiveness of Mexico's ATI system ensures that Mexico's publicly-available ATI data exhibits high volume over both space and time. Third, Mexico's recent history of human rights abuse provides a highly suitable test case for our ATI measurement evaluations, including a wide array of associated resources for external validation. This intersection between the availability of ATI request data, spatio-temporal coverage, and external validation sources makes Mexico an optimal case study for this paper.

Mexico's 2002 *Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental* (LFTAIPG) established a unique online information platform known as INFOMEX.[6] An independent commission (likewise established under LFTAIPG) administers the INFOMEX system. This commission, originally known as Instituto Federal de Acceso a la Información (IFAI), has been referred to as the Instituto Nacional de Access a la Información Pública y Protección de Datos (INAI) since 2015. We provide further background on Mexico's ATI system in the Online Appendix. With the establishment of LFTAIPG and its online request system (INFOMEX), the texts of all Mexican national-level ATI requests, along with associated metadata, were made publicly available starting in June 2003. Even in instances where requesters submit written or verbal requests, agency officials enter the relevant information into INFOMEX.

This publicly available information corresponds to 1) all ATI requests made to Mexican federal government agencies, 2) other branches of the Mexican government, and 3) constitutionally autonomous bodies. INAI made requests pertaining to the latter two categories publicly available when coverage of the INFOMEX system expanded in 2016. These latter

[5] Table A.15 in our Online Appendix provides a list of additional ICT-enabled ATI systems for nation-states and similar units across the world.

[6] This online system was re-named as the Plataforma Nacional de Transparencia Gobierno Federal (PNT) in 2016. We continue to refer to the system as INFOMEX below for convenience.

two categories encompass to entities such as Mexico's Supreme Court and National Commission of Human Rights, whereas the requests falling in the former category were made available beginning in 2003 and encompass agencies and ministries such as Mexico's Secretary of the Interior and Secretary of Defense. The requests themselves correspond to queries made to these federal government bodies. Private Mexican citizens frequently make these queries, as do journalists, businesses, academics, and NGOs (Bookman and Guerrero Amparán, 2009). Topically, the requests cover, for example, queries for specific information relating to government spending, the environment, education, or the military (Berliner, Bagozzi and Palmer-Rubin, 2018).[7]

Coding Human Rights Abuses from ATI Requests

ATI Request Sample

We downloaded all available ATI requests and associated metadata from Mexico's INFOMEX web interface, for the June 2003 through June 2018 period. The corresponding sample includes 1,518,979 total ATI requests. Our focus is on the ATI *request* texts. These texts appear under a single variable field in our downloaded ATI request data, and correspond to each requester's open-ended description of the specific information they seek.

Although most requesters describe the nature of their requests within INFOMEX's primary request field, some requesters also include additional contextual information within two supplemental request text fields. First, in an "otros datos" field, requesters at times include additional textual information supporting their request.[8] "Otros datos" entries occur in 32.30% of all requests, and were merged into our primary request text field when they exist. Second, a smaller subset of requests (11.67%) include a portion or all of the request as an uploaded attachment (e.g., a Microsoft Word document or a PDF). We separately downloaded these attachments, digitized all attachment texts, and appended that

[7] Here and below, we omit requests for personal information, which INFOMEX also administers, given the confidential nature of this information and the lack of direct relevancy to our research objectives.

[8] E.g., a textual description of, or reference for, a news story, law, or document that their primary request made mention to.

text, where applicable, onto the main request text field.[9] Because some of these attachments contain massive spreadsheets or technical manuals totaling in the thousands of pages, we then truncated all combined textual entries from the thousandth character string onward. This truncation affects fewer than 0.02% of our requests. These steps created a corpus that included *all* available public ATI requests for the period June 2003-June 2018.

Human Coding

With this full sample of requests in hand, we next used keywords to identify the subset of all retained requests that could *potentially* pertain to human rights abuses.[10] Given that we anticipated narrowing this identified subset down further via human coding, we were intentionally over-inclusive in the keywords that we used to generate this initially identified subset of potential human rights related requests. Specifically, we employed a set of 41 n-grams or n-gram roots—typically unigrams or bigrams—whose usage within the text of an ATI request indicated that the request was potentially related to human rights abuse.

Our identification of these n-grams or n-gram roots followed a two step process. First, we qualitatively assessed a range of primary material relating to human rights abuses in Mexico[11] alongside past studies that have sought to either (i) identify Mexico's ATI requests for security-related requests based upon keywords (Almanzar, Aspinwall and Crow, 2018), or (ii) summarize the topics of Mexico's ATI requests via candidate words (Berliner, Bagozzi and Palmer-Rubin, 2018). This identified an initial set of keywords which we used to query and then retain any Mexican ATI request that contained at least one keyword.[12] From this initial request set, we summarized all unigrams, bi-grams, and tri-grams and reviewed the most frequent terms[13] to identify any keywords that may have been missed from our

[9] A negligible share of attachments were missing or were corrupted, and were hence not included in our analyses.

[10] For a similar application to country reports on human rights practices, see Cordell et al. (2019*a*).

[11] Specifically: CDHDF (2015), CDHDF/OAS (2015), OU-DN (2016), and USAID (2018).

[12] All keywords and queried texts were standardized to lower-case for this step.

[13] For this step, we summarized all (1) unigrams that appeared in at least 5,000 of the remaining request documents, (2) bigrams that appeared in at least 2,500 of the remaining documents and (3) trigrams that appeared in at least 2,500 of the remaining documents. These thresholds were chosen to ensure that we ended up with > 100 but $< 1,000$ unigrams to evaluate each case.

initial qualitative approach. This process identified several additional candidate keywords, which we added to our final keyword list, as depicted in Table A.1 of the Online Appendix. Using these final keywords, we then subset our full Mexican information request corpus to encompass only those requests containing at least one instance of a keyword on our list, yielding a total of 187,145 candidate human rights requests for human coding.

Our supervised coding scheme sought to then identify the subset of these 187,145 requests that actually pertained to human rights abuse. Altogether, we favored this approach over a wholly unsupervised approach in the interest of maintaining quality control over our coded abuse cases. We began by human coding a random sample of our 187,145 requests. Our human coding tasks separately coded binary indicators for whether (= 1) or not (= 0) a given request pertained to a human rights abuse perpetrated by (a) a state based actor, (b) a non-state based actor, or (c) an unknown actor. State based perpetrators encompassed any governmental actor, including the military and police. Non-state perpetrators encompassed citizens, businesses, private armed forces, and criminal groups. We designated all remaining cases as unknown. Instances where a request alluded to multiple perpetrator types received a coding of "1" in each perpetrator category. This coding scheme then also facilitated the post-coding creation of a more general "human rights abuse" indicator for any perpetrator type, hereafter referred to as "HRA," and coded as "1" for any perpetrator type. Separately, we also coded whether or not each identified HRA case pertained to a discrete human rights abuse incident—such as the 2014 Ayotzinapa massacre—versus a more general HRA.

For each measure, we developed a detailed coding rubric *a priori* so as to ensure that our human rights codings were both consistent and credible. To this end, we defined a "human rights abuse" to exclude requests for information on procedures or legislation, such as requests seeking to know whether the Mexican government had ratified a particular international human rights treaty or had enacted domestic human rights laws or services. We then more specifically define a "human rights abuse" as one relating to instances of disappearances, extrajudicial killings, political imprisonment, torture, or limitations upon freedom of assembly, association, movement, speech, or electoral self-determination—in each case as defined by the CIRI Human Rights coding scheme (Cingranelli and Richards, 2010), with one important adjustment: we also include instances of human rights abuse at the hands of non-government actors, unlike CIRI.[14] We provide (Spanish and English-translated) example requests that were coded as “1” for our four binary human rights indicators—and as “0” across all indicators—in the Online Appendix.

The above criteria do not encompass several additional types of human rights (abuses) that CIRI codes, or that arise more generally in the context of ATI requests. For example, our coding scheme does not code requests related to CIRI categories encompassing freedom of religion, workers’ rights, or women’s rights. After reviewing an initial sample of relevant requests, and without knowing a requester’s identity, we determined that coding categories related to the rights of particular identity groups was likely to produce sub-optimal intercoder reliability and poor overall coding consistency. Many such request cases, for example, request broad employment or salary statistics for a particular actor or group—requiring the coder to infer whether requester sought this information with regards to equality for that entity—or pertained to an employee airing frustrations about their particular work schedule with little clear evidence of a systematic abuse of workers rights. We likewise do not code instances where a requester alludes to abuses to their actual right to access public information, or to similar references made in relation to individual data privacy concerns, as a human rights abuse. Requesters make such references frequently within Mexico’s ATI system, but typically in a hypothetical manner—often as a reminder that the information being requested must be provided.

To apply our human coding scheme to our identified ATI request texts, we first drew a random sample of 3,150 ATI requests from our initial keyword sample of 187,145 requests for coding. Human coding involved two coders, who are also among this paper’s co-authors. The coders first used a sample of 100 of these requests for pre-coding practice and to inform

[14] However, we continue to exclude requests that directly pertain to violence between criminal organizations themselves, such as conflicts between rival cartels.

the coding rules discussed above. Following pre-coding, a separate sample of 1,000 requests was jointly coded. After the first 200 of this 1,000 joint coding sample were completed, the two coders held an initial calibration meeting to review their codes. The remaining 800 requests from this joint coding sample were then separately coded by each coder. Across both coders, we found that 0.5% (nonstate perpetrator), 0.8% (state perpetrator), 3.2% (unknown perpetrator), and 0.8% of all keyword-identified ATI requests were human coded as pertaining to a particular human rights abuse subset. This ensured that 4.2% of all human-coded human rights abuse requests pertained to our "any human rights" indicator.

For this jointly coded sample, we calculate Cohen's Kappas to assess inter-coder reliability. These statistics are reported in Table A.2 of the Online Appendix. Our Cohen's Kappas for the state- and non-state-perpetrator indicators, and for the incident level indicator, have been characterized within extant research as "good" (Steiner et al., 2004; Bächtiger and Hangartner, 2010, Note 5), with values of 0.66, 0.67, and 0.63 respectively. Cohen's Kappas for the unknown-perpetrator category, and for the joint human rights abuse indicator, are "excellent" (Steiner et al., 2004; Bächtiger and Hangartner, 2010, Note 5), with respective values of 0.84 and 0.89. Given these levels of inter-coder reliability, each human coder then coded an additional sample of 1,025 requests. This ensured a total human coded sample of 3,050 requests,[15] which, as noted above, were randomly drawn from our 187,145 (keyword-identified) candidate requests.

Supervised Text Classification

We next use our 3,050 hand-labeled requests to identify an appropriate set of supervised classifiers for our data, and to select appropriate tuning parameters for each classifier. In this instance, our binary human labels are the outcomes of interest, and our features correspond to a document-term-matrix (DTM) of unigrams appearing within the requests found within our training sample. Prior to creating this DTM, all request texts were pre-processed to remove stopwords, punctuation, sparse terms, numbers, individual letters, and placeholders

[15] That is, after discarding the initial 100 cases used for practice.

for blank entries,[16] and all remaining words were stemmed and converted to lower case. These steps are consistent with past automated analyses of Mexico's ATI request texts (Berliner, Bagozzi and Palmer-Rubin, 2018). Given our sample size, and following extant research (e.g., Lee, Liu and Ward, 2018), we then implemented our in-sample supervised classification exercises within a three-fold cross-validation framework.

These in-sample cross-validation assessments evaluated three machine learning classifiers: naive Bayes, random forests, and HyperSMURF. Naive Bayes classifiers employ Bayes' rule to perform probabilistic classification whilst treating all (DTM) features as independent. For each dichotomous human rights coding mentioned above, we use cross-validation and areas under the receiver operating characteristic curve (AUCs) to select an optimally performing naive Bayes classifier in terms of prior[17] and smoothing parameter.[18] Random forests use classification trees to identify the optimal features within random samples of one's data for binary partitions of one's outcome of interest. At each node, a predictor that provides the best partition is then selected. This selection is then repeated, with replacement, for subsequent random samples using additional classification trees. The generated predictions are then combined via majority vote to generate binary classifications that are relatively robust to overfitting (Breiman, 2001; Liaw and Weiner, 2002). We use cross-validation to select an appropriate number of classification trees for our random forest classifiers within each binary classification task, evaluating commonly used sizes of 10, 100, and 500.

One potential limitation for the random forests and naive Bayes classifiers proposed above is poor performance when dealing with imbalanced outcomes (e.g., outcomes with far fewer 1's than 0's). Our final supervised classification approach, HyperSMURF, was developed to address this particular problem within the context of rare genetic diseases (Schubach et al., 2017). The application below—to the best of our knowledge—is one of the method's first applications to a social science domain. As described below, HyperSMURF implements a

[16] E.g., instances where a requester entered in "xxxxxxx" in the main request text field when uploading their main request as an attachment instead.

[17] I.e., uniform or doc/term-frequency based.

[18] Across the set: 1, 5, 10.

hyper-ensemble (i.e., an ensemble of ensembles) of random forests in an imbalance-aware manner for a given supervised classification task. All of our binary human rights variables are highly imbalanced for our human-labeled data, with these variables exhibiting only 1-5% 1's and 95-99% 0's. In light of this imbalance, we suspect that HyperSMURF will provide a more appropriate and more competitive alternative to classifying our binary codings—both in sample and out-of-sample—than will either naive Bayes or random forests.

To perform classification in an imbalance-aware manner, HyperSMURF randomly partitions the observations in one's more imbalanced outcome category, which in our case corresponds to identified instances of concern over human rights abuses. It then applies a synthetic minority oversampling technique (SMOTE) to each partition, so as to generate additional synthetic instances of this rarer outcome category. The application of SMOTE addresses the inherent imbalance in our binary outcomes of interest, in that it ensures that our resultant training data contain an increased number of (synthetically generated) instances of human rights abuse concern (Schubach et al., 2017). The original partitioned instances of human rights abuse concern, along with these synthetic instances, are next combined with a comparable number of sampled zero cases for each human rights measure in order to construct a balanced set of parallel training datasets. A collection of h corresponding random forests are then run in parallel on these datasets, and their predictions are hyper-ensembled (Schubach et al., 2017). For each variable of interest, we use cross-validation to select an appropriate HyperSMURF specification across ranges of both h and the number of the features randomly selected within each h.[19]

The steps described above allow us to select optimal classifiers for each of our five variables of interest. We consider multiple out-of-sample classification statistics—specifically, AUC, area under the precision recall curve (AUC-PR), precision, recall, F1 scores, and overall accuracy—for each classifier and each binary outcome. All classification statistics were derived from three-fold cross-validations that utilize all 3,050 in-sample cases. Each classifier tended to perform commensurately in classifying our HRA class, with AUCs ranging

[19] Considering ranges of 10, 100, and 500; and 100 and 250; respectively.

from 0.81-0.90 and total accuracy ranging between 0.73-0.84. However, our three classifiers exhibited lower classification accuracy for our remaining variables—at times noticeably so. For example, across our two rarest perpetrator-specific variables[20] our AUCs range from 0.54-0.81, and overall accuracy declines to 0.54-0.81. In comparing precision and recall, we likewise find moderate-to-high recall, but notably low precision, across our perpetrator and incident specific variables—suggesting that classifications of these variables exhibit higher false positive rates than do our HRA classifications. In light of these trends, and given the fact that many of our perpetrator-specific human rights abuses were identified as "unknown perpetrator" in any case, we conclude that our combined HRA measure is the most internally consistent construct for our coding tasks, and primarily focus on this combined measure in the validation exercises below.

We next seek to determine the ideal classifier(s) for our full out-of-sample classification tasks. Across our cross-validation results for each variable, all three classifiers perform comparably across each variable of interest, with three exceptions. First, naive Bayes performs poorly in classifying state and nonstate perpetrated human rights abuses, especially in terms of AUC, precision, F1-score, and accuracy. Second, random forests underperform relative to our other classifiers on AUC-PR for state perpetrated human rights abuses, non-state perpetrated human rights abuses, and human rights abuse incidents. Finally, HyperSMURF outperforms naive Bayes and random forests across most binary outcomes for our two most preferred classification statistics (AUC and AUC-PR), suggesting that this SMOTE-based method successfully addresses our class imbalance issues better than standard approaches. These exceptions notwithstanding, all three classifiers exhibit unique strengths in terms of both specific classification statistics and abilities to classify some of our specific variables of interest over others. In light of this, we favor an ensemble of all three classifiers for each variable within our final (out-of-sample) supervised classifications.

Having identified a primary human rights measure of interest (HRA), and an optimal ensemble of classifiers, we next return to our full sample of 187,145 potential human rights-

[20] State and non-state perpetrated human rights abuses.

related requests. Based upon the tasks above, 3,050 of these are now human labeled and 184,095 remain unlabeled. We pre-process this full set of potential human rights abuse requests in the manners described above before classification, and convert all remaining unigrams to a DTM. For each of our five variables, we next re-train and re-run our three classifiers on our full set of 3,050 labeled cases for that variable, and then use the parameters from these training models to classify all remaining 184,095 request texts.[21] Our naive Bayes, random forests, and HyperSMURF classifications for each variable are then ensembled to produce a single measure using majority vote. Across our full 187,145 request sample, we find that this approach identifies 37,970 requests pertaining to HRA's. The approach comparably identified 37,628, 31,173, 40,848, and 36,392 cases for our state perpetrated, nonstate perpetrated, unknown perpetrator, and incident-level human rights abuse variables, respectively.[22] As above, this suggests that the latter measures at times exhibit a moderate degree of overprediction, relative to our primary HRA indicator. Hence, for each indicator, we retain all identified human rights abuse cases and—as mentioned earlier—focus primarily on our HRA indicator during validation.

Validation

There are two types of validation for coded text data: internal validation and external validation (Bagozzi et al., 2019). Internal validation assesses whether one's coding approach accurately recovers the true (non)instances of a construct of interest *within the actual text data being coded* by comparing one's codings to a sample of "gold standard" codings of that same text data. External validation evaluates whether one's codings accurately reflect external events and related "on the ground" measures of the construct of interest, as coded from sources that are *distinct from* the original text data. We perform both types of validation below.

[21] For each model and variable, we use the tuning parameter values identified in the cross-validation exercises above. We then dichotomize each resulting prediction according to the optimal cutpoint that was identified for a given classifier and variable during cross-validation.

[22] Additional descriptive tables and plots—both over time and by target Mexican Federal Agency—for these full human rights classifications appear in Tables A.3-A.7 and Figure A.1 of the Online Appendix.

Internal Validation: Artículo 19

Our internal validation focuses on a set of "gold standard" ATI requests that the NGO Artículo 19 (hereafter A19)—the Mexican Chapter of the International NGO Article 19—has identified, coded, and archived. Globally, Article 19's campaigns work to understand, interpret, and promote new policies and laws pertaining to human rights at both the national and international levels. Artículo 19 collaborates in this vein with INAI. Together, they run an initiative known as Proyecto Memoria y Verdad (Project of Memory and Truth, hereafter PMV), which began in 2015. Part of PMV's work seeks to identify, compile, and highlight Mexican ATI requests and related information concerning grave human rights violations in Mexico from 1960 onward.

To achieve these overarching goals, PMV (i) promotes the non-repetition of serious human rights abuses in Mexico, (ii) improves the right to the truth in such contexts, and (iii) facilitates ATI for human rights abuse victims, investigative bodies, jurisdictional bodies and/or guarantors of human rights, courts and any other interested party (Memoria Y Verdad, 2016*b*). In these endeavors, PMV has assembled curated datasets of ATI requests pertaining to 15 major human rights abuse incidents or campaigns that have occurred in Mexico from 1960-present.[23] Upon determining that public information related to these 15 cases was inaccessible, of poor quality, and/or incomplete, A19 and PMV identified specific information gaps, and performed an exhaustive search to identify and collect relevant information pertaining to relevant ATI requests, ATI responses, review resources, INAI resolutions, multimedia materials, and reports from international organizations and NGOs (Memoria Y Verdad, 2016*a*). Each identified piece of information was then systematically classified, individually analyzed, and categorized according to multiple dimensions—including its relevance to PMV's 15 human rights violation cases (Memoria Y Verdad, 2016*a*).

We focus on the ATI requests that PMV identified. PMV compiled these requests into 15 spreadsheets; one for each major human rights abuse case. These spreadsheets include a request identifier, the date of the request, the request's target agency, and up to 50 additional

[23] These 15 cases are listed in the Online Appendix.

variables for various request and response characteristics. The Online Appendix provides brief background summaries for PMV's 15 human rights abuse cases, the range of request dates associated with each human rights abuse, the total number of ATI requests identified under each case, and additional relevant notes. We use the PMV-identified ATI requests from 14 of these 15 human rights abuse cases for internal validation,[24] and assess the extent to which our own coded ATI requests recover the ATI requests that PMV identified.

To do so, we assess how well our HRA codings classify the PMV's identified ATI requests for our full sample of 1,518,979 requests.[25] We evaluate classification performance based on precision, recall, F1-Scores, and total accuracy. This evaluation is an exceptionally high bar for internal validation. Our binary records of PMV-identified requests were not used to train our own codings of known human rights-related ATI requests, and we did not include proper nouns related to any of PMV's 15 human rights abuse events in the initial keyword-based subsetting of our ATI request sample. The latter point is especially relevant, given that many of PMV's identified ATI requests do not mention human rights violations explicitly; rather they simply refer to the name of the human rights abuse incident when they solicit information related to that event. Moreover, PMV only identified 1,068 total unique ATI requests associated with the 14 human rights abuse cases. Since we are attempting to correctly classify 1,068 cases out of our full sample of 1,518,979 requests, we are trying to predict an out-of-sample binary outcome with only 0.07% 1's, and 99.93% 0's. Hence, standard rules of thumb for precision and recall do not apply in our case, and we accordingly focus on relative comparisons of precision and recall rather than their absolute values.

To assess HRA's classification performance in this context, we evaluate this measure's *relative* classification performance against three plausible baselines. First, we create a binary record for any of the 1,518,979 requests that were identified as potentially related to a human rights abuse by our initial keyword method. This measure is thus equal to 1 for our 187,145 keyword-request cases and zero otherwise; and allows us to assess the value added of our

[24] We omit one case because the compiled spreadsheet of ATI codings for that case remained unavailable for download at the time of writing. See the Online Appendix for further details.

[25] The Online Appendix reports results from our additional human rights measures.

human-coding steps relative to a more naive keyword-only approach. Next, we construct two random baselines for comparison, hereafter denoted ξ. For our first ξ, we generate random binary human rights abuse classifications with probability $\frac{1}{2}$. For the second ξ, we generate comparable random binary classifications with probability equal to the mean of our true binary PMV sample proportion $\bar{y} = 0.0007$. As such, $\xi = \bar{y}$ provides us with a random guessing baseline that preferences overall accuracy, whereas $\xi = \frac{1}{2}$ provides us with a random guessing baseline that instead maximizes the identification of our less common class (i.e., PMV's actual human rights abuse ATI codings).

We present our classification results for our HRA codings, our human rights abuse keyword sample on the whole,[26] $\xi = \bar{y}$, and $\xi = \frac{1}{2}$ in Table 1. Comparable results using our additional human rights abuse indicators appear in Table A.8 of the Online Appendix. Turning to Table 1, we find that our HRA indicator exhibits substantially higher precision than any of our baseline comparisons, with a precision value that is roughly triple that of the human rights keyword sample, and that is seven to nine times that of $\xi = \bar{y}$ or $\xi = \frac{1}{2}$.[27] Thus—for the 1's recorded by each of these approaches—a notably higher share correspond to PMV's human rights abuse cases amongst our HRA codings, in relation to our alternative baselines. Recall (the proportion of PMV cases that our approaches correctly predict as 1's) in turn indicates that roughly 23% of all PMV cases are recovered by HRA, 38% by our keyword sample, 50% by $\xi = \frac{1}{2}$, and 0.1% by $\xi = \bar{y}$. This suggests that—relative to HRA—a larger share of all PMV cases lie within the 1's for $\xi = \frac{1}{2}$ and our full human rights keyword sample; whereas a far smaller share fall within the 1's on $\xi = \bar{y}$.

However, the remaining classification statistics in Table 1—as well as the precision values discussed above—suggest that the relatively higher recall values on $\xi = \frac{1}{2}$ and on our human rights keyword sample come at the cost of a substantially higher share of false positives, in

[26] Which represents a ceiling on the PMV cases that our HRA indicator can recover.

[27] Precision in this case denotes the fraction of an approach's predicted human rights abuses that were in fact PMV cases. Because our HRA and keyword approaches were constructed to code human rights abuses that extend well beyond the specific abuse cases coded by PMV, the share of "false positives" (instances where a predicted human rights abuse was not related to one of PMV's 14 abuse events) in the present comparisons is naturally very high. This in turn ensures that all corresponding precision values are very low.

comparison to HRA. This can be observed in the F1-Scores in Table 1, which indicate that HRA's combined levels of precision and recall are three-to-thirteen times larger than (i.e., superior to) those of any of our baseline models. This can also be seen by the percentage of PMV cases correctly classified (Accuracy) in the final column of Table 1, wherein HRA correctly classifies 97% of all sample cases in comparison to only 88% of all cases for the keyword indicator, and only 50% of all cases for our coinflip indicator ($\xi = \frac{1}{2}$).

Table 1: Internal Validation Classification Statistics

	Precision	**Recall**	**F1 score**	**Accuracy**
Any Human Rights Abuse (HRA)	0.64	22.66	1.24	97.46
Human Rights Abuse Keyword Sample	0.22	38.39	0.44	87.66
$\xi = \frac{1}{2}$	0.07	50.36	0.14	50.03
$\xi = \bar{y}$	0.09	0.09	0.09	99.86

Note: $N = 1,518,979$.

In summary, HRA recovers a notable share of a separately recorded and verified sample of ATI requests pertaining to a set of 14 specific human rights abuse cases. Table 1 further suggests that our HRA indicator minimizes the rate of false positives obtained, relative to the full samples of potentially human rights related requests that we identified via keywords and to random guessing. As demonstrated in Table A.8 of the Online Appendix, our findings for each of the additional human rights indicators that we coded—pertaining to specific perpetrators or human rights abuse incidents—reinforce these conclusions. Moreover, as the example requests in our Online Appendix and our external validations below highlight, our approach also captures an extensive variety of additional human rights-based requests that extend well beyond the incidents considered by PMV.

Internal Validation: Extant ATI Topics

We next internally validate our human rights classifications against a second distinct ATI-based set of measures: the twenty fully unsupervised thematic topics identified for Mexican ATI requests from 2003-2015 by Berliner, Bagozzi and Palmer-Rubin (2018). These topics encompass themes ranging from (e.g.) taxes and finance; health statistics; education; the environment and land; and military, police, and crime. For requests overlapping during the 2003-2015 period, Table A.9 in the Online Appendix reports bivariate correlations between these twenty topics' document-level posterior probabilities and each of our human rights indicators. Across all of our human rights indicators, we find negligible correlation coefficients (-0.045 ↔ 0.045) for 19 of our 20 topics. However, for one topic—Topic 16: military, police, and crime—we consistently find large, positive, and statistically significance correlation coefficients of up to 0.351 (for our HRA indicator). Thus, our HRA measure is internally valid relative to a second, wholly unsupervised, request-level measure, in that the former is strongly positively associated with the one security-related topic identified by Berliner, Bagozzi and Palmer-Rubin (2018), but not strongly associated with any of those authors' other 19 (non-security) topics.

Taken together, the above findings suggest that our measurement approach—and our identified HRA ATI requests—each exhibit a notable degree of internal validity. With this in mind, we next turn to evaluating how our aggregated HRA requests conform with a pair of extant, and externally coded, measures of human rights for the case of Mexico.

External Validation: Spatial Variation

To externally validate our HRA measure against subnational data on human rights abuses for Mexico, we consider two of the most widely used global event datasets with available data through 2018: The Integrated Crisis Early Warning System dataset (ICEWS; Boschee et al., 2015) and the Georeferenced Event Dataset (GED; Sundberg and Melander, 2013). These datasets record political events from an extensive array of international and local news sources, as well as from NGOs in the case of GED. Previous research has used these datasets to study human rights violations (e.g., Fjelde and Hultman, 2014; Wood and Sullivan, 2015; Sharma et al., 2017), including validation assessments (Bagozzi et al., 2019). We subset each event dataset to only contain instances of human rights abuses against civilians arising from relevant source actors in Mexico, for at least a municipality-level of geo-location precision.

We then combine these data with our HRA indicator at the municipality-day level.[28] Full details on our event data aggregation decisions are included in the Online Appendix.

These aggregation steps generate event records at the municipality-day level. For the purposes of initial comparison, we collapse these municipality-day event counts to the municipality level, and likewise generate Mexican municipality-level counts of our HRA ATI requests. We first visually compare these 2003-2018 municipality-level counts via municipality maps. Given that each set of counts is highly skewed,[29] we log each count before plotting these quantities on maps. Next, and because the scale of our (logged) counts differs by several orders of magnitude—wherein at the municipality level our ATI, ICEWS, and GED human rights abuse measures exhibit ranges of $0 \leftrightarrow 4925$, $0 \leftrightarrow 366$, and $0 \leftrightarrow 4$ respectively—we place all three sets of (logged) municipality-level counts on a consistent 0-1 scale for plotting and subsequent comparison using minimum-maximum normalization:

$$\text{Scaled Count}_i = \frac{\ln(\text{Count}_i + 1) - \ln(\text{Count}_{min} + 1)}{\ln(\text{Count}_{max} + 1) - \ln(\text{Count}_{min} + 1)}$$

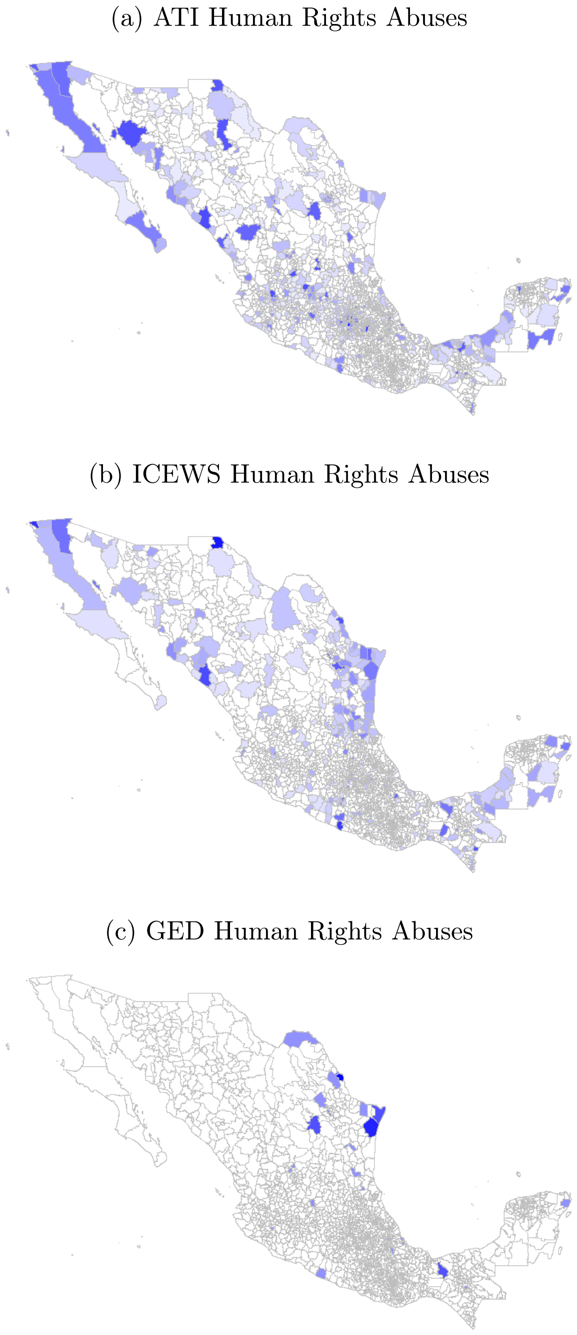
where “Count” denotes a particular count measure of interest (e.g., the HRA ATI counts or our ICEWS event counts), i is a given municipality, ln is the natural logarithm, and min and max are the minimum and maximum municipality counts for a given measure across the entire 2003-2018 period. Our resultant ATI- and event data-scaled counts of human rights abuses are then plotted at the municipality-level in Figure 1 below.

Based on Figure 1, our ATI-based HRA measure exhibits a striking level of similarity with the human rights abuses derived from ICEWS, whereas GED exhibits far more sparsity and hence less comparability to either of the HRA or ICEWS human rights abuse measures. The latter finding is expected, given that GED only records fatal human rights abuses with identifiable perpetrators, whereas ICEWS and our HRA measure incorporate a wider variety

[28] In limiting these external validation comparisons to Mexican municipalities, we omit roughly 8% of our classified ATI requests—corresponding to ATI requests that arose from requesters based outside of Mexico or requesters that did not provide sufficient geographic information for this level of aggregation.

[29] Where at this level of aggregation, our HR-any counts exhibit skewness of 22.48, our ICEWS counts exhibit skewness of 26.68, and our GED counts exhibit skewness of 15.06.

Figure 1: Municipality-level Scaled Human Rights Abuses, 2003-2018



(a) ATI Human Rights Abuses

(b) ICEWS Human Rights Abuses

(c) GED Human Rights Abuses

of (potential) physical confrontations that do not yield any fatalities. Returning to Figure 1, we can also note several interesting discrepancies. Our ATI-based measure appears to capture more intense levels of human rights abuses than either ICEWS or GED within municipalities falling in Mexico's North-West and North-Central States—where conflict and crime associated with drug and human smuggling is known to be rife—most notably in Baja California, Baja California Sur, Sinaloa, Sonora, and Chihuahua. We also can observe that our ATI-based measure identifies abuses within a larger number of municipalities throughout Mexico's central and south central states than do either ICEWS or GED. Together these trends suggest that measures based on direct citizen communications may capture more breadth in human rights abuses than media-based measures. At the same time, ICEWS and GED do appear to capture relatively higher rates of human rights abuses in Mexico's North-Eastern States of Tamaulipas and Nuevo León. This suggests that future studies of human rights may benefit from jointly leveraging the measures considered here—a point we elaborate further upon in the Online Appendix. This point not-withstanding, the striking similarities between HRA and ICEWS in Figure 1 suggest that the former is indeed a valid measure of human rights abuses at this level of aggregation, and possibly one that provides more geographic coverage and variation than standard event data-based approaches.

External Validation: Spatio-temporal Variation

Figure 1 does not take into account the temporal variation in our respective measures of human rights abuses. We therefore evaluate a series of pairwise correlations amongst our HRA measure and our ICEWS- and GED-based measures of human rights abuses across multiple levels of spatio-temporal aggregation. In each aggregation, we standardize all measures using the minimum-maximum standardization formula presented above. We then calculate Pearson's correlations among bivariate pairings involving (i) our ICEWS and GED human rights abuses measures (as a baseline for comparison), (ii) our ICEWS and HRA measures, and (iii) our GED and HRA measures. We specifically assess each of these correlations at municipality-day, municipality-week, municipality-month, municipality, and monthly aggregations.[30] The results from these correlation exercises are presented in Table 2 below.

Looking across the columns in Table 2, we find that the correlations between our HRA measure and each event data measure are positive and are statistically significant at the $p < .01$ level in nine of ten possible instances. The correlations between the ICEWS and GED human rights abuses are also positive and statistically significant at the $p < .01$ level in four of five possible cases. We further find that the positive correlations involving ICEWS and HRA are, on average, over twice as strong as those involving ICEWS and GED in four of our five aggregations: Municipality-Day, Municipality-Week, Municipality-Month, and Municipality. The exception is the purely-monthly data aggregation, in which case neither the ICEWS-GED pairing nor the ICEWS-HRA pairings are statistically significant. However, for this aggregation, our HRA measure continues to exhibit a statistically significant correlation with GED. Further, the size of this HRA-to-GED correlation is approximately five times that of the non-significant correlation between GED and ICEWS at this same level of aggregation.

These correlations strongly suggest that the HRA measure is an externally valid measure of human rights abuse at multiple levels of sub-national and sub-annual validation. Indeed, the highest correlation of any pairing in Table 2 is that involving ICEWS and HRA at the municipality level, which is equal to 0.37. Hence, HRA is likely a valid measure of human rights abuses for the case of Mexico. It is also likely to offer richer variation than either ICEWS or GED, given that our HRA measure includes 34,543 discrete instances of 1's for our data, in comparison to 2,322 1's for ICEWS and only 34 1's for GED.[31]

The above quantities also reinforce our earlier points as to why our HRA measure is typically not correlated as highly with GED as it is with ICEWS: our GED data only include fatal human rights abuse events with identifiable (e.g., government) abusers, whereas the ICEWS data include all material human rights abuse events. The latter encompass both fatal and non-fatal events, including those arising from unidentified abusers. HRA, by com-

[30] We apply one-a-day filtering to ICEWS in order to address duplicate events, which imposes an artificial ceiling on our daily ICEWS events. Hence, our daily-level correlations should be interpreted cautiously.

[31] Our HRA measure records at least one human rights abuse for 685 unique municipalities; whereas ICEWS and GED only report violations in 324 and 25 municipalities, respectively.

parison, not only captures material (including both fatal and non-fatal) events related to human rights abuses (like ICEWS), but also includes (request) instances where a *potential or suspected* human rights abuse may have arisen. Thus, for scholars interested in subnational human rights, measures of human rights abuse obtained from ATI requests can be considered valid relative to global event data measures, whilst also offering researchers with far more information pertaining to abuses, and potential abuses, that do not make it into (international) media and NGO reports.

Table 2: Pearson's Correlations Between HRA, ICEWS Human Rights Abuses, and GED Human Rights Abuses

	Muni-Day	**Muni-Week**	**Muni-Month**	**Municipality**	**Monthly**
ICEWS & GED	0.0105**	0.0225**	0.0422**	0.2267**	0.0393
HRA & ICEWS	0.0208**	0.0821**	0.1537**	0.3689**	-0.1417
HRA & GED	0.0025**	0.0023**	0.0051**	0.1624**	0.1934**
N	13,535,613	1,928,745	444,717	2,457	181

Note:$* = p < .05$, $** = p < .01$ All variables have been standardized using min-max standardization.

The Online Appendix evaluates the robustness of the above validation tests in several manners. First, we omit all ICEWS, GED, and HRA-based data that come from municipalities falling within Mexico's Federal District. NGOs based in the Federal district likely file many human rights-related requests from the Federal District that seek information pertaining to human rights abuses elsewhere. Hence, removing cases from Federal District-based requests addresses this potential form of measurement error in our HRA aggregations. Second, we aggregate our municipality-level HRA, ICEWS, and GED scaled-events to the state-level to ensure that our findings also hold at this much coarser level of spatial aggregation. Third, we reevaluate our comparisons after retaining a maximum of one HRV-coded ATI request per municipality-day, to address potential overcounting of human rights abuses in our ATI data. While some findings weaken under these alternate configurations and lower N's, Figures A.2-A.4 and Tables A.10-A.12 illustrate that our conclusions generally hold under these alternate frameworks for comparing our ICEWS, GED, and HRA data. Following this, a series of count-based regression analyses offered in Tables A.13-A.14 then provide additional insights into the (under-reporting) correlates that each measure (i.e., ICEWS, GED, and HRA) exhibits at the Municipality-Month level.

Finally, recall that our introduction emphasized the promise of ATI-oriented human rights data in relation to current country-year human rights measures. While the latter measures retain advantages relative to our data in terms of cross-national comparability, we argued above that ATI data offer strengths in measuring *within-country* variation. To illustrate this, Figure 2 below externally validates our HRA-based data—in this case measured as a percentage of all monthly ATI requests—against one prominent country-year level human rights measure that has coverage for our full 2003-2018 period: the Political Terror Scale (PTS; Wood and Gibney, 2010). Both measures exhibit similar overall trends, with lower levels of human rights abuse from 2003-2008, an upward trend from 2009-2012, and then a fairly constant level of abuses thereafter (aside from declines at the very ends of our series). Yet, our ATI data provide substantially more variation in abuses within each of these time-windows, with several notable HRA-spikes in 2004, 2005, 2012, and 2015 that the PTS' annual data miss entirely. As such, these findings help to further demonstrate (i) the external validity of our HRA data and (ii) its relative within-country strengths.

Conclusion

This paper assesses the relative merit of using access-to-information (ATI) requests to systematically code human rights abuses at fine-grained spatio-temporal scales. Extant quantitative measures of human rights abuses are typically either bounded to the country-year level of aggregation, or are susceptible to media reporting biases due to the primary sources that they rely upon for coding. We expand upon this current measurement and understanding of human rights, and advance insights into how these rights are protected by national governments. As we show, text-based ATI requests offer uniquely disaggregated records of human rights abuses, and supervised coding of these requests in turn yields *externally valid*

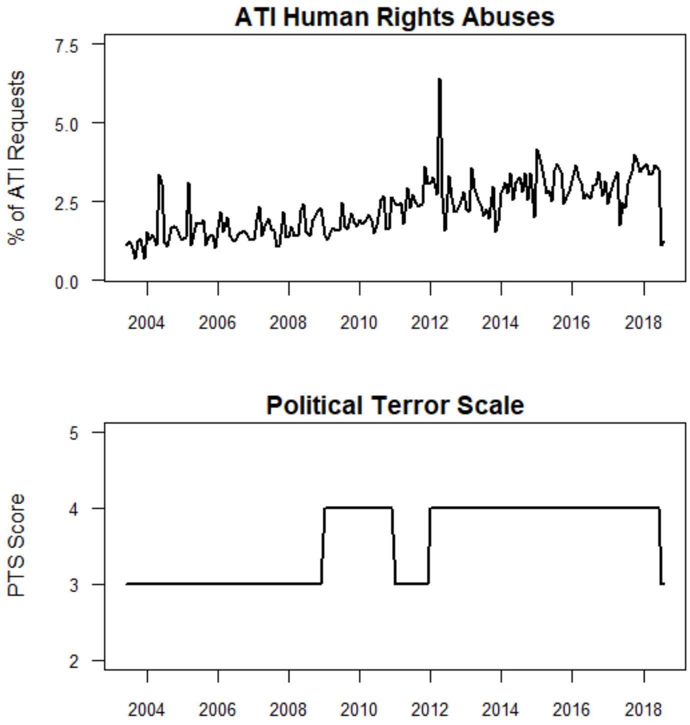


Figure 2: Comparison of Monthly Human Rights Abuses in Mexico

records of domestic human rights abuses across both time and space. This approach provides researchers, non-profits, and governments with a better grasp of the fine-grained nature of human rights abuses. We demonstrate this through the application of an innovative supervised machine classification approach to a novel dataset of federal ATI requests for the case of Mexico from 2003 to 2018. We further illustrate the *internal validity* of our approach—and of our coded human rights abuse cases—with the aid of "gold standard" ATI requests pertaining to high profile human rights abuses that were NGO-identified and coded.

This study thus provides the first successful quantitative effort to code human rights abuses from ATI request texts, along with validation of these coded data. In this respect, our proposed method has important policy implications and its future application stands to help human rights defenders identify potentially unidentified cases of abuse. Our results also highlight the broader promise of efforts to measure public problems using large-scale textual records from platforms for ICT-enabled citizen-government interactions. As the availability and usage of such platforms increases, this approach will become increasingly applicable and useful in measuring human rights concerns across multiple contexts. Such innovations will directly complement recent calls for big data innovations within global efforts to measure sustainable development outcomes by the United Nations and others (United Nations, 2017). Finally, the machine learning methods introduced above—in particular HyperSMURF—also stand to benefit peace and conflict research more broadly. Indeed, given the rarity of many forms of political violence, HyperSMURF will likely be indispensable to future researchers interested in conflict forecasting and/or conflict early warning.

Author's Note

All replication materials, including data and code, are available on the *Journal of Conflict Resolution* website.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Benjamin Bagozzi and Sarah Ellington's research was supported by the University of Delaware General University Research Fund. Brian Palmer-Rubin's research was supported by the Marquette University Committee on Research. Aaron Erlich's research was supported by Social Science and Humanities Research Council (SSHRC) Grant #430-2018-1069 and Fonds de recherche du Québec - Société et culture (FQRSC) Grant #253243.

References

Almanzar, Tanya, Mark Aspinwall and David Crow. 2018. "Freedom of information in times of crisis: The case of Mexico's war on drugs." *Governance* 31(2):321–339.

Bächtiger, André and Dominik Hangartner. 2010. "When Deliberative Theory Meets Empirical Political Science: Theoretical and Methodological Challenges in Political Deliberation." *Political Studies* 58:609–629.

Bagozzi, Benjamin E., Patrick T. Brandt, John R. Freeman, Jennifer S. Holmes, Alisha Kim, Agustin Palao Mendizabal and Carly Potz-Nielsen. 2019. "The Prevalence and Severity of Underreporting Bias in Machine and Human Coded Data." *Political Science Research and Methods* 7(3):641–649.

Berdou, Evangelia and Cathy Shutt. 2017. "Shifting the spotlight: understanding crowdsourcing intermediaries in transparency and accountability initiatives." *Making All Voices Count Research Report* .

Berliner, Daniel, Benjamin E. Bagozzi and Brian Palmer-Rubin. 2018. "What Information Do Citizens Want?: Evidence from One Million Information Requests in Mexico." *World Development* 109:222–235.

Bookman, Zachary and Juan-Pablo Guerrero Amparán. 2009. "Two Steps Forward, One Step Back: Assessing the Implementation of Mexico's Freedom of Information Act." *Mexican Law Review* 1(2):25–49.

Boschee, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz and Michael Ward. 2015. "ICEWS Coded Event Data." `http://dx.doi.org/10.7910/DVN/28075`. Harvard Dataverse, V4.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45(1):5–32.

CDHDF. 2015. "Derechos Humanos en México Y América Latina: Una Visión Desde El Consejo De La CDHDF." Comisión de Derechos Humanos del Distrito Federal.

CDHDF/OAS. 2015. "Situación de Derechos Humanos en México." Comisión de Derechos Humanos del Distrito Federal Y Organización de los Estados Americanos. OEASer.LVII, Doc. 44/15.

Cederman, Lars-Erik and Kristian Skrede Gleditsch. 2009. "Introduction to Special Issue on "Disaggregating Civil War"." *Journal of Conflict Resolution* 53(4):487–495.

Chatfield, Akemi Takeoka and Christopher G. Reddick. 2018. "Customer agility and responsiveness through big data analytics for public value creation: A case study of Houston 311 on-demand services." *Government Information Quarterly* 35(2):336–347.

Cingranelli, David L. and David L. Richards. 2010. "The Cingranelli and Richards (CIRI) Human Rights Data Project." *Human Rights Quarterly* 32(2):401–424.

Cingranelli, David L., David L. Richards and K. Chad Clay. 2014. "The CIRI Human Rights Dataset." http://www.humanrightsdata.com.

Clark, Ann Marie and Kathryn Sikkink. 2013. "Information Effects and Human Rights Data: Is the Good News about Increased Human Rights Information Bad News for Human Rights Measures?" *Human Rights Quarterly* 35(3):539–568.

Conrad, Courtenay R., Jillienne Haglund and Will H. Moore. 2014. "Torture Allegations as Events Data: Introducing the Ill-Treatment and Torture Specific Allegation Data." *Journal of Peace Research* 51(3):429–438.

Cordell, Rebecca, K. Chad Clay, Christopher J. Fariss, Reed M. Wood and Thorin Wright. 2019*a*. "Disaggregating Repression: Identifying Physical Integrity Rights Allegations in Human Rights Reports." Working Paper.

Cordell, Rebecca, K. Chad Clay, Christopher J. Fariss, Reed M. Wood and Thorin Wright. 2019*b*. "Recording Repression over Space and Time: Identifying Allegations in Annual Country Human Rights Reports." Working Paper.

Davenport, Christian and Patrick Ball. 2002. “Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995.” *Journal of Conflict Resolution* 46(3):427–450.

Erlich, Aaron, Danielle F. Jung, James D. Long and Craig McIntosh. 2018. “The Double-edged Sword of Mobilizing Citizens Via Mobile Phone in Developing Countries.” *Development Engineering* 3:34–46.

Erlich, Aaron, Stefano G. Dantas, Benjamin E. Bagozzi, Daniel Berliner and Brian Palmer-Rubin. 2021. “Multi-label Prediction for Political Text-as-Data.” *Political Analysis* .

Fariss, Christopher J. 2014. “Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability.” *American Political Science Review* 108(2):297–316.

Fjelde, Hanne and Lisa Hultman. 2014. “Weakening the Enemy: A Disaggregated Study of Violence against Civilians in Africa.” *Journal of Conflict Resolution* 58(7):1230–1257.

Fumega, Silvana and Fabrizio Scrollini. 2018. “Exploring the role of digital civil society portals in improving Right to Information regimes.” *U4 Anti-Corruption Resource Centre* 1.

Gleditsch, Kristian Skrede, Nils W Metternich and Andrea Ruggeri. 2014. “Data and progress in peace and conflict research.” *Journal of Peace Research* 51(2):301–314.

Greene, Kevin T., Baekkwan Park and Michael Colaresi. 2019. “Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects.” *Political Analysis* 27(2):223–230.

Grossman, Guy, Melina R Platas and Jonathan Rodden. 2018. “Crowdsourcing accountability: ICT for service delivery.” *World Development* 112:74–87.

Hill, Daniel, Will Moore and Bumba Mukherjee. 2013. "Information Politics v. Organizational Incentives: When are Amnesty International's 'Naming and Shaming' Reports Biased?" *International Studies Quarterly* 52(2):219–232.

Innes de Neufville, Judith. 1986. "Human Rights Reporting as a Policy Tool: An Examination of the State Department Country Reports." *Human Rights Quarterly* 8(4):681–699.

Lee, Sophie J., Howard Liu and Michael D. Ward. 2018. "Lost in Space: Geolocation in Event Data." *Political Science Research and Methods* p. 1–18.

Liaw, Andrew and Matthew Weiner. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

Memoria Y Verdad. 2016*a*. "Metodología para la construcción de casos." `http://www.memoriayverdad.mx/index.php/acerca-de.html`. Accessed on 7/22/2019.

Memoria Y Verdad. 2016*b*. "¿Qué Es Memoria Y Verdad?" `http://www.memoriayverdad.mx/`. Accessed on 7/22/2019.

OU-DN. 2016. "20 Claves Para Conocer Y Comprender Los Derechos Humanos." Oficina en México del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (ONU-DH).

Pak, Burak, Alvin Chua and Andrew Vande Moere. 2017. "FixMyStreet Brussels: socio-demographic inequality in crowdsourced civic participation." *Journal of Urban Technology* 24(2):65–87.

Peixoto, Tiago and Micah L. Sifry. 2017. *Civic Tech in the Global South.* The World Bank.

Potz-Nielsen, Carly, Robert Ralston and Thomas R. Vargas. 2018. "Recording Abuse: How the Editing Process Shapes our Understanding of Human Rights Abuses." Working Paper.

Raleigh, Clionadh, Andrew Linke, Håvard Hegre and Joakim Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature." *Journal of Peace Research* 47(5):651–660.

Saenz, Rodolfo D. 2017. "Confronting Mexico's Enforced Disappearance Monsters:How the ICC Can Contribute to the Process of Realizing Criminal Justice Reform in Mexico." *Vanderbilt Journal of Transnational Law* 50:45–112.

Schubach, Max, Matteo Re, Peter N. Robinson and Giorgio Valentini. 2017. "Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants." *Scientific Reports* 7(2959).

Sharma, Kiran, Gunjan Sehgal, Bindu Gupta, Geetika Sharma, Arnab Chatterjee, Anirban Chakraborti and Gautam Shroff. 2017. "A complex network analysis of ethnic conflicts and human rights violations." *Scientific Reports* 7(8283).

Sjoberg, Fredrik M, Jonathan Mellon and Tiago Peixoto. 2017. "The effect of bureaucratic responsiveness on citizen participation." *Public Administration Review* 77(3):340–351.

Steiner, Jürg, André Bächtiger, Markus Spörndli and Marco R. Steenbergen. 2004. *Deliberative Politics in Action: Analysing Parliamentary Discourse*. Cambridge University Press.

Sundberg, R. and E. Melander. 2013. "Introducing the UCDP georeferenced event dataset." *Journal of Peace Research* 50(4):523–532.

United Nations. 2017. "Big Data for Sustainable Development." https://www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html.

USAID. 2018. "México: Derechos Humanos." USAID Del Pueblo De Los Estados Unidos De América.

von Borzyskowski, Inken and Michael Wahman. 2019. “Systematic Measurement Error in Election Violence Data: Causes and Consequences.” *British Journal of Politica Science* .

Weidmann, Nils B. 2015. “On the Accuracy of Media-based Conflict Event Data.” *Journal of Conflict Resolution* 59(6):1129–1149.

Weidmann, Nils B. 2016. “A Closer Look at Reporting Bias in Conflict Event Data.” *American Journal of Political Science* 60(1):206–218.

Wilkinson, Daniel. 2019. “Mexico: The Other Disappeared.” Human Rights Watch. `https://www.hrw.org/news/2019/01/15/mexico-other-disappeared`. Accessed on 8/19/2019.

Wood, Reed M. and Christopher Sullivan. 2015. “Doing Harm by Doing Good? The Negative Externalities of Humanitarian Aid Provision during Civil Conflict.” *The Journal of Politics* 77(3):736–748.

Wood, Reed and Mark Gibney. 2010. “The Political Terror Scale (PTS): A Re-introduction and a Comparison to CIRI.” *Human Rights Quarterly* 32(2):367–400.