Disaggregating Repression: Identifying Physical Integrity Rights Allegations in Human Rights Reports

Rebecca Cordell, K. Chad Clay, Christopher J. Fariss, Reed M. Wood, and Thorin Wright[1]

Most cross-national human rights datasets rely on human coding to produce yearly, country-level indicators of state human rights practices. Hand-coding the documents that contain the information on which these scores are based is tedious and time consuming but has been viewed as necessary given the complexity and detail of the information contained in the text. However, advances in automated text analysis have the potential to streamline this process without sacrificing accuracy. In this research note, we take the first step in creating this streamlined process by employing a supervised machine learning automated coding method that extracts specific allegations of physical integrity rights violations from the original text of country reports of human rights. This method produces a dataset including 163,512 unique abuse allegations in 196 countries between 1999 and 2016. This dataset and method will assist researchers of physical integrity rights abuse because it will allow them to produce allegation-level human rights measures that have previously not existed, and provide a jumping-off point for future projects aimed at using supervised machine learning to create global human rights metrics.

Accepted for publication at *International Studies Quarterly* 17/5/2021

[1] Authors Note: We were supported for this manuscript by the National Science Foundations (award numbers 1627464, 1626775, 1627101). We would also like to thank the editors of *International Studies Quarterly*, two reviewers and Jacqueline Davis for their helpful comments as we worked on this manuscript. We would also like to thank Stephen Bagwell, Shelby Hall, Matthew Rains, and K. Anne Watson for their research assistance.

Introduction

Quantitative datasets on state repression and human rights practices such as the Political Terror Scale (PTS) and the CIRI Human Rights Data Project (e.g., Gibney, et al. 2018; Cingranelli, Richards, and Clay 2014) have traditionally relied heavily on human coders. Such datasets have proven invaluable for scholars investigating cross-national patterns of state and state-sponsored human rights abuses.[2] Yet, largely as a result of the research questions they were developed to address and the time intensive nature of hand-collecting data from extensive amounts of text, these indicators effectively aggregate evidence about specific repressive state practices into a single composite ordinal measure (PTS) or a small number of such measures (CIRI). The categorization processes used to construct the PTS and CIRI measures make use of a substantial amount of qualitative information; however, the variables these processes ultimately produce are rather coarse and retain scant information about the specific actions that inform the measures and/or how human coders used this information. A consequence of the aggregate nature of these indicators is that they are much less informative than the texts used to produce them. In short, the hand-coding process used to generate the measures included in two of the most widely used datasets on state repression result in the exclusion of useful details about the range of human rights abuses that occur within a given country during a year.

Herein we propose an automated allegation extraction technique that facilitates the collection, coding, and retention of data on a number of different dimensions of human rights violations. This process allows us to create more nuanced datasets that capture a variety of behaviors and actions that are currently understudied and largely ignored in cross-national quantitative human rights research. Specifically, we develop an automated method for extracting

[2] Many other datasets rely on the human coding of content of annual human rights reports (e.g., Conrad, Haglund, and Moore 2013; 2014, Eck and Hultman 2007).

physical integrity rights allegations from annual country human rights reports produced by Amnesty International (AI), Human Rights Watch (HRW), and the US State Department (SD) for 1999-2016. We generate a dataset of allegations that will serve to help scholars of repression investigate new questions at a much finer grained level of detail than existing data currently do and overcome some of the acknowledged problems of existing cross-national repression data (e.g. Wood and Gibney 2010; Cingranelli and Richards 2010; Conrad, Haglund, and Moore 2014; Fariss 2014).

Our method has several advantages over existing approaches to data collection from documents. First, our method serves to greatly reduce the time necessary to collect information from human rights reports, producing a dataset of allegations that would take a team of trained researchers years to code by hand. Second, the resulting dataset of allegations lends itself to several new and innovative human rights data projects that move far beyond existing cross-national sources. With limited additional work, our method and data could be used to extract allegations of abuse from a variety of different information sources. Finally, our approach is flexible enough to allow the collection of different types of allegations beyond physical integrity rights.

Motivation

Many of the important empirical patterns related to state repression and human rights abuse identified by scholars derive from analyses of standards-based datasets like the PTS (Gibney, et al. 2018) and the CIRI Human Rights Data Project (Cingranelli, Richards, and Clay 2014).[3] These projects relied on the information contained in annual reports produced by the US State

[3] While "physical integrity rights violations" and "repression" are not perfect synonyms, the use of the term "repression" as a shorthand for physical integrity rights abuse is common in the existing literature (e.g. Poe, Tate, and Keith 1999). As such, we use the term "repression" interchangeably with "physical integrity rights violations."

Department, Amnesty International, and Human Rights Watch.[4] As such, these reports—and the datasets that relied on the information within them—have been invaluable resources on which much of the field's knowledge of cross-national repression has been built.

Scholars acknowledge however, that information contained in these reports does not capture the totality of repressive events (e.g., Hill, Moore, and Mukherjee 2013; Conrad, Haglund, and Moore 2014). Rather, they contain a series of allegations that represent a subset of the overall occurrence of repressive acts over the time periods they cover. Furthermore, the relationship between the actual level of repression and the number of repressive acts alleged by reports is non-constant across space and time. For example, differences in media coverage, non-governmental organization scrutiny, US strategic interests and foreign policy objectives, and other factors influence the convergence (or divergence) of the reported number of repressive events from the true number of such events (Poe, Carey, and Vazquez 2001). As such, the count of allegations is a biased undercount of the true number of repressive acts in any country in any year (Conrad, Haglund, and Moore 2014, 434). Accordingly, some scholars have suggested that attempts to use the data from these reports to produce conclusions about the causes and consequences of repression should either account for this bias in their analysis (Bagozzi, Hill, Moore, and Mukherjee 2015) or use measurement models that treat the reported allegations of repressive acts as *generated* by the unobservable true level of abuse rather than as an unbiased measurement in their own right (Schnakenberg and Fariss 2014; Fariss 2014). As with the teams that code the PTS and CIRI, we confront this challenge when disaggregating the content of the report.

Current measures in existing datasets do not fully leverage all of the information about the occurrence of repression from the annual reports. In part, this reflects the purpose for which these

[4] The PTS includes reports from HRW beginning in 2013.

datasets were created. The creators of standards-based datasets explicitly envisioned them as useful ways to provide annual snapshots or summaries of the human rights conditions within a state. The PTS provided a single composite score that informally captured multiple dimensions of state and state-sponsored physical integrity abuses, including its scope, intensity, and range (see Carlton and Stohl 1981; Wood and Gibney 2010). CIRI added nuance by disaggregating the scope dimensions (e.g., torture, killings, etc.) into individual indicators that could be analyzed separately or combined into a single additive scale. In both cases, coders rely on all of the relevant information contained in the annual reports to assign a country a given score but this information used in each report is not retained. Despite the richness of the information included in the reports, coders only utilize the specific events and behaviors in the reports to inform the broadly descriptive measure(s) included in the dataset. By contrast, the Ill-Treatment and Torture (ITT) dataset (Conrad, Haglund, and Moore 2013; 2014), based on every allegation of torture made by Amnesty International from 1995–2005, highlights the nuance of the information included in the source data. For example, it acknowledges that the reports contain information regarding the identity of victims, the agency or group responsible for the abuse, the location and timing of the abuse, and other potentially useful details about the event.

As ITT illustrates, allegation-level data can serve to both increase our ability to navigate the problem of the biased undercount in human rights reporting and help us produce more detailed disaggregated data on repressive acts. Yet, no existing dataset catalogues or codes all allegations included in the annual reports produced by the State Department, Amnesty International, and Human Rights Watch. In the subsequent sections we therefore describe existing text analysis applications in political science and then introduce an automated procedure to extract allegations of physical integrity violations from annual human rights reports.

Text Analysis and Human Rights

Following recent innovations in machine learning and the increased digitalization of political texts, automated text analysis has become widely used in political science. To overcome the resource costs of analyzing large corpuses of texts by hand, these methods offer opportunities to pursue new research objectives and replicate human coding processes in a more systematic and rapid manner. Scholars have used online blogs, social media data, press releases, political speeches and newspaper articles to sort documents into sentiment and topic-related classes on a range of issues (Hopkins and King 2010; Grimmer and King 2011, Schrodt and Van Brackle 2013, King et al. 2013, Jamal et al. 2015; Windsor 2018).

Text Analysis Methods

Several recent studies use novel automated text analysis techniques to better understand empirical patterns and trends in state respect for human rights (e.g., Fariss et al. 2015, Greene, Park and Colaresi 2019; Park, Greene and Colaresi 2020a; Park, Greene and Colaresi 2020b). However, none of these research projects attempted to replicate the hand-coding approach traditionally employed to produce commonly used human rights measures. To move towards this research objective, we develop an automated classification method that extracts physical integrity rights allegations from annual country reports produced by Amnesty International, Human Rights Watch and the US State Department for 1999-2016. Automated classification programs can be used to organize documents into categories that the researcher defines (supervised approaches) as well as allowing the machine to discover new conceptual structures itself (unsupervised approaches) (Grimmer and Stewart 2013; Lucas et al. 2015).

When the categories are known and specified in advance, dictionary methods provide a straightforward way of sorting and coding texts based on whether content in the text fulfills the pre-specified conditions. Once developed, dictionaries provide a cheap and simple approach to segmenting, organizing and summarizing texts by specifying a list of words that are commonly associated with certain concepts that the researchers are trying to measure such as tone, topic prevalence, or ideology (e.g. see Cordell, Clay, Fariss, Wood and Wright 2020; Murdie, Davis and Park 2020). For this process to be effective, researchers must possess a sufficient substantive knowledge of the conceptual categories, and the meanings of the words and text features specified must match the context in which they are being analyzed.[5] There are high pre-analysis and post-analysis costs to building a dictionary, from selecting the words to mediating the extent to which the model under and over fits the data. Moreover, the model must be thoroughly validated to ensure that the words contained in the dictionary accurately map onto the text and concepts that the researcher is trying to be measure (Quinn et al. 2010; Grimmer and Stewart 2013).

By contrast, supervised learning approaches for classification tasks use hand-coded data (training data) to train a model to replicate the human process of splitting and coding texts into predefined categories. The accuracy of the model can be validated by examining the extent to which it correctly predicts out-of-sample data (test data) that the machine was not exposed to during the training process. Similar to the dictionary-based approached, relevant features of the text are identified by humans in advance but the model itself decides which aspects are more integral for mapping the conceptual categories onto the data (Greene, Park and Colaresi 2019; Park, Greene and Colaresi 2020a; Park, Greene and Colaresi 2020b).

[5]Most dictionaries are custom built according to the topics analyzed and sources used but there are also a series of standardized dictionaries that provide key words for a variety of categories (e.g. Hart 2000).

Finally, unsupervised modeling approaches used to classify texts are run independent of any training data that includes categories known to the researcher or indicates the way a text is structured (Quinn et al. 2010; Bagozzi and Berliner 2018; Potz-Nielsen, Ralston and Vargas 2018). The benefits of employing this method include minimal pre-analysis and the ability to uncover new features within the text that are theoretically useful that the researcher may not have anticipated. However, interpreting and validating the results through experimental, substantive and statistical evidence is an important step to ensure that the model identifies theoretically relevant concepts and correctly selects sections of the text that pertain to the categories (Grimmer and Stewart 2013; Lucas et al. 2015).

Previous research suggests that machine-coding methods can produce the same levels of accuracy as human coded approaches (Bagozzi et al., 2019). While there has been some debate on the superiority of supervised versus unsupervised machine learning methods for analyzing texts (e.g. Hillard et al. 2008), the automated approach adopted should be selected based on its suitability to achieve the desired research objective. Therefore, we use a supervised machine learning approach[6] that uses training data containing 31,061 unique examples of sentences containing physical integrity rights allegations collected by human coders to assign a probability and a binary classification to each sentence on the likelihood of it being a physical integrity rights allegation.[7]

Human Rights Applications

[6] We use a supervised machine learning method as opposed to a dictionary-based approach because it is less likely to produce false positives. For example, a sentence may contain a key word relating to physical integrity rights but may not be a physical integrity rights allegation. By drawing upon examples from our training data of sentences that contain key terms and are indeed physical integrity rights allegations, our supervised machine learning model can more accurately classify which sentences are and are not physical integrity rights allegations based on other (non-dictionary) terms included in the sentences.

[7] These outcomes are interdependent as the binary classification each model produces is derived via the model's probability distribution.

Scholars have previously used machine coding approaches to explore potential biases in annual country reports and uncover lexical patterns in the coding of human rights protection scores. For example, Bagozzi and Berliner (2018) utilize an unsupervised structural topic model to measure the salience of human rights categories over time in the US State Department country reports. They find variation in the prevalence of terms associated with specific human rights topics over time and that prevalence of these terms increases among US military allies, aid recipients, and trade partners. Park et al. (2020a) apply an aspect-based sentiment analysis to identify specific features of hierarchical human rights reports such as positive or negative coverage or intensity of coverage. Park et al. (2019) evaluate how the issues and topics in reports co-evolve over time in human rights documents produced by 18 NGOs and two UN human rights bodies. Their analysis reveals an increase in information over time, the emergence of new topics (including LGBT rights, children's rights and religious freedom), and inter-relatedness of topics. Expanding the focus to Amnesty International and US State Department annual country reports, Potz-Nielsen et al. (2018) use an event data coding approach to sort sentences into allegations related to civil and political rights, economic, social and cultural rights, and physical integrity rights for a small number of countries. As anticipated, they find differences in the amount, density, and type of information found within the documents; highlighting the impact of organizational dynamics on the data generating process. Additionally, Fariss et al. (2015) construct a large document term matrix for all human rights reports produced from 1974-2014 by several reporting agencies to explore word frequency patterns over time and identify the most important key words contained in the reports for the PTS coding, CIRI human rights variables, and Hathaway torture scale coding.

We build on existing automated approaches to measuring human rights by reducing the time necessary to collect information from human rights reports that can be implemented as new

reports are released each year. The resulting dataset of allegations is the first of its kind and lends itself to several new and innovative human rights data projects that move far beyond existing cross-national sources. For example, these allegations (alongside further training data) could automate the classification of violation type, type of actor responsible for the violation (i.e., state versus non-state), location of the violation, and intensity of the violation at the allegation-level. Moreover, the method presented in this paper could also extract allegations of abuse from several different information sources and is flexible enough to allow the collection of different types of allegations beyond physical integrity rights, such as civil liberties, labor rights, worker's rights, or gender rights.

Data and Methods

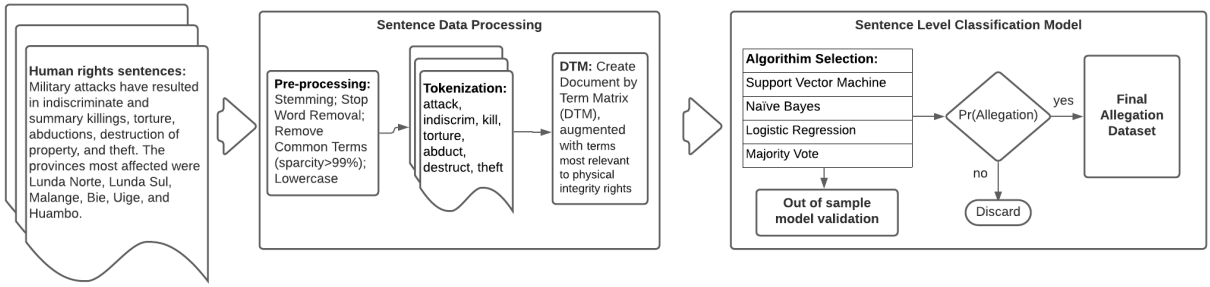
We develop an automated method for extracting physical integrity rights allegations from annual country human rights reports produced by Amnesty International, Human Rights Watch and the US State Department for 1999-2016. Human rights scholars continue to rely upon these texts to develop hand-coded measures of countries' respect for human rights over space and time. We implement a text-as-data approach to replicate the first stage of this hand-coded process and segment the reports into allegations describing physical integrity rights (disappearances, torture, killing, and political imprisonment) by country-year at the sentence level. Future researchers can therefore avoid the costs of extracting allegations directly from the reports by hand and ensure that the information used in any new measures is reproducible. Since the process for extracting allegations is automated, we will also be able to continue digitizing annual human rights reports and producing new allegation data relating to physical integrity rights violations organized by report, country and year as they are released.

Data

Our digital corpus of primary source human rights documents includes the raw text of 7,445 annual human rights reports for 196 countries produced by Amnesty International, Human Rights Watch and the US State Department for 1999-2016. Each report provides a detailed overview of a country's human rights behavior for a given year. They include both general statements that summarize the intensity of human rights violations as well as detailed descriptions of human rights abuses at the event-level. For example, the 2011 Amnesty International report for India summarizes the intensity of abuses in the country by stating that "torture and other ill-treatment, extrajudicial executions, deaths in custody and administrative detentions remained rife" (2011, 166). However, the report also describes numerous specific instances of abuse, such as "[i]n May, Adivasi leader Laxman Jamuda was killed when police fired at people protesting against the acquisition of Adivasi lands for a proposed Tata Steel project in Kalinganagar, Orissa" (2011, 168). The reports are "highly structured" with specific sections relating to different categories of human rights violations that enable a country's performance on human rights to be assessed over time (Fariss et al. 2015, 3).[8] Diagram 1 displays the data processing, model fitting, and evaluation process that we use to generate our sentence level allegation dataset. We describe each of the steps in detail below.

Diagram 1:

[8]Although the changing standards in the accountability of human rights has led to important differences in reports over time including the length of documents, topical attention, spatial focus and language (Clark and Sikkink, 2013; Fariss, 2014).



Data Preprocessing

We begin by pre-processing the documents. First, we add line numbers to all 7,445 ASCII text files to enable users to trace the original location of each physical integrity rights allegation within the text. Second, we segment each text into sentences as most physical integrity rights allegations only contain one sentence.[9] To do this we use the Open Natural Language Processing (NLP) *Maxent_Sent_Token_Annotator* command that uses a probability model to detect the boundaries of sentences.[10] This model is trained on a corpus with annotated sentence boundaries and splits sentences according to the features of words to the left and right of punctuation marks (Reynar and Ratnaparkhi 1997).[11]

Next, we create a line number and identification variable for each sentence and clean the text. We develop regular expression algorithms to correct all strings concatenated or separated in error and remove unnecessary white space, punctuation, numbers, hyperlinks, page formatting language encoded text, non-ascii characters, and stop words (e.g. “are”, “at”, “by”, “from”, that”, “the”). We also convert all text to lower case to streamline the classification process and avoid

[9] Occasionally an allegation contains more than one sentence. To account for this, we create a binary variable in our dataset that indicates whether the sentence before or after contains a term related to physical integrity rights abuse. In addition, users are also able to view the content of surrounding sentences of allegations through our line number variable.

[10] We apply a regular expression algorithm to correct strings concatenated in error by a period.

[11] A punctuation mark is not a sufficient indicator for the end of a sentence. For example, a period can have many different uses including a decimal point, ellipsis, and abbreviation.

duplication for capitalized and small letter versions of the same word. For the same reason, we apply a Porter stemming algorithm to reduce all words to their base root form (e.g. "killed", "kill", and "killing" would all be condensed to "kill") as these terms all capture the same concept of interest (Porter 1980). We use this cleaned version of our corpus for analysis but we also retain the original text to improve readability and enable alternative uses of the allegation data in future research. Finally, we create a variable that counts the total number of words in each sentence to help identify and remove observations in our dataset that are not in fact sentences.[12] Together, these preprocessing steps convert our corpus of text files to 2,013,199 sentences.

Classification Method

Our classification method for identifying physical integrity rights allegations in our corpus of primary human rights documents uses a supervised machine learning approach. First, we train supervised machine learning models on training data collected by human coders that assigns the value of 1 to sentences that the model predicts as a physical integrity rights allegation, and 0 otherwise. Second, we extract probabilities from the best performing machine learning algorithm; assigning a value to each sentence between 0 and 1; with a higher value indicating that the sentence is more likely to include material on physical integrity rights abuse.[13] The probability estimate allows future researchers to determine their own probability threshold for classifying sentences as physical integrity rights allegations.

Supervised Machine Learning Models

[12] We remove sentences containing fewer than two words as they do not contain useful information and are primarily the result of our sentence splitter segmenting a hyperlink into multiple sentences after each period.

[13] In order to transform the SVM model binary classification to probability estimates, we use the Platt scaling method that fits a logistic regression to the SVM models scores to generate probability distributions over classes.

We train supervised machine learning models using human coded training data extracted from the Amnesty International, Human Rights Watch, and US State Department for six pilot countries for 1999-2016 (Angola, Belarus, Mexico, Nigeria, Philippines and the UK) and all countries in 2012. Our training data includes a total of 195,924 sentences, with human coders identifying 16% as physical integrity rights allegations (31,061 sentences), and 84% as non-physical integrity right sentences (164,863 sentences). Table 1 displays the distribution of our training data over space and time. This data was constructed by a team of graduate and undergraduate research assistants that extracted all allegations of physical integrity rights abuse in the annual human rights reports produced by Amnesty International, Human Rights Watch, and US State Department for six pilot countries for 1999-2016 (Angola, Belarus, Mexico, Nigeria, Philippines and the UK), all countries in 2012, and a random sample.

Our pilot countries vary on primary indicators of country-level human rights performance including regime type, economic development, population level, and civil conflict involvement, and have full coverage across Amnesty International and US State department reports (Hill and Jones 2014). By including reports for all countries in 2012, for all years for our pilot countries, and a random sample our data captures variation in the different types of physical integrity rights abuses that take place around the world and accounts for change in language and topical attention over time. Allegation observations in our training data contain the quote from the report, in addition to country, year, report, report page number, and line number from which each allegation was obtained.

Table 1: Training data

Country	Year	Number of Allegation Sentences
Angola	1999-2016	2,456
Belarus	1999-2016	2,889

Mexico	1999-2016	6,259
Nigeria	1999-2016	2,455
Philippines	1999-2016	2,704
United Kingdom	1999-2016	1,144
All countries	2012	12,615
Random sample	1999-2016	539

Physical integrity rights allegations collected by human coders from Amnesty International, Human Rights Watch and US State Department annual country human rights reports. *N = 195,924 sentences.*

In our instructions for coders, we define an allegation, or "informative statement," as a sentence, or group of sentences, that provides information about the enjoyment of physical integrity rights (i.e., freedom from disappearance, extrajudicial killing, torture, ill-treatment, and arbitrary arrest and imprisonment) in the country being discussed in the assigned human rights report. In most cases, an allegation will only contain one sentence; each sentence in a report that contains information about enjoyment of physical integrity rights will be part of an observation in this first stage of data collection. However, some observations in our training data are based on multiple sentences if additional information was necessary to add to or complete the information provided by the first sentence. Consequently, we can always delete extraneous information after it has been captured and have thus far encouraged our research assistants to collect additional text when in doubt. Our human-coded allegation extraction thus likely over captures allegations, including some statements that are effectively uninformative on the level of physical integrity rights abuse. We include our physical integrity rights data collection instructions in Appendix A.

To improve efficiency, we pre-process the clean stemmed version of the text by reducing it to only the most common terms in our corpus (removing less frequent terms whose sparsity is greater than 99%). Because this process removes many terms relevant to physical integrity rights, we create a dictionary of key words and combine these with the common terms for the clean stemmed version of the text. We then create a Document-Term-Matrix (DTM) that records the

word count of common terms and key words that appear in each sentence in our training data (413 unique terms in total) and add a binary variable that indicates whether human coders coded the sentence as a physical integrity rights allegation.

We identify key terms relevant to physical integrity rights using our training data containing 31,061 unique examples of allegations collected by human coders. To identify relevant terms in the training data that are used to describe disappearances, extrajudicial killings, torture and ill-treatment, and political imprisonment and other forms of arbitrary arrest and detainment, we create a DTM that produces a word count for each term included in our training data – with the columns corresponding to the unique words and the rows corresponding to the sentences. This procedure produces a list of the most frequent terms contained for allegations in the training data.[14] We then select terms from this list as physical integrity rights key words based on their relevance to physical integrity rights abuse (e.g. where "abuse" is more relevant than "according").[15] Appendix B displays the unique keywords terms that we extract from the training data (172 terms).

Using the training data DTMs of physical integrity rights key words and the most common terms in our corpus, we use three different supervised machine learning algorithms (Support Vector Machine [SVM], Naïve Bayes and Logistic Regression) to predict which sentences in our dataset are physical integrity rights allegations. We train the models on 80% of the training data (156,740 sentences) using repeated k-fold cross validation. This procedure splits the training data into 5 k-subsets, with each subset removed from the sample and trained on all other subsets – repeated 5 times. Each model finds patterns in the training data that relate word count of key terms

[14] We exclude numbers, stop words, and sparse terms from the text that appear in less than 5% of the allegations.

[15] We add terms that appear frequently in the physical integrity rights allegations in our training data but are not direct physical integrity rights terms in order to better assess the interaction between terms. For example, it may be the case that sentences with the term "kill" are more likely to be physical integrity rights sentences when the term appears with "polic" or "secur" (i.e. when the police and security forces kill individuals as opposed to deaths that take place in communal settings).

and common words contained in a sentence to its binary classification as a physical integrity rights allegation. Sentences that the model classifies as containing information on physical integrity rights are assigned a value of 1; 0 otherwise. We then extract probabilities from the best performing machine learning algorithm to assign a probability to each sentence, with a higher value indicating that the sentence is more likely to include material on physical integrity rights abuse. Appendix C displays examples of how our model codes some sentences in our training data along with the probability of an allegation and binary measure (Final Allegation Dataset), with the key words highlighted bold in the sentence and the word count of common terms and key words in the Document-Term-Matrix.

Results

In order to evaluate the accuracy of the supervised machine learning models, we compare the classes that our models and human coders assign to these sentences in-sample (on 20% of the training data, 39,184 sentences) and out-of-sample (test data from our human rights corpus for India from 1999-2016 and a random sample for 1999-2016, 33,069 sentences). The results from our analysis show that this method achieves between 84-88% in-sample accuracy and 81-84% out-of-sample accuracy. Table 2 and 3 displays the out-sample and out-of-sample accuracy of each model, measured as the mean accuracy over each of the 5 k-sub-samples. The Logistic Regression algorithm has the greatest out-of-sample accuracy in correctly predicting classes of sentences identified by human coders as physical integrity rights allegations at 84.3%. We also use an ensemble method that combines the predictions from each of the models via majority vote.[16] This approach performs slightly better on generating out-of-sample predictions than any one of the

[16] We use a hard voting ensemble that sums the predictions of the SVM, Naïve Bayes and Logistic Regression and predicts the class for each sentence with the most votes. For a similar approach, see Greene, Park and Colaresi (2018).

machine learning algorithms alone. The precision (the ratio of correct positive predictions to the total predicted positives) of this measure is 41%, the recall (the ratio of correct positive predictions to the total number of true positives and false negatives) is 58%, and the F1 score (that takes into account the balance of precision and recall) is 49%.[17] The area under the curve for the Receiving Operator Characteristic (AUC ROC), which reflects ability of the model to separate physical integrity rights allegations from non-physical integrity rights allegations, is 68%. The AUC for the Precision and Recall (AUC PR), indicating the average precision for each recall threshold, is 39%.[18]

Table 2: Out-sample model accuracy with 5x repeated k-fold cross validation

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC ROC	AUC PR
Support Vector Machine	0.883	0.422	0.715	0.531	0.695	0.446
Naïve Bayes	0.842	0.467	0.497	0.482	0.689	0.355
Logistic Regression	0.870	0.356	0.659	0.462	0.661	0.386
Majority Vote	0.876	0.411	0.672	0.510	0.687	0.420

N = 39,184 sentences.

Table 3: Out-of-sample model accuracy with 5x repeated k-fold cross validation

Algorithm	Accuracy	Precision	Recall	F1Score	AUC ROC	AUC PR
Support Vector Machine	0.841	0.421	0.575	0.486	0.677	0.391
Naïve Bayes	0.807	0.493	0.460	0.475	0.684	0.352
Logistic Regression	0.843	0.369	0.593	0.454	0.657	0.379
Majority Vote	0.843	0.415	0.585	0.485	0.675	0.394

N = 33,069 sentences.

The ensemble supervised machine learning approach classifies an additional 1,734 sentences (5% of the out-of-sample test data) as physical integrity rights allegations that our team

[17] *Accuracy* is the proportion of 1s and 0s correctly classify or predicted from the model. *Precision* is the proportion of true positives predictions relative to the total number of positive predictions from the model. *Recall* is the proportion of true positives predictions relative to the total number of true predictions and missed predictions from the model. The *F1 Score* is the harmonic mean of the *Precision* proportion and the *Recall* proportion, which is equivalent to *Accuracy* when the proportion of 1s and 0s are equal.

[18] Area Under the Receiver Operating Characteristic (AUC ROC) is the average of the proportion of true positive predictions for all prediction thresholds from 0 to 1. AUC PRC Area Under the Precision Recall Curve is the average recall proportion for all recall thresholds from 0 to 1.

of human coders did not identify as physical integrity rights allegations. The reason is twofold. On the one hand, there are many sentences that our automated approach successfully identifies as containing information on physical integrity rights abuses on killings, torture, political imprisonment and detention. This result demonstrates how this automated allegation extraction technique can be used to overcome human error associated with identifying physical integrity rights allegations in human rights reports by hand (i.e., missing cases of allegations in the text).[19] On the other hand, there are some sentences that our automated approach identifies as being a physical integrity rights allegation but that are not directly related to the concept including information on human rights investigations and intercommunal conflict. This is because some terms included in our physical integrity rights dictionary have multiple meanings (e.g., "cut", "return", "releas") or only relate to physical integrity rights under some circumstances, with human coders being better at taking in the overall context of sentences (e.g., "die", "violenc", "kill").

Conversely, the ensemble supervised machine learning approach does not classify 3,447 sentences (10% of the out-of-sample test data) as physical integrity rights allegations that our team of human coders do identify as physical integrity rights allegations. First, our human coders over-capturing allegations because we encouraged our research assistants to collect statements when in doubt since we can always delete extraneous information after it has been captured. Second, this could be because our machine coded approach was too strict in its classification of physical integrity rights allegations and unnecessarily excluded sentences containing relevant information. For example, some of the physical integrity rights allegations missed by our model include more general terms in our dictionary such as "humanright", "alleg" and "violat" rather than sentences

[19] While we still consider human coded data as the gold standard, the supervised machine learning method allows us to go above and beyond standard practices used by researchers to account for human error and disagreement among coders (e.g., using multiple trained coders to code each observation).

that contain more specific terms relating to physical integrity rights such as "tortur" and "disappear". Future researchers can overcome this issue by utilizing our probability estimate to lower or raise the threshold for including sentences that might relate to physical integrity rights.

Our ensemble model (the classification approach with the highest accuracy) creates a dataset of physical integrity rights allegations that assigns a binary classification to all 2,013,199 sentences in our data via majority vote from our individual machine learning algorithms. The procedure indicates 8% of sentences in our corpus of human rights documents contain information related to physical integrity rights abuse (163,512). To accompany our physical integrity rights binary variable, we extract probabilities from the logistic regression model (the single algorithm that produces the highest accuracy) that captures the likelihood of it being a physical integrity rights allegation. The probability estimate provides future researchers with the flexibility to determine their own probability threshold for classifying sentences as physical integrity rights allegations. The probability threshold simultaneously produces the greatest accuracy, precision, recall and F1 score is 50%. This is the probability threshold used by the Logistic Regression binary measure to code sentences in our corpus as physical integrity rights allegations (see Table 2 and 3). Researchers can also decrease or increase the probability threshold if they are most interested in increasing precision (a lower false positive rate) or recall (a lower false negative rate). For example, a 45% probability threshold achieves the same level of accuracy (84%) (determined by evaluating the percentage of correctly identified physical integrity rights allegation in the out-of-sample test data) but increases the precision at the cost of decreasing the recall. Conversely, a 55% probability threshold also achieves 84% accuracy but increases the recall at the cost of decreasing the precision. While the predicted probabilities are informative, a good measure should reduce

both the false positive and false negative rate which is why we choose the majority vote binary measure to create our dataset of physical integrity rights allegations.

The majority of physical integrity rights allegations in our new allegation dataset, which were identified by the ensemble majority vote method, are extracted from the human rights reports produced by the US State Department (120,618 allegations), followed by Amnesty International (28,758 allegations), and Human Rights Watch (14,136 allegations). The greatest number of allegations are extracted from reports produced for the year 2001 (12,242 allegations), with the lowest number originating from reports produced for the year 2013 (6,598 allegations). Figure 1 displays the number of allegations identified by our classification model over time, by organization.[20]

We illustrate the top 10 countries with the greatest number of allegations identified by our model over the entire period (1999-2016). We should note that the total number of allegations is not a measure of the relative severity of one country compared to another but rather the relative level of monitoring that each country receives in these particular reports.[21] These countries include Israel (3,870 allegations), India (3,429 allegations), Russia (3,425 allegations), Colombia (3,102 allegations), Democratic Republic of Congo (2,943 allegations), China (2,923 allegations), Mexico (2,873 allegations), Pakistan (2,795 allegations), Iraq (2,560 allegations) and Indonesia (2,556). The country with lowest number of allegations extracted from reports are is Tuvalu (63 allegations) and the average number of allegations extracted is 834 (e.g., Gambia). Figure 2 displays the total number of allegations identified by our classification model, by country.[22]

[20] An annual report was not produced by Human Rights Watch in 2003 or Amnesty International in 2013.
[21] While the United States receives a large amount of coverage in the AI and HRW reports, it is not covered at all by the USDS reports, which provides the vast majority of global allegations. This greatly reduces the US's relative coverage overall.
[22] Country reports for China contain allegations for Macau, Hong Kong, Taiwan, and Tibet; country reports for Israel contain allegations for Palestine and the occupied territories; country reports for Morocco contain allegations for

Figure 1: Number of physical integrity rights allegations over time by organization

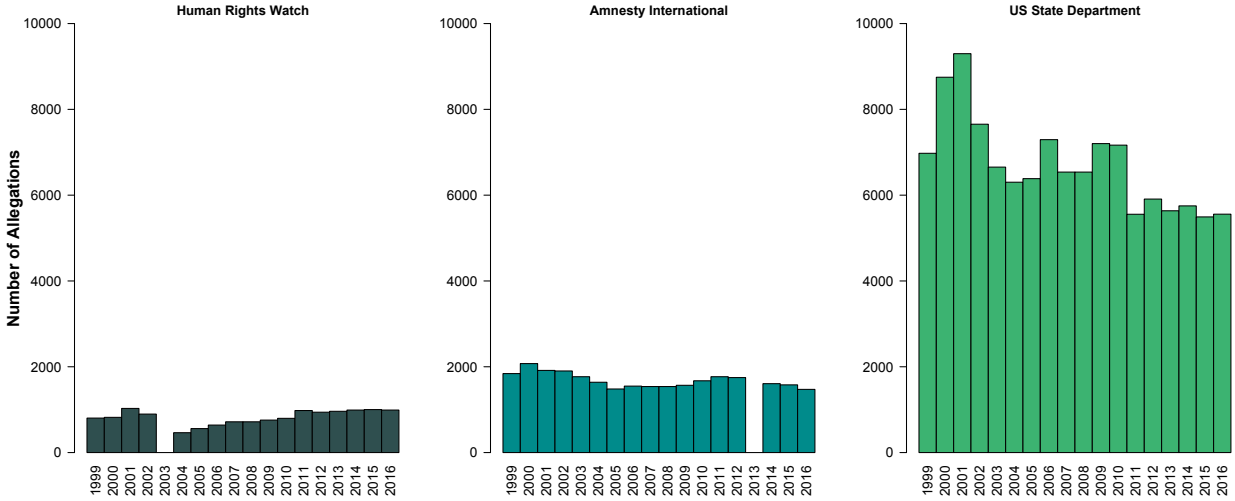
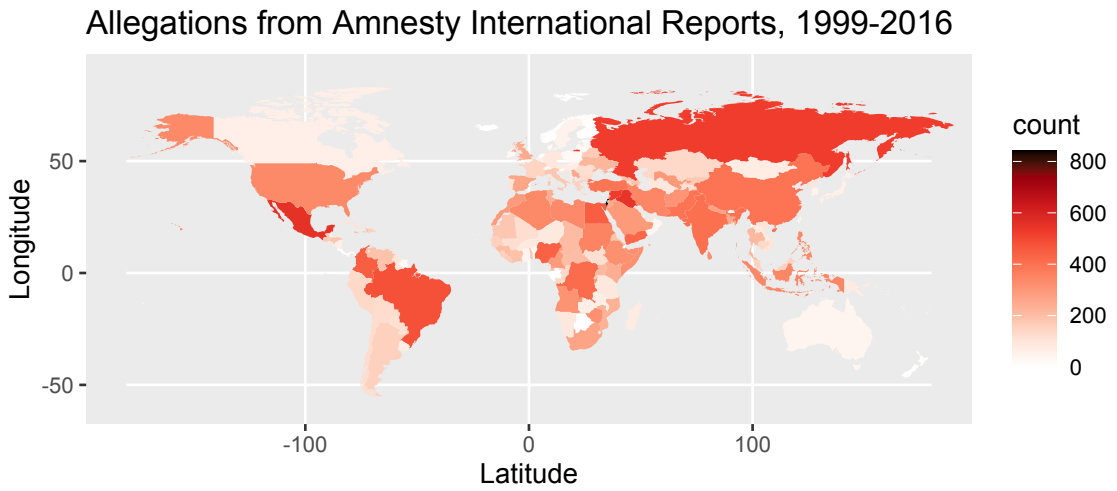


Figure 2: **Number of allegations by country**



Western Sahara; and country reports for the US contain allegations for Puerto Rico. See an aggregate world Map in Appendix D.

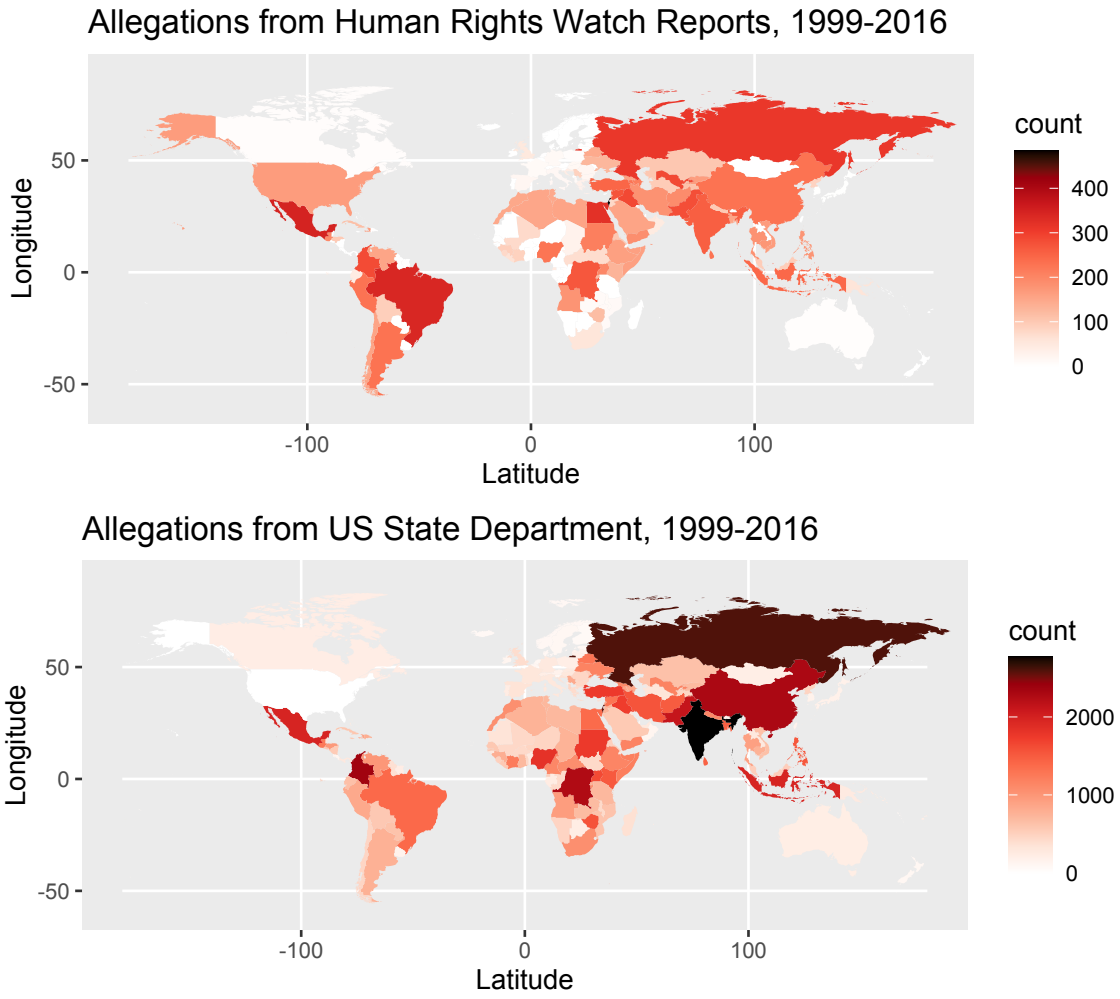


Table 4 displays the top 20 physical integrity rights terms with the greatest frequency in the allegation dataset. The term with the highest frequency is "kill" (mentioned 44,227 times), the term with the lowest frequency is "criminalit" (mentioned once) and the average number of times a term occurs is 3,839 (e.g. "mistreat"). The average number of key words in an allegation is 4 terms, with the minimum of 0 and the maximum of 43. We use allegations and term counts in a series of country-year level validation assessments in Appendix E.

Table 4: Top 20 keywords with the greatest frequency in the allegation data

Keyword	Frequency

kill	44227
arrest	34876
tortur	29927
forc	29641
detent	21075
prison	19978
secur	19498
arbitrari	17765
alleg	17746
beat	17367
detain	17262
humanright	15839
abus	14161
charg	14107
attack	13754
death	11984
disappear	11700
suspect	10957
sever	10054
treatment	9909

Top 20 keywords with the greatest frequency in the allegation sentences. We use these allegations and term counts in a series of country-year level validation assessments in Appendix E.

Conclusion

In order to capture physical integrity allegations contained in three sets of human rights monitoring reports, we employed a supervised learning approach to extract and verify the allegations. Existing data projects have long relied on human coders to produce annual standards-based scores on country-level repressive activities. In order to overcome the costs of analyzing large corpuses of texts by hand, these methods enable us to replicate human coding processes in a more systematic and timely manner, with high degrees of accuracy. Indeed, it took approximately 4,000 person-hours to hand code the training data used to produce our dataset. At that rate, obtaining the same data using human coders would have taken somewhere around 44,000 person hours overall.

The sentence allegation data are useful for several purposes. By developing a more fine-grained approach to gathering data, such as pulling out the allegations and assembling them into a dataset as we have done here, other researchers and teams will be able to investigate and generate their own estimates for new questions related to physical integrity abuse (see additional validation evidence in Appendix E). Researchers may also employ similar methods to generate allegations of abuse that pertain to different categories of human rights beyond physical integrity. This is because the sentence level allegations from human rights monitoring reports are the key building blocks for all standards-based human rights variables. By systematically identifying these sentence level allegations within each report, any measurement project can train human coders or an algorithm to focus exclusively on these allegations when developing variables for the country-level or other units of analysis.

Our new data and approach contribute to the study of human rights violations, which suffer from an information problem stemming from incomplete reporting, thus resulting in event counts that are usually biased under counts.[23] Because of this, using standards-based information from monitoring reports makes sense for measuring human rights performance in relation to other source of event-based information (e.g., Fariss, Kenwick, and Reuning 2020). However, prior attempts at directly measuring human rights performance with these reporting sources are also not leveraging the level of detail contained in these reports, limiting the applications of variables generated from the reports. We systematically leverage information at a level of detail not

[23] This is an active area of research (e.g., Cordell et al 2020; Fariss 2014; Fariss Kenwick, and Reuning 2020; Greene, Park and Colaresi 2019; Park, Greene and Colaresi 2020a; Park, Greene and Colaresi 2020b) that our new dataset makes a substantively important and practical contribution to.

previously used. With sentence level information, we can estimate new standards-based variables of human rights performance in a transparent and reproducible way.[24]

We also provide some cautionary advice for future users of this sentence level dataset. First, our measurement approach and data provide researchers with the ability to shrink or expand the concept and resulting measure. For example, scholars could take a more expansive view of physical integrity rights and identify an expanded set of sentences for their measure. On the other hand, scholars could take a stricter legal view for their concept and measure. Second, we urge scholars not to use counts from the reports as direct measures of abuse. Instead, we advise that measurement models be used to both account for uncertainty and make different levels of coverage comparable across cases. The reports themselves are not a census of events and our sentence level data is not either. Any measures created using the sentence level data should be thoroughly validated. Third, we do not currently identify current and past references to human rights violations, which is additional information that future research may wish to identify and incorporate into measurement models that combine these sentence level data. Finally, we hope that by identifying and classifying the sentence level data from the Amnesty International, Human Rights Watch and US State Department reports, we will be able to extend our approach to a much broader set of information sources.

In future research we plan to use these allegations alongside additional training data to automate the classification of violation type, type of actor responsible for the violation (i.e., state versus non-state), location of the violation, and intensity of the violation at the allegation-level. Alongside other automated coding processes, event data coding (who did what to whom, when

[24] All existing human rights coding projects are in principle reproducible. However, given the number of person-hours necessary to replicate these large county-year datasets, such replications have not occurred in practice. Our approach systematically automates a major task of the human coding process making future replications and extensions much more cost effective and time efficient.

and where). This approach can be used to develop a variety of new data for human rights researchers seeking to pull allegations from new documents or to extract different types of allegations from our sources.

References

Amnesty International. 2005. *Amnesty International Report 2005 The State of the World's Human Rights*. Available at: https://www.amnesty.org/en/documents/pol10/0001/2005/en/.

Amnesty International. 2011. *Amnesty International Report 2011 The State of The World's Human Rights*. Available at: https://www.amnesty.org/en/documents/pol10/001/2011/en/.

Bagozzi, Benjamin E and Berliner, Daniel. 2018. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports". *Political Science Research Methods* 6(4): 661-677.

Bagozzi, Benjamin E, Brandt, Patrick T, Freeman, John R, Holmes, Jennifer S, Kim, Alisha, Mendizabal, Agustin Palao and Potz-Nielsen, Carly. 2018. "The Prevalence and Severity of Underreporting Bias in Machine-and Human-Coded Data". *Political Science Research Methods* 7(3): 641-649.

Bagozzi, Benjamin E, Daniel W. Hill Jr, Will H. Moore and Bumba Mukherjee. 2015. "Modeling Two Types of Peace: The Zero-inflated Ordered Probit. ZiOP. Model in Conflict Research". *Journal of Conflict Resolution* 59(4): 728-752.

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 Sikkink, Kathryn. 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 M. Wright. 2020. "Changing Standards or Political Whim? Evaluating Changes in the Content of US State Department Human Rights Reports Following Presidential Transitions" *Journal of Human Rights* 19(1): 3-18.

Eck, Kristine, and Lisa Hultman. 2007. "Violence Against Civilians in War." *Journal of Peace Research* 44 (2): 233–46.

Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability". *American Political Science Review* 108(2):297–318.

Fariss, Christopher J, Linder, Fridolin J, Jones, Zachary M, Crabtree, Charles D, Biek, Megan A, Ross, Ana-Sophia M, Kaur, Taranamol, Tsai Michael. 2015. “Human Rights Texts: Converting Human Rights Primary Source Documents into Data”. *PLOS ONE*: 1-19.

Fariss, Christopher J., Michael R. Kenwick, and Kevin Reuning. 2020. “Estimating one-sided-killings from a Robust Measurement Model of Human Rights” *Journal of Peace Research* 57(6):801-814.

Gibney, Mark, Linda Cornett, Reed M. Wood, and Peter Hascke. 2018. *The Political Terror Scale*. Available at: www.politicalterrorscale.org.

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: 223-230.

Grimmer, Justin, and King, Gary. 2011. “General Purpose Computer-Assisted Clustering and Conceptualization”. *Proceedings of the National Academy of Sciences* 108(7): 2643–2650.

Grimmer, Justin and Stewart, Brandon M. 2013. “Text-as-data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts”. *Political Analysis* 21(3): 1-31.

Hart, Roderick P. 2000. *Diction 5.0: The Text Analysis Program*. Thousand Oaks, CA: Sage-Scolari.

Hill, Daniel W. Jr, Will H. Moore and Bumba Mukherjee. 2013. “Information Politics Versus Organizational Incentives: When Are Amnesty International's “Naming and Shaming” Reports Biased?” *International Studies Quarterly* 57(2): 219–232.

Hill, Daniel W. Jr and Zachary M. Jones. 2014. “An Empirical Evaluation of Explanations for State Repression”. *American Political Science Review* 108(3): 661-687.

Hillard, Dustin, Purpura, Stephen and Wilkerson, John. 2008. “Computer-Assisted Topic Classification for Mixed-Methods Social Science Research”. *Journal of Information Technology & Politics* 4(4): 31-46.

Hopkins, Daniel and King, Gary. 2010. “Extracting Systematic Social Science Meaning from Text”. *American Journal of Political Science* 54(1): 229-247.

Human Rights Watch. 2015. *Human Rights Watch World Report 2015 Events of 2014*. Available at: https://www.hrw.org/world-report/2015.

Jamal, Amaney A, Keohane, Robert O, Romney, David and Tingley, Dustin. 2015. “Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses”. *Perspectives on Politics* 13(1): 55-73.

King, Gary, Pan, Jennifer and Roberts, Margaret E. 2013. “How Censorship in China Allows Government Criticism but Silences Collective Expression”. *American Political Science Review* 107: 1-18.

Lucas, Christopher, Nielsen, Richard A, Roberts, Margaret E, Stewart, Brandon M, Storer, Alex and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics". *Political Analysis* 23: 254-277.

Murdie, Amanda, David R. Davis and Baekkwan Park. 2020. "Advocacy Output: Automated Coding Documents from Human Rights Organizations. *Journal of Human Rights* 19(1): 83-98.

Park, Baekkwan, Amanda Murdie and David R Davis. 2019. "The Co(evolution) of Human Rights Advocacy: Understanding Human Rights Issue Emergence Over Time". *Cooperation and Conflict* 54(3): 313-334.

Park, Baekkwan, Kevin T. Greene, Michael Colaresi. 2020. "Human Rights are (Increasingly) Plural: Learning the Changing Taxonomy of Human Rights from Large-scale Text Reveals Information Effects" *American Political Science Review* 14(3): 888-910.

Porter, Martin F. 1980. "An Algorithm for Suffix Stripping". *Program* 14(3):130–137.

Potz-Nielsen, Carly, Ralston, Robert, Vargas, Thomas R. 2018. "Recording Abuses: How the Editing Process Shapes our Understanding of Human Rights Abuses". *Working paper*: Available at: https://3b6611d6-c4e8-4edd-a036-4479a2b74783.filesusr.com/ugd/c245b0_42173ba3ab884b659924d31ef90a3e7b.pdf.

Quinn, Kevin M, Monroe, Burt L, Colaresi, Michael, Crespin, Michael H and Radev Dragomir, R. 2010. "How to Analyze Political Attention with Minimal Assumption Costs". *American Journal of Political Science* 54(1): 209-228.

Reynar, Jeffrey C. and Adwait Ratnaparkhi. 1997. "A Maximum Entropy Approach to Identifying Sentence Boundaries". In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Available at: https://www.aclweb.org/anthology/A/A97/A97-1004.pdf.

Schnakenberg, Keith E. and Christopher J. Fariss. 2014. "Dynamic Patterns in Human Rights Practices." *Political Science Research and Methods.* 2(1): 1-31.

Schrodt, Philip A and David Van Brackle. 2013. "Automated Coding of Political Event Data". In *Handbook of Computational Approaches to Counterterrorism*, ed. V.S. Subrahmanian. New York: Springer Press.

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

US State Department. 2000. *Country Reports on Human Rights Practices Bureau of Democracy, Human Rights and Labor 1999 – Angola.* Available at: https://2009-2017.state.gov/j/drl/rls/hrrpt/1999//index.htm

Windsor, Leah. 2018. The Language of Radicalization: Female Internet Recruitment to Participation in ISIS Activities. *Terrorism and Political Violence*: 1-33.