Web Appendix for "Human Rights are (Increasingly) Plural: Learning the Changing Taxonomy of Human Rights from Large-scale Text Reveals Information Effects"

Baekkwan Park[1], Kevin Greene[2], and Michael Colaresi[3]

[1]Assistant Professor , East Carolina University , `baekkwan.park@gmail.com`
[2]PhD Candidate , University of Pittsburgh , `ktg19@pitt.edu`
[3]William S. Dietrich II Professor , University of Pittsburgh , `mcolaresi@pitt.edu`

March 23, 2020

A Details of Sentiment Analysis

To measure the overall sentiment of the State Department reports over time we utilize both a dictionary-based and a supervised learning approach. For the dictionary sentiment analysis we use a pre-made dictionary that labels the positive or negative sentiment for a large list of words. As there is no sentiment dictionary specific to human rights, we use the sentiment dictionary AFINN-111 (Nielsen, 2011). The AFINN-111 dictionary contains a sentiment score for thousands of words.[1] To measure sentiment we begin at the sentence level and sum the sentiment scores (1,-1) of all the words in a sentence. If the sum is greater than 0, the sentence is classified as positive. If it is less than 0, it is classified as negative. We then take the average of these sentiment scores across all the sentences in a given year. To provide a more concrete example take the sentence "The government welcomed and regularly granted visas to international ngos and other human rights monitors including members of amnesty international and human rights watch." Here only the words "welcomed" and "granted" are included in the the AFINN sentiment dictionary. Because both are labeled as denoting positive affinity and there are no words with negative affinity in the sentence, the sentence is classified as being positive. This process would be repeated for each sentence in the corpus.

The classifier method, which utilizes a support vector machine (SVM), uses the term frequency–inverse document frequency (tf-idf) counts of the words (represents as bigrams) in each sentence to learn the mappings between the word features and the sentiment scores. To train our sentiment classifier, we first randomly sample 4000 sentences from the State Department Reports and code them for positive (1), neutral (0), or negative (-1) judgments on state human rights practices. We then create a document-term matrix where each sentence is represented as the term frequency–inverse document frequency (tf-idf) counts of the words (represents as bigrams) in each sentence and a corresponding handcoded sentiment value. Using these sentences we learn the mappings from language in the text, to sentiment scores. After tuning the model using cross validation, we use the highest performing model to predict the sentiment scores for the remainder of

[1] The dictionary can be found here http://corpustext.com/reference/sentiment_afinn.html

the unlabeled sentences. From here we calculate the average sentiment by taking the mean of the expected value of the sentiment scores[2] from all the sentences in a given year.

A.1 Annotating Sentiment

As noted above the sentiment coding was conducted by randomly selecting 4000 sentences from our corpus and then assigning each sentence a value of -1 (negative), 0 (neutral), 1 (positive). Sentiment values were assigned using the following criteria.

Negative Sentiment

1. The text refers to clear ineffectiveness in protecting an aspect of human rights. ("The Cambodian human rights committee, which the government established in 1998, largely was inactive throughout the year, and its activities were not credible.")
2. The text refers to clear violations of an aspect of human rights ("There were instances of arbitrary arrests and detention.")

Positive Sentiment

1. The text refers to clear support for an aspect of human rights. ("The government welcomed and regularly granted visas to international ngos and other human rights monitors, including members of amnesty international and human rights watch.")
2. The text is clear there was no restriction of an aspect of human rights ("The government did not refuse visas to international NGO human rights monitors.")

[2] For a sentence that was classified with probability .8 for 1, .1 for -1 and .1 for 0 the expected sentiment would be $(.8 * 1) + (.1 * -1) + (.1 * 0) = .7$

Neutral Sentiment

1. The text refers to a simple fact, rather than a judgment on human rights ("Baha'i, Christian, Zoroastrian, and Jewish communities constitute less than 1 percent of the population.")
2. The text refers to a possible restriction on human rights but that is unclear without additional context ("The ministry of defense may ban works about sensitive security issues.")

A.2 Results

We use the 4000 annotated training sentences to train a support vector machine (SVM). This machine learning model is compared against the sentiment dictionary AFINN. The accuracy of the dictionary approach is 46% compared to 61% for the classifier. Full evaluation metrics are presented below.

	Precision			**Recall**			**F1-Score**		
Label	**-1**	**0**	**1**	**-1**	**0**	**1**	**-1**	**0**	**1**
	0.64	0.50	0.65	0.70	0.49	0.57	0.67	0.49	0.61

Table A.1: Evaluation Metrics for Sentiment Classifier

B Measuring Latent Available Information Density Over Time

Previous work has theorized that the increasingly dense information available to human rights organizations has in turn led to additional violations being recorded and then composed in reports released by the State Department and other human rights groups (Clark and Sikkink, 2013; Fariss, 2014). In the main text of the paper we extend existing theory and find evidence that not only are more rights being judged in later years, but that the rights being judged are on more specific aspects of existing rights. Similar to previous work we suggest that increases in available information are driving these changes. While previous work has laid out compelling theoretical explanations for this process and noted examples from particular cases, thus far no work has explicitly modeled the changes in the information environment that has been expected to be making this denser information available over time. In particular, the spread of information communication technology (ICT) such as the number of Internet users or individuals with mobiles phones, provides additional opportunities for NGOs to collect evidence on human rights activities.

We are interested in measuring both changes in the available information density and the deepening of taxonomies of human rights judged in texts. While our paper and other parts of this appendix detail how we extract the implicit evolving taxonomy of rights from texts, here we detail our new measure of available information density.

B.1 Indicators

To build this measure we gather a variety of indicators of ICT globally for the years 1977-2014. The first variable is the number of mobile cellular subscriptions per capita calculated drawn from the International Telecommunication Union, World Telecommunication/ICT Development Report and database, retrieved from the World Bank.[3] The spread of cell phone use not only allows calls from previously unreachable locations, but recent innovations such as Ushahidi, which following electoral violence in Kenya, developed a web-platform where users can report the location of violence or election interference, allows for new and more specific information to be collected that may have been overlooked in the past.

The second variable is the number of individuals using the Internet per capita, drawn from the International Telecommunication Union, World Telecommunication/ICT Development Report and database.[4] Access to the Internet opens a number of channels for increased human rights relevant information. In recent years the internet has been used to organization revolutions in Egypt and Tunisia, as well as providing a new outlet for the expression of social and political rights the world over.[5]

[3] International Telecommunication Union, World Telecommunication/ICT Development Report and database, accessed June 15, 2019.

[4] International Telecommunication Union, World Telecommunication/ICT Development Report and database, accessed June 15, 2019.

[5] https://www.hrw.org/world-report/2017/country-chapters/the-internet-is-not-the-enemy

The third variable is the number of monthly active Facebook users worldwide calculated by Facebook Quarterly Earnings Slides Q1 2019, from Statistica, for years after 2008 and from the Wall Street Journal for years before 2008.[6] The number of users on Facebook increases the potential for a story that would have been contained locally, to spread though social media receiving global coverage. An instance of a Libyan military commander ordering a summary execution of ten men that was posted on Facebook, gained attention from a wide reaching audience, including the International Criminal Court, who issued a warrant for his arrest.[7]

The fourth variable is the inverse of the highest panchromatic resolution of non-military earth-imaging satellites (in cm) calculated from the Satellite Imaging Corporation.[8] In recent years human rights organizations have increasingly relied on satellite-imaging to document human rights abuses, particularly in locations where placing investigators is restricted by authoritarian governments.[9] In the past the technology was used in Croatia and Bosnia to find locations of mass graves, while more recently in Syria satellite-images where used to uncover a crematorium built to dispose of evidence of human rights abuses.[10] As the resolution of satellites improves higher quality images can be captured, providing stronger evidence of rights violations.

B.2 A Bayesian Latent Variable Model of Available Information Density

Let T index the number of time periods we are interested in, and P the number of observable proportions. $\mathbf{X}$ is then a $T \times P$ matrix of our observable measures. We transform each column j of $\mathbf{X}_j$ into $\mathbf{Z}_j$ using the logit transform, and then standardize the columns. We denote $\mathbf{Z}$ as the matrix of transformed inputs on the standardized log-odds scale.[11]

We assume that θ_t follows a random walk and projects into our observed measures through a

[6] https://www.wsj.com/articles/facebooks-timeline-15-years-in-11549276201

[7] https://www.buzzfeednews.com/article/meghara/facebook-youtube-icc-war-crimes

[8] The raw data is available here https://www.satimagingcorp.com/satellite-sensors/. We created a maximum resolution across available satellites for each year in our data.

[9] https://www.hrw.org/news/2017/11/30/new-satellite-imagery-partnership

[10] https://www.scientificamerican.com/article/how-satellite-images-can-confirm-human-rights-abuses/

[11] In practice, for numerical stability we use a small epsilon value to ensure we are not dividing by 0 or taking the log of 0.

loading matrix $beta$, which is a row vector of length P. We define the model and priors as:

$$\begin{aligned}
Z_t &\sim N(\beta\theta_t, \sigma_z) \\
\sigma_z &\sim \frac{N}{2}(0, \sigma_c) \\
\beta &\sim N(.5, .5) \\
\theta_t &\sim N(\theta_{t-1}, \sigma_\theta), \ \forall \ 1 < t \leq T \\
\theta_1 &\sim N(\mu_0, .01)
\end{aligned}$$

The prior for σ_c encodes plausible values for the measurement noise (σ_z) for each series. The priors for β encode the knowledge that we expect all the loadings to be positive. μ_0 is set such that latent available information is likely to be lowest in the first year being estimated. These priors for β and μ_0 identify the model. Specifically, $\mu_0 = -8$, which is on the log-odds scale, $\sigma_\theta = 1$, and $\sigma_c = 1$. The estimated θ's are then rescaled such that the minimum value is 0 and the maximum value is 1. We use Stan version 2.19 (Carpenter et al., 2017) and PyStan 2.19.0.0 for this analysis (Stan Development Team, 2018). All diagnostics, including traceplots and the number of effective sample size do not turn up any flags. Further, there were no divergences or other warnings. We present bivariate plots of the posterior samples for β and σ_z. Both are vectors of length P. We can see that there are not strong correlations in the posteriors except for between the zeroth series (which is mobile accounts) and the first series (which is internet access). A future version of the model could leverage these correlations with non-independent priors.

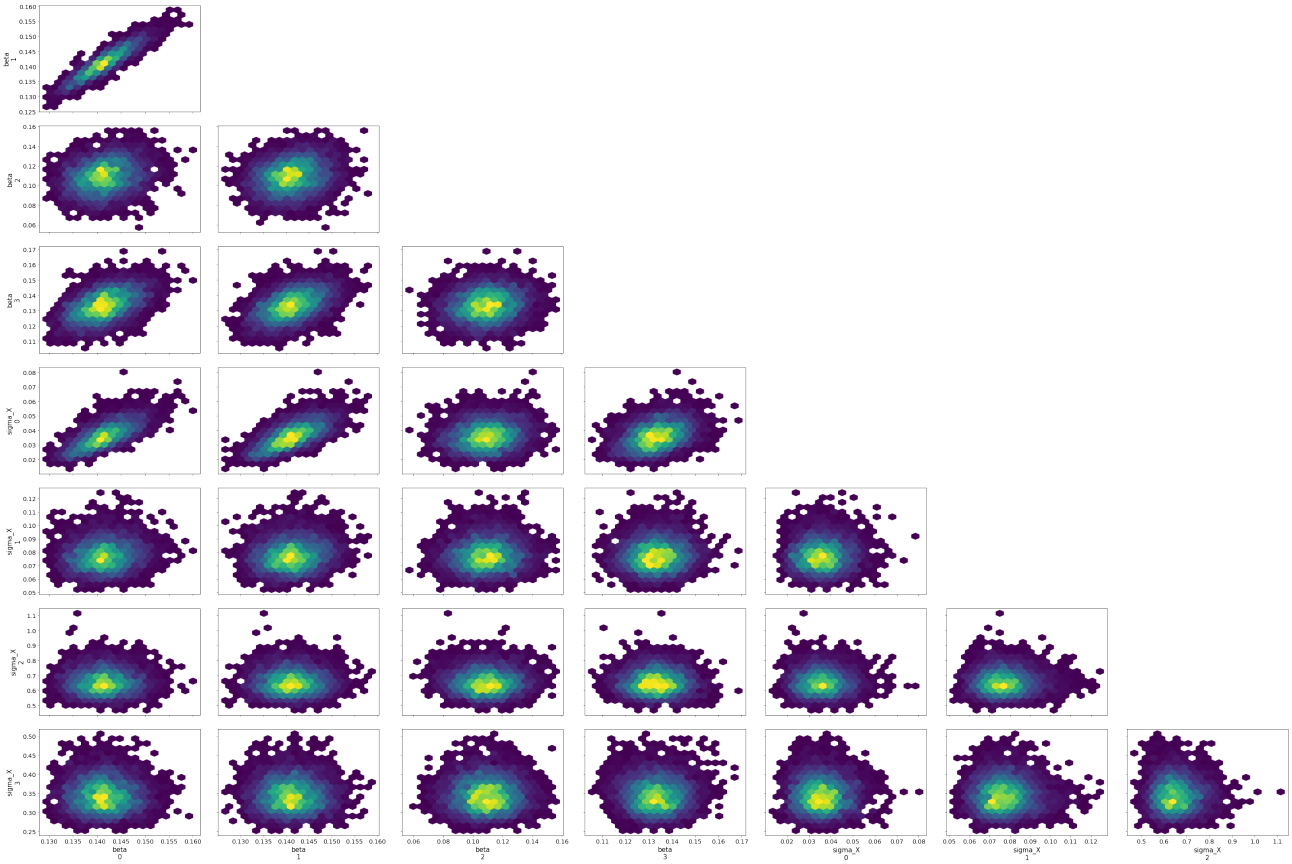


Figure B.1: Bivariate density plot of posterior distributions for β and σ_z

A plot our latent available information density over time is presented below. It is important to note that our latent variable model takes into account the cumulative effects of the informational measures, as opposed to simply being an average across them. We have set up a github repo that will allow researchers to use and extend our model and measures.

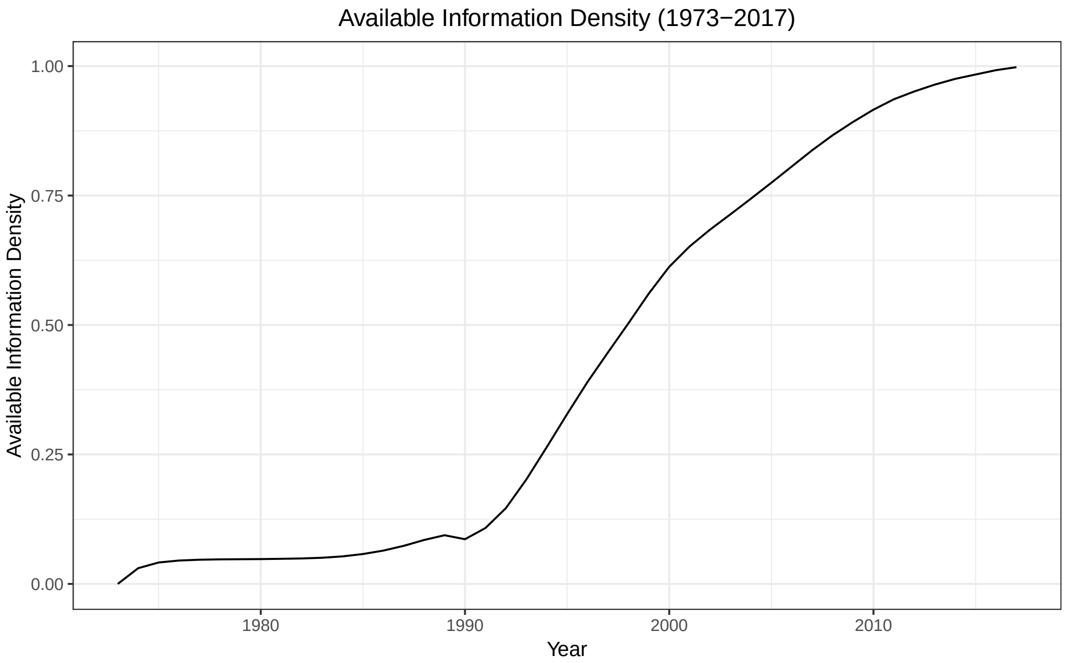


Figure B.2: Our Latent Available Information Density measure over time.

C Information Communication Technology and Bureaucratic Change

We view our work as being complimentary to previous work assessing changes in the bureaucracy of the State Department. In particular past work has treated the State Department reports similarly to those compiled by human rights NGOs, and generally conducted their analyses on these reports or used the State Reports as evidence in support of their theory. This is partly due to that fact that the State Department themselves rely on human rights NGOs as the primary sources used to compile their reports. Thus as there are greater numbers of NGOs and they focus on a greater number of violations, there are in turn more detailed source information for the State Department to use to compile their own reports. Previous work has corroborated this finding, for instance Bagozzi and Berliner (2016) find that more recent State Department reports focus on a greater number of human rights issues, while Cordell et al. (2019), citing Fariss (2014), explicitly state that the reporting and information gathering of the State Department reports have improved over time.

Cordell et al. (2019) theorize that the State Department reports are longer, more detailed, and cover more aspects of human rights due to bureaucratic inertia, increased information availability, and a changing standard of accountability. That is, these forces work in conjunction. Thus while the administration in power in the US effects the coverage of some aspects of human rights, our theory is compatible with that presented by Cordell et al. (2019). In particular bureaucratic inertia can lead to reports that are longer over time, while an administration's preferences may alter the content of the reports, but without more information available to be encoded in the reports and a changing standard that makes more fine grained human rights violations more salient, we would not observe the increased focus on fine grained aspects of human rights in later years.

To further validate that the results we present in the main paper are not simply driven by bureaucratic effects we evaluate the predictive performance of our latent available information density (AID) measure against other potential factors leading to increased complexity in human rights reporting. The first is our latent AID measure, the second, to account for bureaucratic effects, is the administration model (Admin) which includes indicators for the US Presidential in power, the third model simply represents the performance of a model that uses linear time as an input feature, finally, because changes in AID and bureaucracy may very well work in tandem to change the content of human rights reports we interaction the AID and the Admin variables. In each model we aim to use these yearly features to predict the target, the average implicit node depth for that year, estimated from our model in the main paper. We also conduct an additional test where the model instead predicts the implicit node depth of the following year.

The comparison is conducted using leave-one-out cross validation. Leave-one-out cross validation works by dividing the data into N mutually exclusive partitions, the model is fit on all but one observation, and the remaining observation is then predicted. This is repeated until every observation has been predicted. The performance metrics shown are mean squared error (MSE) and mean absolute error (MAE). In both cases, lower scores represent better performance. All of the trained models are fit with ordinary least squares. Across both metrics we see that a model fit with our latent available information density measure produces superior predictive performance relative to a model only accounting for time or bureaucracy. This accords with the narrative in Apodaca (2019) who suggests that bureaucratic changes after 1977 have not had step effects on the content of the reports. Instead, there has been an evolution in improved accuracy, including in the collection of relevant information. However, as noted by Cordell et al. (2019) and consistent with the theory presented in the main text, a model including both AID, Admin, and their interaction leads to the best predictive performance. Thus while changes in human rights reporting are not driven solely by administration effects, they do seem to have an influence. One interesting avenue for future research would be to more deeply understand these connections.

	1 step ahead		2 step ahead	
	MSE	MAE	MSE	MAE
AID	.045	.167	.039	.159
Admin	.073	.206	.080	.217
Time	.068	.244	.070	.230
AID*Admin	.027	.133	.027	.132

Table C.1: Predictive performance using leave-one-out cross validation based on mean squared error and mean absolute error. The target implicit node depth is led 1 and 2 time periods.

D The US State Department Annual Human Rights Reports Corpora with Explicit Taxonomic Meta-data, 1977-2016

D.1 Data

The data for our analyses are taken from the State Department's Annual Country Reports on Human Rights Practices. The reports cover a wide variety of civil, political, and economic rights. The reports are required to be created each year based on the Foreign Assistance Act of 1961 and the Trade Act of 1974. According to the State Department they are among the most widely read US Government document each year, and have an impact on the allocation of foreign aid, asylum cases, and are one means of the US laying out it's human rights priorities.

The documents from 1999-2016 were scraped from the State Department's website. For the period 1977-1998, we use the documents from Fariss et al. (2015). Because the earlier documents are based on optical character recognition (OCR) scans of the primary documents, we have spent considerable time correcting thousands of errors in the documents.

D.2 Human Rights Reports and Meta-Data

While there are fine-grained labels in later years, the most specific in 2015/2016 (See Table D.1), earlier years only label the approximate general location. To ensure comparability over time we have taken several steps. First, as the exact name for each category has slightly changed over time we match the different names from the early reports to the category names of the 2015/2016 categories as much as possible. For example, in 1985, "*Political Killing*" was used to describe extrajudicial killings, but in 2016, the report uses "*Arbitrary Deprivation of Life and other Unlawful Politically Motivated Killings*" instead. In order to make these two different names consistent, we label it as "*Extrajudicial Killing*." The 1982 report uses "*Invasion of the Home*" as the header for the section handling arbitrary and unlawful searches of homes. The 2015 report calls it "*Arbitrary Interference with Privacy, Family, Home, or Correspondence*." We tag both of them as "*Privacy*."

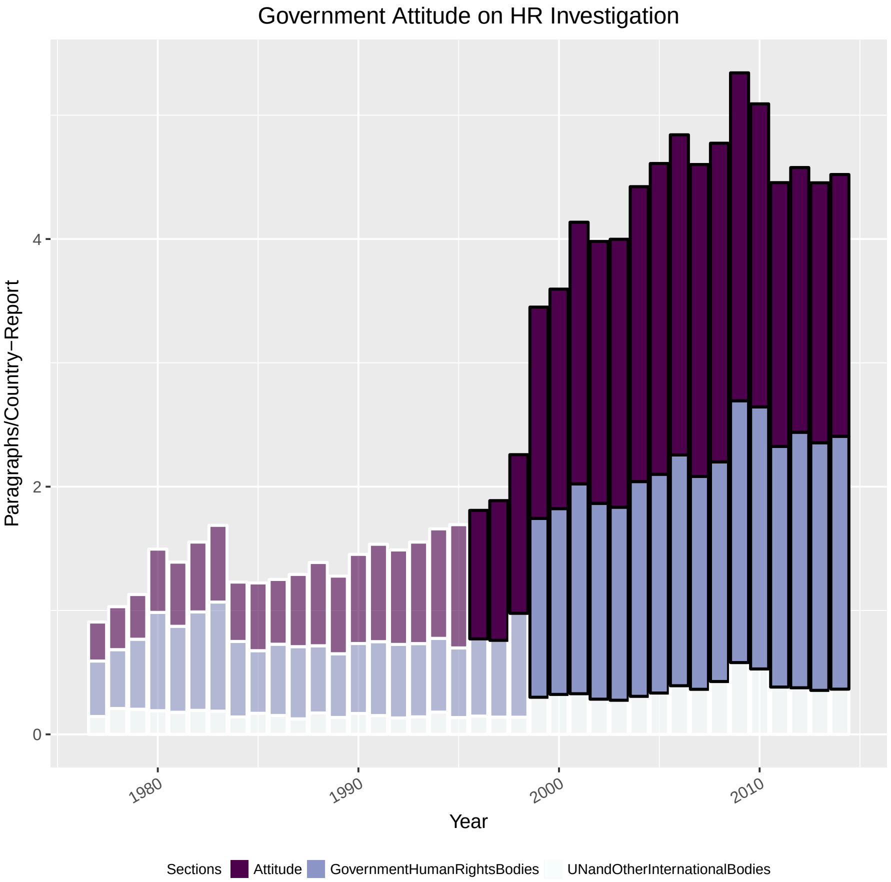
Second, the number of explicit labels for human rights violations has increased considerably over time. As discussed in the main paper, there were only about 11 categories in 1977, but there are 112 categories in 2015/2016. Thus, many of the explicit labels from 2015/2016 do not exist in those early years. We try to do range-approximation labeling based on the contents between the early years and the later years. For example, in 1977, there's is a very general category of "*Arbitrary Arrest or Imprisonment*", but in 2015/2016, there are 7 subcategories under "*Arbitrary Arrest or Detention*": "*Role of the Police and Security Apparatus*", "*Arrest Procedures and Treatment of Detainees*", "*Arbitrary Arrest*","*Pretrial Detention*","*Amnesty*","*Detainee's Ability to Challenge Lawfulness of Detention before a Court*","*Protracted Detention of Rejected Asylum Seekers or Stateless Persons*." Thus, when we match 2015/2016 labels to 1977 label, these 7 labels belonging to the *Arbitrary Arrest or Imprisonment*” section in 1977.[12]

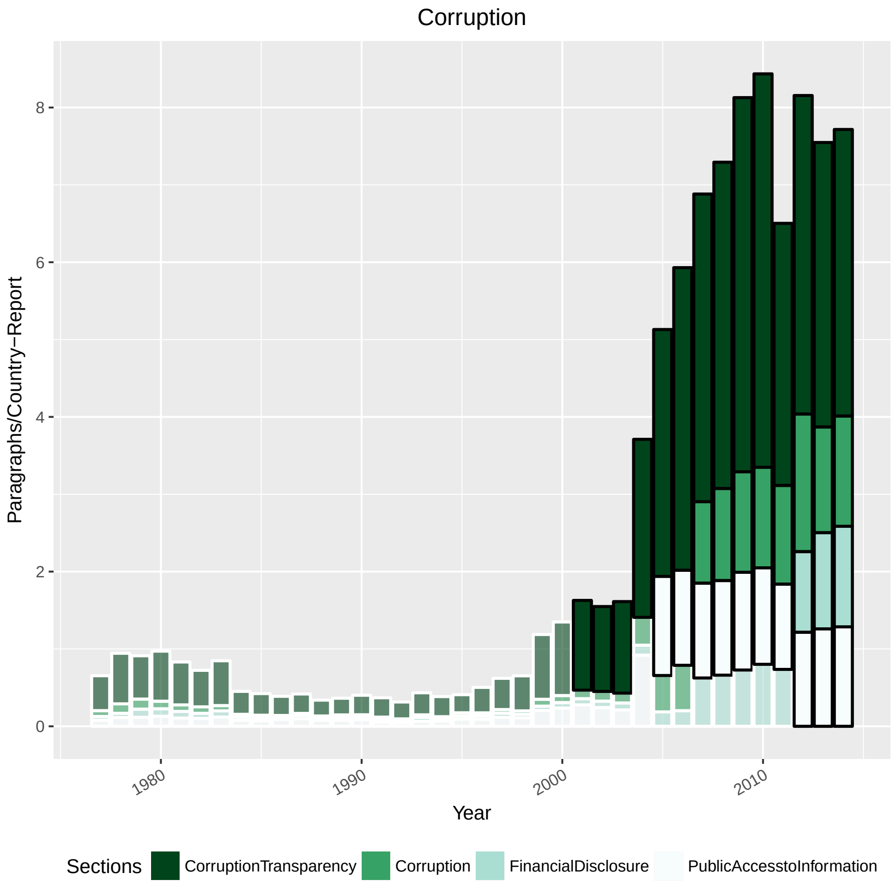
[12] This range-approximation labeling is particularly useful for *PRE* calculations.

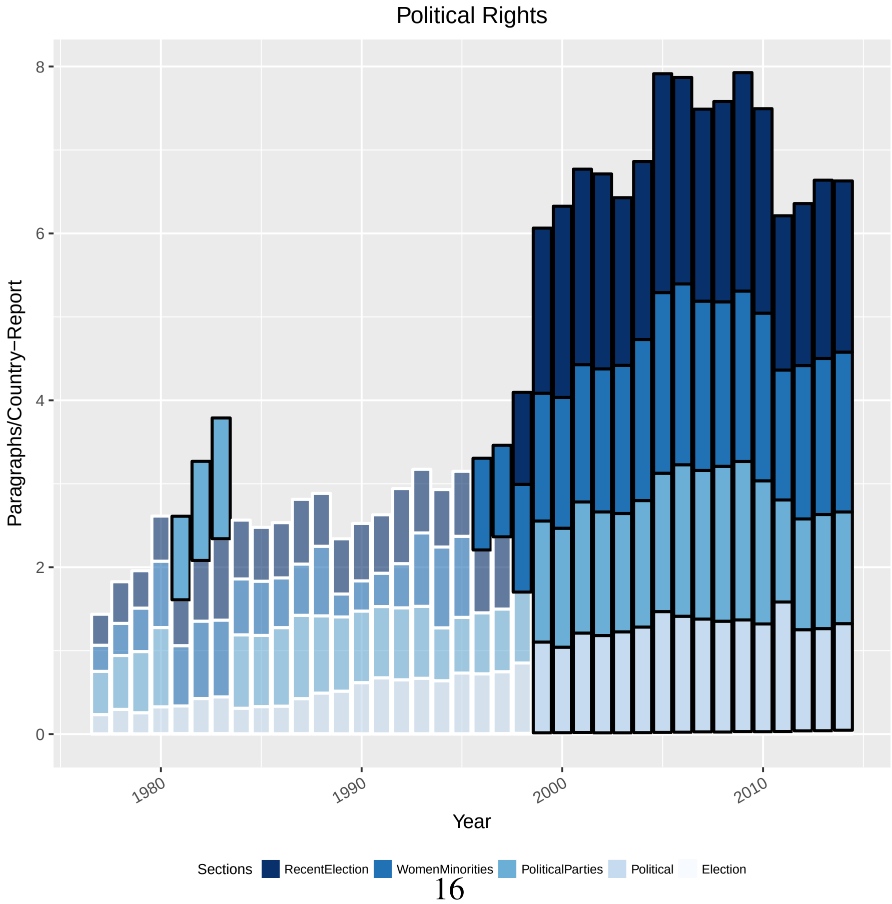
Table D.1: Key to 2015/2016 Aspects, Sections and Labels (112)

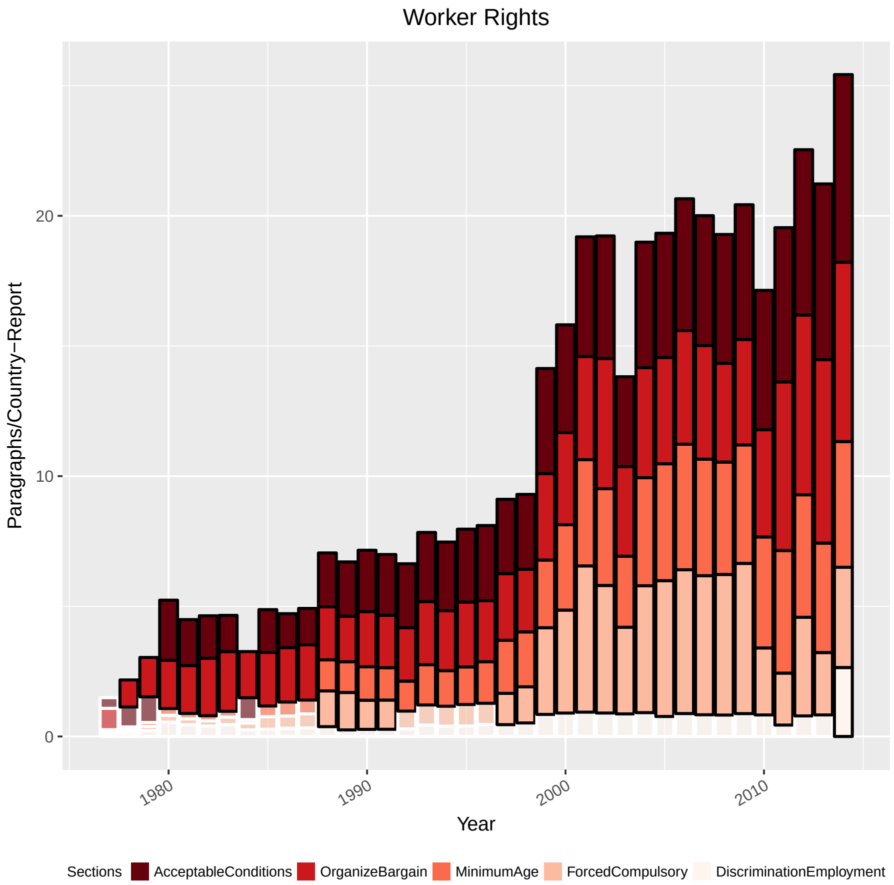
Label	**Section (Specific Aspect)**	**Label**	**Section (Specific Aspect)**
sec 000	Section Attitude	sec 056	Section Discrimination Children MedicalCare
sec 001	Section Attitude GovernmentHumanRightsBodies	sec 057	Section Discrimination Children SexualExploitationofChildren
sec 002	Section Attitude UNandOtherInternationalBodies	sec 058	Section Discrimination HIVandAIDSSocialStigma
sec 003	Section Civil Assembly	sec 059	Section Discrimination Incitement
sec 004	Section Civil Assembly FreedomofAssembly	sec 060	Section Discrimination IndigenousPeople
sec 005	Section Civil Assembly FreedomofAssociation	sec 061	Section Discrimination NationalRacialEthnicMinorities
sec 006	Section Civil Movement	sec 062	Section Discrimination OtherSocietalAbusesandDiscrimination
sec 007	Section Civil Movement AbuseofMigrantsRefugees	sec 063	Section Discrimination PeoplewithDisabilities
sec 008	Section Civil Movement Citizenships	sec 064	Section Discrimination SocietalDiscriminationSexualOrientation
sec 009	Section Civil Movement EmigrationandRepatriation	sec 065	Section Discrimination TraffickinginPersons
sec 010	Section Civil Movement Exile	sec 066	Section Discrimination Women
sec 011	Section Civil Movement ForeignTravel	sec 067	Section Discrimination Women Discrimination
sec 012	Section Civil Movement IncountryMovement	sec 068	Section Discrimination Women FGM
sec 013	Section Civil Movement InternallyDisplacedPersons	sec 069	Section Discrimination Women GenderbiasedSexSelection
sec 014	Section Civil Movement ProtectionofRefugees	sec 070	Section Discrimination Women HarmfulTraditionalPractices
sec 015	Section Civil Movement ProtectionofRefugees AccesstoAsylum	sec 071	Section Discrimination Women RapeandDomesticViolence
sec 016	Section Civil Movement ProtectionofRefugees AccesstoBasicServices	sec 072	Section Discrimination Women ReproductiveRights
sec 017	Section Civil Movement ProtectionofRefugees DurableSolutions	sec 073	Section Discrimination Women SexualHarassment
sec 018	Section Civil Movement ProtectionofRefugees Employment	sec 074	Section Integrity ArrestDetention
sec 019	Section Civil Movement ProtectionofRefugees FreedomofMovement	sec 075	Section Integrity ArrestDetention ArrestDetain
sec 020	Section Civil Movement ProtectionofRefugees Nonrefoulement	sec 076	Section Integrity ArrestDetention ArrestDetain AbilitytoChallenge
sec 021	Section Civil Movement ProtectionofRefugees RefugeeAbuse	sec 077	Section Integrity ArrestDetention ArrestDetain Amnesty
sec 022	Section Civil Movement ProtectionofRefugees SafeCountryofOriginTransit	sec 078	Section Integrity ArrestDetention ArrestDetain ArbitraryArrest
sec 023	Section Civil Movement ProtectionofRefugees TemporaryProtection	sec 079	Section Integrity ArrestDetention ArrestDetain DetentionofRejectedAsylumStatelessPersons
sec 024	Section Civil Movement StatelessPersons	sec 080	Section Integrity ArrestDetention ArrestDetain PretrialDetention
sec 025	Section Civil Religion	sec 081	Section Integrity ArrestDetention PoliceSecurity
sec 026	Section Civil SpeechPress AcademicCultural	sec 082	Section Integrity Denial
sec 027	Section Civil SpeechPress ActionstoExpandPressFreedom	sec 083	Section Integrity Denial CivilJudicialProceduresandRemedies
sec 028	Section Civil SpeechPress InternetFreedom	sec 084	Section Integrity Denial PoliticalPrisoners
sec 029	Section Civil SpeechPress Status	sec 085	Section Integrity Denial PropertyRestitution
sec 030	Section Civil SpeechPress Status ActionstoExpandPressFreedom	sec 086	Section Integrity Denial TrialProcedures
sec 031	Section Civil SpeechPress Status CensorshiporContenRestrictions	sec 087	Section Integrity Disappearance
sec 032	Section Civil SpeechPress Status FreedomofPress	sec 088	Section Integrity Extrajudicial
sec 033	Section Civil SpeechPress Status FreedomofSpeech	sec 089	Section Integrity Force
sec 034	Section Civil SpeechPress Status LibelSlanderLaws	sec 090	Section Integrity Force Abductions
sec 035	Section Civil SpeechPress Status NationalSecurity	sec 091	Section Integrity Force ChildSoldiers
sec 036	Section Civil SpeechPress Status NongovernmentalImpact	sec 092	Section Integrity Force Killings
sec 037	Section Civil SpeechPress Status ViolenceandHarassment	sec 093	Section Integrity Force OtherConflictRelatedAbuses
sec 038	Section Corruption	sec 094	Section Integrity Force PhysicalAbuse
sec 039	Section Corruption CorruptionTransparency	sec 095	Section Integrity Privacy
sec 040	Section Corruption FinancialDisclosure	sec 096	Section Integrity Torture
sec 041	Section Corruption PublicAccesstoInformation	sec 097	Section Integrity Torture PrisonDetentionCenterConditions
sec 042	Section Discrimination	sec 098	Section Integrity Torture PrisonDetentionCenterConditions Administration
sec 043	Section Discrimination AntiSemitism	sec 099	Section Integrity Torture PrisonDetentionCenterConditions Improvements
sec 044	Section Discrimination Children	sec 100	Section Integrity Torture PrisonDetentionCenterConditions Monitoring
sec 045	Section Discrimination Children BirthRegistration	sec 101	Section Integrity Torture PrisonDetentionCenterConditions PhysicalConditions
sec 046	Section Discrimination Children ChildAbuse	sec 102	Section Political
sec 047	Section Discrimination Children ChildFGM	sec 103	Section Political Election
sec 048	Section Discrimination Children ChildMarriage	sec 104	Section Political Election ParticipationofWomenandMinorities
sec 049	Section Discrimination Children ChildSoldiers	sec 105	Section Political Election PoliticalParties
sec 050	Section Discrimination Children DisplacedChildren	sec 106	Section Political Election RecentElection
sec 051	Section Discrimination Children Education	sec 107	Section Worker AcceptableConditions
sec 052	Section Discrimination Children HarmfulTraditionalPractices	sec 108	Section Worker DiscriminationEmployment
sec 053	Section Discrimination Children Infanticide	sec 109	Section Worker ForcedCompulsory
sec 054	Section Discrimination Children InstitutionalizedChildren	sec 110	Section Worker MinimumAge
sec 055	Section Discrimination Children InternationalChildAbductions	sec 111	Section Worker OrganizeBargain

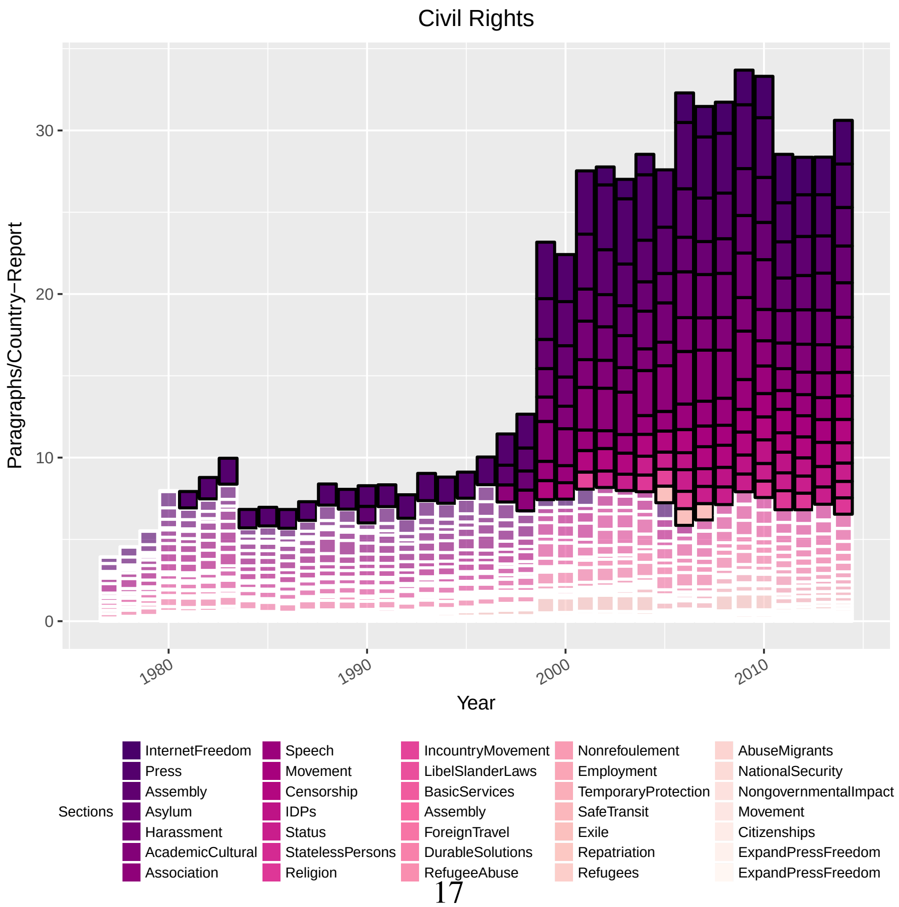
E Attention Across the Implicit Taxonomies of State Department Reports for the Additional Sections




Corruption
Paragraphs/Country−Report
0
2
4
6
8
1980
1990
2000
2010
Year
Sections
CorruptionTransparency
Corruption
FinancialDisclosure
PublicAccesstoInformation


Political Rights
Paragraphs/Country−Report
0
2
4
6
8
1980
1990
2000
2010
Year
Sections
RecentElection
WomenMinorities
PoliticalParties
Political
Election


Worker Rights
Paragraphs/Country−Report
0
10
20
1980
1990
2000
2010
Year
Sections
AcceptableConditions
OrganizeBargain
MinimumAge
ForcedCompulsory
DiscriminationEmployment


Civil Rights
Paragraphs/Country−Report
0
10
20
30
1980
1990
2000
2010
Year
Sections
InternetFreedom
Press
Assembly
Asylum
Harassment
AcademicCultural
Association
Speech
Movement
Censorship
IDPs
Status
StatelessPersons
Religion
IncountryMovement
LibelSlanderLaws
BasicServices
Assembly
ForeignTravel
DurableSolutions
RefugeeAbuse
Nonrefoulement
Employment
TemporaryProtection
SafeTransit
Exile
Repatriation
Refugees
AbuseMigrants
NationalSecurity
NongovernmentalImpact
Movement
Citizenships
ExpandPressFreedom
ExpandPressFreedom

F Depth of Coverage Over Time in State Department Reports

Relatedly, we are interested in whether more attention is being placed on more specific and complex distinctions in later reports, as compared to earlier reports. With our model, we can analyze whether the rights/aspects being judged are, on average, farther down the taxonomy represented by G_{2015_2016}, as compared to earlier years. This can provide more direct evidence that increasingly specific information from HROs, satellites and camera phones produce more fine-grained distinctions in the text of the reports themselves.

Figure F.1 shows not only an increase in total coverage in the reports, but also that there is more coverage of fine-grained rights in later years, compared to earlier years. In 1977 the bulk of the text is found in lower level sections (level 1) rather than deeper subsections (levels 3 and 4). By 2014 however, the proportion of text found in these deeper sections increases dramatically. For the Discrimination sections, roughly half of the text is found in the deepest nested subsection. There is a similar, though, less dramatic trend for the coverage of Physical Integrity rights. Here there is very little content dealing with more specific, deeply nested rights in early years, but considerably more in later years.

These large increases in content for more specific rights demonstrates not only that the reports change over time, but also that they cover more fine-grained and specific protections and violations in later years. This is particularly telling evidence because, while there could have been additional attention to more general rights without concomitant increases in specific human rights, adding specific rights also often adds discussion of the general human rights they are nested within. Thus, the fact that the deeper sections grow along with the increases in the reports is important.

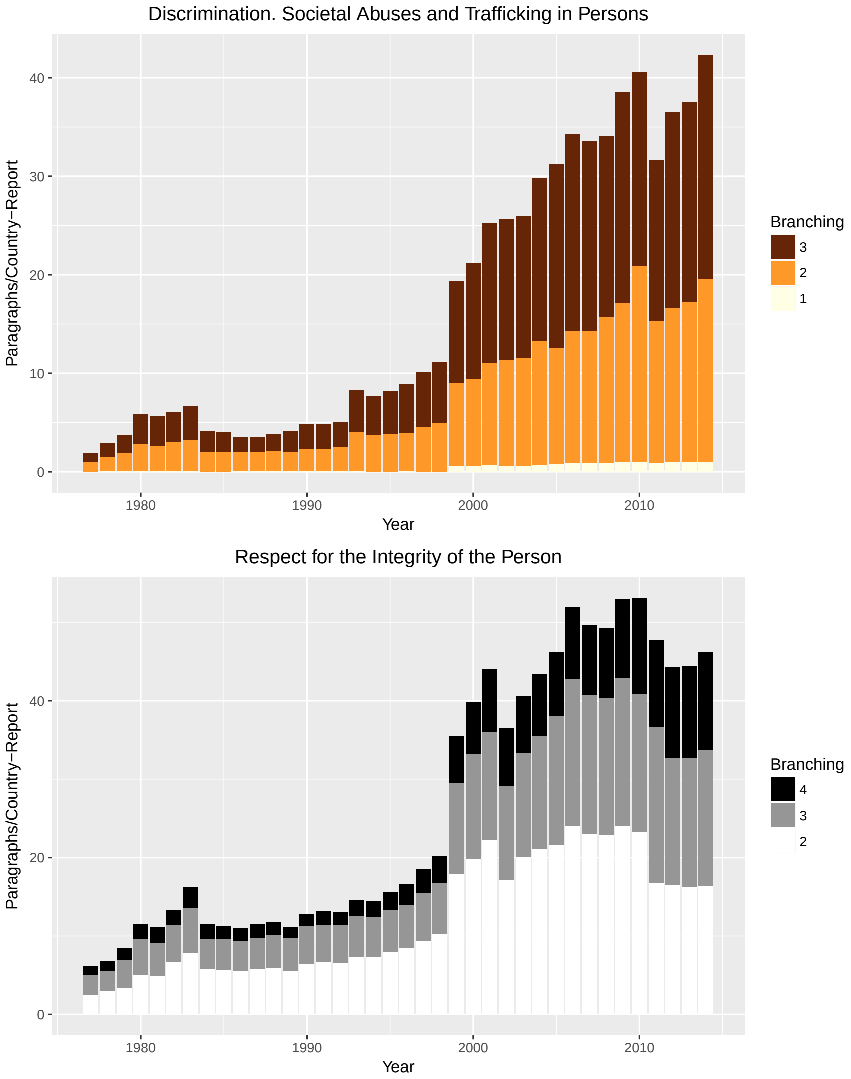


Figure F.1: Count of the depth where paragraphs fall in the State Department report over time. 1 signifies paragraphs falling in the main section (most general) while 4 signifies paragraphs falling in a sub-sub-sub-section (most specific).

G The Relationship Between ABSA and Coding the Taxonomy of Human Rights Being Judged

One increasingly common set of natural language processing tasks is aspect-based sentiment analysis (ABSA) (Liu, 2012; Pang and Lee, 2008; Pontiki et al., 2014). Conventional sentiment analysis

Slot	Description	Product Example	Human Rights Report
who	Who/what is the source of opinion?	Roger Ebert	State Department
feels how	What is direction of the judgment?	negative	frequent violations
about what part	What abstract aspect is being judged?	acting	protection from torture
of an entity or object	Who/what is the target of the judgment?	Ishtar (1987)	North Korea (2014)

Table G.1: The sub-tasks of aspect-based sentiment analysis and examples from a movie review and a human rights report on a particular country, in a given year.

attempts to identify the overall aggregate negative or positive sentiments, judgments or opinions[13] in a given text, such as a movie or product review.

ABSA aims to identify what abstract aspects are being judged positively or negatively on a particular instance of an entity or object. Thus ABSA attempts to answer the question, "who feels how about what part of an entity or object?" using natural language. Answering this question can be broken down into several sub-tasks. Table G.1 provides a summary of each slot in the question, along with an explanation of the sub-task. The canonical examples for ABSA are from movie and product reviews Liu (2012); Pontiki et al. (2014). For example, a reviewer might write something negative about the acting in a particular movie, such as Ishtar, but have enjoyed the script.

There are two crucial conceptual connections between ABSA and our exploration of the evolving taxonomy of aspects that are judged in human rights reports over time. First, before or alongside identification of the valence of an opinion, be it positive, negative or neutral, the aspect that is being judged needs to be extracted from the text of a movie review. We aim to mine the aspects of human rights that are being judged in human rights country-reports. Second, ABSA groups specific aspects that are semantically related into more general categories, forming a hierarchy. Acting can be separated into lead and supporting roles, and each of those types of roles can be split further into more specific male or female parts (Liu, 2012).[14] Figure G.1 includes a hierarchy of a

[13] The literature has used the word sentiment to refer to opinions, judgments and emotions (Liu, 2012). In our set of tasks here, we are interested in judgments of human rights protections and violations but we will continue to use the term sentiment to connect our approach to the larger literature on aspect-based sentiment analysis.

[14] In ABSA, because judgments are on aspects, the sub-task of identifying what aspects are being discussed, and how they nest within categories is of central importance. Aspects must be abstract in the sense that they can apply to more than one target entity. So the concept of a plot is an aspect, but the movie Citizen Kane (1941) is a particular instance/entity that can have a plot. Conventional sentiment analysis can be seen as a special case of ABSA, whereby all the texts being scored in a corpora are assumed to only refer to one general aspect.

subset of aspects that might be used to describe a movie.

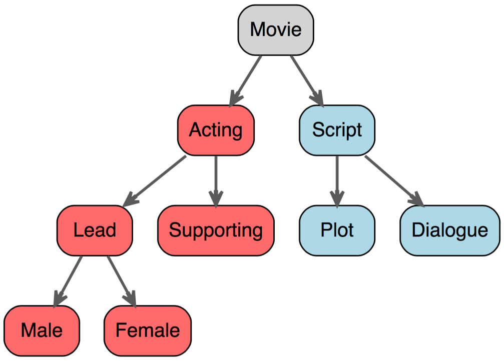


Figure G.1: An example of specific aspects nested within more general categories forming a conceptual hierarchy.

There are more general aspect-categories at the top of the tree, and more specific fine-grained aspects at the bottom. Each node or leaf in the tree, a node with no children, could be expressed in the text with different words. There can also be leaves that refer to general concepts. For example a statement about a movie overall, “It was terrible”, would apply to the root of the tree Liu (2012).

H PULSAR

PULSAR (Parsing Unstructured Language into Sentiment-on-Aspect Representations) is a tool for processing large scale unstructured texts into structured aspect-based sentiment expressions. It is based on a series of grammatical and syntactic rules to identify and extract aspects as noun-phrases and judgments as verb-phrases. In particular, PULSAR generates multi-word expressions (MWE) based outputs, instead of simple Bag-of-Words tokens. Thus, it allows us to identify phrases and words that are semantically more meaningful and more interpretable. Table H.1 illustrates some of the example sentences and the parsed outputs from PULSAR. Unlike the Bag-of-Words (BOW) approach, we can identify the aspect that a given sentiment is attached to. For example, “*citizens privacy rights*” were “*infringed on*”, and there “*NEG were*” (were no) “*politically motivated disappearances*.” See Park, Colaresi and Greene (2018) for further details.

Sentence	Output
The Taliban arbitrarily arrested and detained persons and infringed on citizens privacy rights.	(citizens_privacy_rights, infringed_on), (persons, arbitrarily_arrested)
There were no reports of politically motivated disappearances.	(politically_motivated_disappearances, NEG_were)
There were credible reports that the government or its agents committed arbitrary or unlawful killings.	(unlawful_killings, POS_committed)
Taliban forces were responsible for indiscriminate bombardment of civilian areas.	(indiscriminate_bombardment, were_responsible_for)

Table H.1: Table of outputs from PULSAR for four example sentences from the State Department Human Rights Reports.

I An Example of Taxonomies Growing With Technology: Biology

First of all, we believe that the change in the underlying taxonomy of human rights that we are proposing is similar to the creation and growth of biological taxonomies as new technologies allowed for researchers to perceive increasingly specific and detailed differences and similarities in plants, animals and eventually cellular-level instances over time (Small, 1989). In a sparse information environment, differences between animals or plants would be categorized morphologically leading to general, but coarse, taxonomies. However, more complete information that can be collected across animal specimens allows for those general categories to be increasingly differentiated into more specific concepts. Stace (1991, 3, 43, 87) explicitly notes that the discovery and availability of new technological tools such as the electron microscope and spectrometer led to the discovery of new classes of living beings in biology. The same text also notes that these changes in technology have measurement consequences, reminding us that novel plants and animals that went extinct before the availability of the tools to identify their difference from other morphologically similar species are missing from the current biological record (Stace, 1991, 43).[15]

Biology is not unique in providing an example of groups of individual evolving taxonomies over time in reaction to new and more detailed information. Barner and Baron (2016) provide

[15] This is directly analogous to Fariss (2014)'s argument that past violations may have been missed in previous reports and codings, but these distinct behaviors could have been caught and codified later.

a number of examples of increasing the specificity of concepts as unexpected evidence becomes available (19-21). They further tie this conceptual evolution to the human use of Bayesian inference where prior concept hierarchies limit what is perceived, but new information updates those conceptual taxonomies over time (21). In this way, categorizing human rights that can be written about systematically across countries is no different than other human endeavors to incrementally learn about the world around them as new types of signals are perceivable.[16]

J Supervised Learning Approach to Computing the Implicit Taxonomy of Rights

On the basis of the features from PULSAR, we build a system that can accurately predict the human right concepts being judged within unlabeled text. To formalize how we use our input features to accomplish this task, let each paragraph be indexed by $d \in (1, \ldots, D)$ in a D-length set of all country reports. These are split into orthogonal sets, where D_{train} is the total number of training paragraphs, where specific target labels are available. D_{test} is the number of test paragraphs, beginning at $D_{train} + 1_{th}$ paragraph, where labels are unavailable. These sets can be further partitioned into specific annual reports of length D_t. Denote the string of raw text for paragraph d as m_d. The raw text of each m_d paragraph is encoded into a vector of counts or tf-idf values by PULSAR, such that $PULSAR(m_d) \rightarrow x_{dt}$, where x_{dt} is a V vocabulary-length vector of counts or reals.[17] For convenience, the country of the report where paragraph d appeared is indexed by $n \in (1, \ldots, N)$ and $t \in (1, \ldots, T)$ references the time period of the report where d was found.[18] Denote X_{train}

[16] The farther down a hierarchy, the more specific the defined distinctions. While a crocodile and alligator share many similarities including the same kingdom, phylum, class and order, they have different family and subfamily designations. Further, it takes more expertise and information to process these differences. When applied to rights, we can think about the availability of video evidence of children in dangerous working conditions along with HROs checking up on leads around the globe, leading to new comparisons across countries in how they protect children from these conditions. Precise definitions and concepts, at lower levels of the conceptual hierarchy, facilitate accurate processing of information and reporting.

[17] One can think of the PULSAR output as semi-automatically generated multi-word expressions and tokens that are then counted, and then vectorized.

[18] We suppress the complication that different numbers of countries are included across different years in this notation. A bag of words representations of the paragraphs over-fit the categories, because without filtering proper nouns

as the $D_{train} \times V$ matrix of feature counts/tf-idf values that PULSAR produced where location labels for a given future (or past) taxonomy are available and X_{test} as the $D_{test} \times V$ matrix of feature counts/tf-idf values, again produced by PULSAR, but where location labels are unavailable[19], matching the pre-processed partition of the D paragraphs into D_{train} and D_{test}-length sets.

The labels we are predicting in the training set are the locations in a specific, target human rights taxonomy that was explicitly defined, $G_{\dot{t}}$, for paragraphs in the training set, $m_d, d \in D_{train}$. Let $\dot{t}$ index the year of the taxonomy of interest[20] and let $y_{dt}^{(\dot{t})}$ as $L_i\times$ 1, represent the one-hot encoded $L_{\dot{t}}$-length vector of these labels for a paragraph in the training set.[21] Different explicit taxonomies have different numbers of nodes, as we saw above. Also, denote $Y_{train}^{(\dot{t})}$ as the $D_{train} \times L_{\dot{t}}$ matrix of these vectors for the whole training set.

To be clear, we are using the later, detailed explicit taxonomies, as they provide labels for paragraphs, to learn the implicit taxonomies for earlier years, when those detailed labels were not in use. Thus, the goal of our training is learning a mapping, $\hat{f}_{X,\dot{t}} : X_{train} \rightarrow Y_{train}^{(\dot{t})}$, from the V-length features extracted with PUSLAR from a paragraph to the $L_{\dot{t}}$-length taxonomic labels from $G_{\dot{t}}$, such that $\hat{Y}_{dt}^{(\dot{t})} = \hat{f}_{X,\dot{t}}(x_{dt})$. In the training set, we use the observed features and labels to compute a useful and high performing $\hat{f}_{X,\dot{t}}$. With $\hat{f}_{X,\dot{t}}$, we can now map X_{test} to a distribution across locations in $G_{\dot{t}}$, even when they were not provided, $\hat{Y}_{test}^{(\dot{t})}$.

We use a range of different model representations to compute potential functions for the $\hat{f}_{X,\dot{t}}$. The class of algorithms included logistic regression, naive Bayes, SVM and random forests. We compared models based on cross-validated accuracy in the training set to judge their relative performance. The baseline random probability for predicting one of 112 labels correctly is 0.008. We use $\dot{t} = 2015/2016$ as these years use the most detailed explicit taxonomy we have from the State

they learn that "North Korea" is a term that tells you we are talking about physical integrity abuses. This confuses the target of the judgment with the aspects being discussed. Our parser identifies known named entities under discussion and references to them. More generally, we sweep out proper nouns as they are not abstract aspects by definition.

[19] To be clear, they are unavailable because there is no explicit lables from the training data for those early years that we are trying to predict.

[20] Taxonomies are defined by their structure G, and thus can be the same or different across years.

[21] Thus, we have a binary vector where only one value is turned on for a given paragraph. If $\dot{t} = 2015/2016$, as below, then we have $L_{2015/2016} = 112$. For a paragraph with location 4, the fourth value of $y_{dt}^{(2015/2016)}$ would be 1 and all the other 111 values in the vector are zero.

Department. For comparison, we also utilized $\dot{t} = 1977/1978$, which was the coarsest taxonomy as the training set. These results help us demonstrate that a supervised learning approach is useful to identifying the human rights being judged, relative to a given concept taxonomy. Below we focus on the output from $\hat{f}_{X,2015/2016}$, the models trained on $G_{2015/2016}$ and the 2015 and 2016 annual reports. The test set, where exactly matching explicit labels are missing are then $1977 - 2014$.

Table J.1 presents the accuracy of all of these models, trained on tf-idf features constructed from the PULSAR features.[22]

	Accuracy (CV)	
Model	**1977/1978 Sections**	**2015/2016 Leaves**
LR	0.946	0.858
SVM	0.945	0.857
NB	0.943	0.817
RF	0.932	0.834
Baseline	$0.\overline{250}$	$0.\overline{010}$

Table J.1: Cross-validated accuracy for labels defined first by $G_{1977/1978}$ with three section nodes and then the more complex $G_{2015/2016}$ with 112 nodes, used in 2015/2016. The training windows were 1977/1978 and 2015/2016 respectively. All results are for tf-idf of multiword expressions and words as described in the text. The baseline is the probability of accurately predicting a class for each task by random guessing. Despite having 112 nodes to predict for each paragraph, we reach over 85 percent accuracy for our best performing models, learning $\hat{f}_{X,2015/2016}$.

Our approach yields over 94 percent accuracy on the easy, 1977/1978, single level taxonomy task, but reassuringly, is able to maintain its high performance on the more difficult, multi-level 2016 taxonomy, with 112 concepts to identify. Note the baseline accuracies at the bottom of the table describe the overall difficulties of these tasks. By random chance, one would only classify 1 in 100 paragraphs accurately for $G_{2015/2016}$. Thus, our model produces an 85 times increase in accuracy. The analogous baseline accuracy for our $G_{1977/1978}$ is 1 in 4, with the model producing a bit less than a 4 times increase.

[22] Raw term frequency yielded lower accuracies across model specifications.

J.1 Measuring Changes in the Implicit Taxonomy of Human Rights: Structure, Attention and Sharpness

Our highest performing model supplies predictions of what concept, within the 2015/2016 taxonomy, was being discussed for each paragraph in the corpora in all other years. This allows us to see when and how locations in the most detailed taxonomies *were* and *were not* judged in past reports. In the main paper we explore 3 ideas. First, we track the structure of the taxonomies. What rights are being discussed in at least one paragraph per country in a given annual report. This threshold defines when a taxon exists in the implicit taxonomy estimated for a given year.[23] This definition allows us to cleanly visualize the (in)consistency of the taxonomic structure as available information density has increased.

Second, we measure the amount of attention a label receives. We measure this with the number of paragraphs that our model estimates were discussed in a given annual report. Attention provides us with a way of tracking the consistency in the amount of attention per concept across each annual report as information density has increased over time.[24]

Third, we are interested in whether the density of information, leads to sharper distinctions between concepts. A taxonomy is defined by the distinctions between concepts as one descends the hierarchy. Denser information over time should allow not only more and more distinction, but the distinctions should grow clearer. This would be visible in the sharpness of our models predictions over time. Using information theory as a foundation, as we discuss further below, we are also able to directly measure the average sharpness in our predictions, relative to a baseline, across time. Dense information will be detectable in sharp predictions of the presence of specific concepts. If the text is informative, then the features from the text that uniquely signal a label will show up in past reports. Our model will input that relevant information and produce higher predictions for the relevant concept and lower predictions for others (Colaresi and Mahmood, Forthcoming), even if

[23] Below in our analysis of Human Rights Watch press releases, which are not annual report, we set the threshold to be 5 press releases in a given year. This allows for one node to be included in the first year in our data for that specific corpora.

[24] We use the sum of the probabilities within labels but across paragraphs in a given year to avoid the bias highlighted in King and Lowe (2003).

that concept was not explicitly labeled in that year's taxonomy. Similarly, and perhaps most importantly, if older texts provide less dense relevant information on the conceptual distinctions between rights across the taxonomy, the textual features that uniquely signify a concept in the recent right's taxonomy would then absent, as in a pixelated image. In that case our model will be unable to differentiate among rights, and will produce flatter predictions. Our available information density theory suggests that there should be less sharp predictions in the past, when less information was available, and sharper distinction more recently, when additional information was available. We can directly measure the informative-ness detected by our model using a version of average sharpness in each year that we define below.

Our model provides a signal $P(y_{dj} = 1|X_d)$, the probability that paragraph d from the test set is judging right j; which we will truncate to p_{dj} for brevity. We have predictions for all the D_{test} paragraphs in the test set and for all the 112 rights in taxonomy $G_{2015/2016}$. The Shannon entropy or surprisal is a measure of the absence of information in a given message and is denoted as $-log_2(p_{dj})$. The absence of information is conceptualized as how surprised you would be if you found out $y_{dj} = 1$. If you were surprised, then you did not have this information already. If, prior to a message about an event, you were more and more certain that $p_{dj} \to 1$, a smaller and smaller amount of information would be gained from the message that $y_{dj} = 1$. These would be maximally sharp predictions, represented by spikes in one or a few labels, and very low values for others. Formally,

$$\lim_{x \to 1} -log_2(x) = 0$$

Shannon also defined the expectation over the surprisal across all possible messages for all possible events. In our case this is,

$$H(p_d) = \sum_{j \in G_{2015/2016}} p_{dj}(-log_2 p_{dj})$$

This is simply the average entropy for a paragraph d across all the $L_{2015/2016}$ labels. We can think of average entropy as measuring the sharpness versus the flatness of our predictions for a paragraph across the labels. Flatter vectors of predictions, where we approach all p_{dj} being equal, and thus

$p_{dj} = \frac{1}{L_{2015/2016}}$, approach the maximum possible entropy, and thus supply less information on their own. Sharper predictions, where our model favors one category over others, have lower entropy and thus convey more information.

We are interested in measuring the average sharpness[25] in a paragraph for each annual report. We expect sharp predictions in recent years, as the text should supply information on the distinctions that our model encodes. However, going back in time, if information on specific and complex distinction was not available, then our model cannot sharply predict locations. To facilitate comparisons across different potential taxonomies $G_{\dot{t}}$, with distinct $L_{\dot{t}}$, we utilize a rescaled version of average sharpness, here denoted as $\bar{S}(p|q|)$,

$$\bar{S}(p||q) = log_2 L_{\dot{t}} - \frac{1}{D_t} \sum_{d=1}^{D_t} H(p_d)$$

, in a given year, which is equal to the KL-divergence of our predictions from maximum entropy predictions, $q_j = \frac{1}{L_{\dot{t}}}$, $\forall\ j \in\ G_{\dot{t}}$. See below for the proof of this equality.

J.2 Proof that Average Sharpness is the Expected KL Divergence from the Maximum Entropy Distribution over Labels

Here we provide a short proof that average sharpness ($\bar{S}$) is equal to the Expected KL-divergence of p, where p is $D_t \times L_{\dot{t}}$ matrix of predictions for all documents in year with p_d representing row d, from q, where q_d is a $L_{\dot{t}}$ row-vector with each element being $\frac{1}{L_{\dot{t}}}$ and q matrix of stacked vectors

[25] We use the expectation across paragraphs in a year so that the length of the reports do not build in a bias to find more information in later reports.

for the set. Then,

$$\begin{aligned}
E(KL(p||q)) &= E\left[\sum_{j\in G_i} p_{dj}\ log_2\left(\frac{p_{dj}}{q_{dj}}\right)\right] \\
&= E\left[\sum_{j\in G_i} p_{dj}\ log_2\left(\frac{p_{dj}}{\frac{1}{L_i}}\right)\right] \\
&= E\left[log_2 L_i - \sum_{j\in G_i} p_{dj}\ log_2 p_{dj}\right] \\
&= E\left[log_2 L_i - H(p_d)\right] \\
&= \frac{\sum_{d=1}^{D_t} log_2 L_i - H(p_d)}{D_t} \\
&= log_2 L_i - \frac{1}{D_t} H(p_d) = \bar{S}(p||q) \\
& q.e.d.
\end{aligned}$$

J.3 Implicit vs. Implicit Taxonomies

In the main paper we also use our model to detect an alternative form of change in the structure of the country reports, where the drafters of the reports might label the same lexical features as different categories across years or where they might spread content across the explicit taxonomy. For example, even if the same aspects were being discussed across years, those aspects might be labeled as being in a different part of the hierarchy at different time points. Intimidation of the opposition might switch from physical integrity rights to political rights. We can check for this form of inconsistency by measuring the accuracy of our predicted label (taken from another year) for that paragraph relative to the actual location of the paragraph in that year's set of report. While the labels will not match one-to-one. The parent section or sub-sections headings should. In addition, our model has the possibility of detecting the judgment of rights outside the explicit taxonomy for a particular paragraph. It may be the case that information related to worker's rights is included in the sibling section on civil rights. Our model could thus serve as a guide to help human coders more efficiently read human rights reports by finding rights that are being judged in the text, but would be missed if only explicit labels guided the coding.

K Results for the the Low Resolution Taxonomy, $G_{1977/1978}$ for the State Department

We also explore the consistency of the lowest resolution, oldest taxonomy, and how that projected forward in time. This supplies a placebo test for our sharpness analysis. It should not be possible to detect future distinctions in concepts with the previous, older and coarser taxonomy. At the lowest resolution, the section-level of the 1977/1978 documents, there are 4 sections (See Table K.1). We therefore train a classifier on the binary indicator $y_{din,1977/1978} \in (0, 1)$, where $i \in (1, \dots, 4)$ indexes the possible sections within which paragraph d written about country n in 1977/1978 could be located. We use logistic regression, naive Bayes, SVM and random forests to explore the mapping $x_{dn,1977/1978} \rightarrow y_{din,1977/1978}, \; \forall \, i$.

The computed models supply us with $\hat{p}_{1977/1978}(y_{din,1977/1978}|x_{dnt})$, estimates of what 1977/1978 aspect categories are being referenced given the input features from a given paragraph, written for a specific country-year report. In other words, we can then use the feature vectors for paragraphs across all other years, with the trained model, to predict what aspect-category a paragraph would have been labeled as, had it been written in 1977/1978. Since we also have the actual labeled aspect-category for each paragraph, we can assess how accurately our model trained on these low resolution features in 1977/1978 classifies paragraphs in later years. We can also track the amount of information our model supplies in each successive year by computing the average Shannon entropy of the predictions across the labels in each year. If our model is completely uncertain about where to place paragraphs, the probabilities will be flat and there will be high uncertainty. On the other hand, if the model is nearly certain of a category, then the entropy will be minimized.

Figure K.1 presents the number of expected paragraphs for each low resolution aspect-category as a stacked bar chart over time. We see that there are observations in each section moving for-

Label	Section
S_1	Section Integrity
S_2	Section Civil
S_5	Section Attitude
S_7	Section Worker

Table K.1: Key to 1977/1978 Aspects, Sections and Labels (4)

ward in time. We also see that the amount of attention has changed in several of the sections. The section “Respect for the Integrity of the Person” (gray), also known as physical integrity rights and the section “Respect for Civil Liberties (pink)” have grown dramatically. The amount of predicted attention to “Worker’s Rights” (red) and “Governmental Attitude Regarding International and Nongovernmental Investigation of Alleged Violations of Human Rights” (gray-blue) have grown a little. Each of the general topics continues to be discussed, at least somewhere in the documents across each year. We detect increases in attention over time, as expected.

We also calculate the average sharpness across the years as discussed above. The higher the average sharpness of the model predictions in year, the less information our model contains, on average, about the labels in $G_{1977/1978}$. Tracking the average sharpness of our model trained to identify low resolution taxonomies allows us to measure whether a simple taxonomy has remained consistent moving forward in time.

The plot of average sharpness for low taxonomy over time is presented in Figure K.2. Since there are four possible classes, the maximum entropy, when all the classes were equally likely, would be approximately 2. average sharpness declines from around 1.82 in 1979 to 1.58 in 2016. Although it suggests that our model trained in 1977/1978 is better able to make prediction in earlier years opposed to later years’ reports, of which to classify paragraphs in these low taxonomy categories, the changes are not dramatic. The fact that average sharpness stays relatively constant and even declines suggests that the important lexical features in the texts that identify these general aspects (Section-levels) are present in early and latter years. However, reassuringly our model is not able to learn later signals, from the earlier texts.

Figure K.3 presents the overall accuracy for the low taxonomy models as well as the per-class

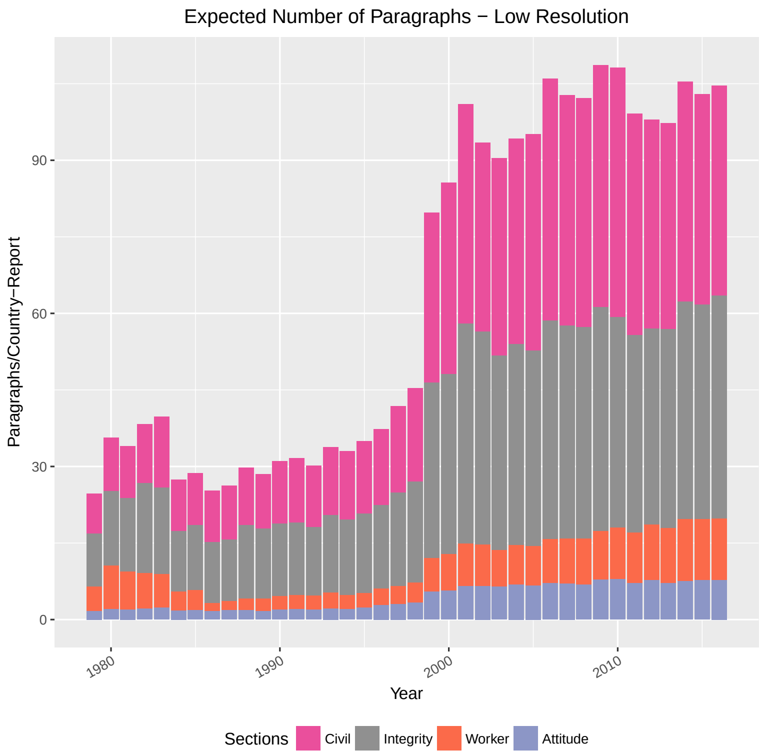


Figure K.1: The expected number of paragraphs on each low resolution aspect-category from 1979 to 2016. The sections are, "Respect for the Integrity of the Person" (gray), "Respect for Civil Liberties" (pink), "Worker's Rights" (orange), "Governmental Attitude Regarding International and Nongovernmental Investigation of Alleged Violations of Human Rights" (gray-blue).

accuracy from 1979 to 2016. The mean accuracy decreases dramatically from 0.92 in 1979 to 0.60 in 2015. This suggests that while we are identifying similar aspects in recent and past years, the content is often located in different parts of the hierarchy.

It is noteworthy that the section "Worker's Rights" decreased quite dramatically right after mid-1980 and stayed low until 2016. Substantively, the section in early years focused on the aspects of social and economic rights. Thus, a lot things were discussed under the section. Over time, as the hierarchical structure has grown and the section has paid attention to more specific and relevant rights in the section. This case emphasizes the importance of aspect-based analysis. Finding consistent aspects of human rights across time necessitates defining a given years mapping from language to its section labels, learning a model for that model, and then applying it to country reports in others years.

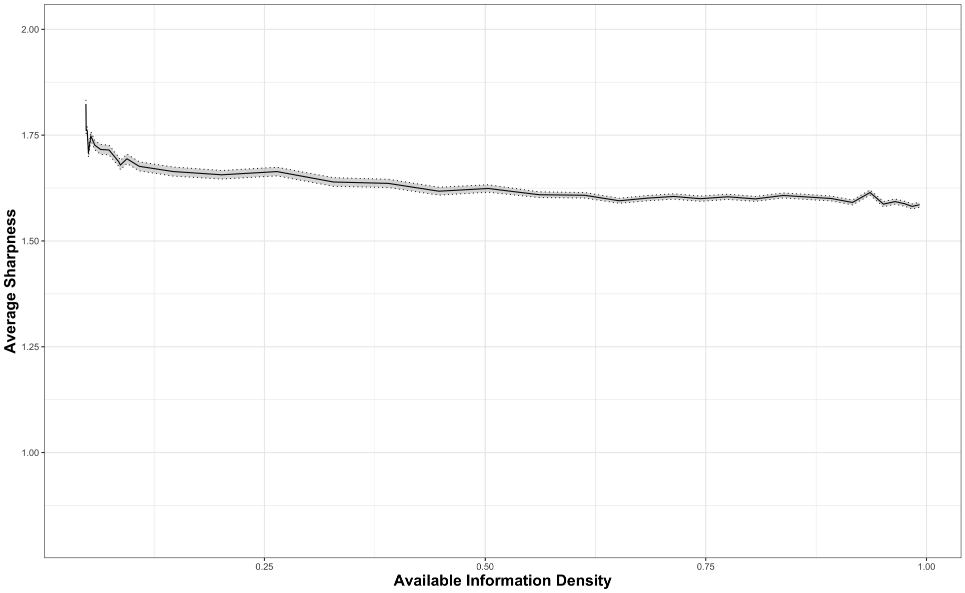


Figure K.2: Average sharpness over Available Information Density for low resolution (section) predictions.

L Proportional Reduction in Error (PRE)

Because the taxonomic structure has changed over time, not all the higher taxonomic labels in later years exist in earlier years. This makes it difficult to evaluate the model performance, which is trained in 2015/2016, over the earlier years such as accuracy rates. First, in the simplest case, there are 112 aspects in 2015/2016 that are unchanged from a leaves in year t. In this case j in 2015/2016 refers to the $i = j$ label value in year t. Here, $j \in (1, \ldots, C_t)$ indexes the a distinct aspect in year t. We use C_t to reference the number of leaves (aspects) in year t. Thus, $C_{2015/2016} = 112$ for our purposes. For these leaves, we can get an actual label, such as "Workers Rights_Minimum_Age" that matches the precision of the predicted class, as the address to the leaf did not change from t to $2015/2016$.

Second, there are leaves in $2015/2016$ that grew from ancestors in year t that themselves were leaves. In this case, each member of the set of all children of the ancestor in $2015/2016$ refers to j, the parent in year t. For example, final leaf "Physical Integrity_Arbitrary Arrest and Detention" in 1999 grew into 4 leaves in $2015/2016$. Therefore, when any of these four separate $2015/2016$ labels is predicted by the model, the only place that information could have been placed in t that

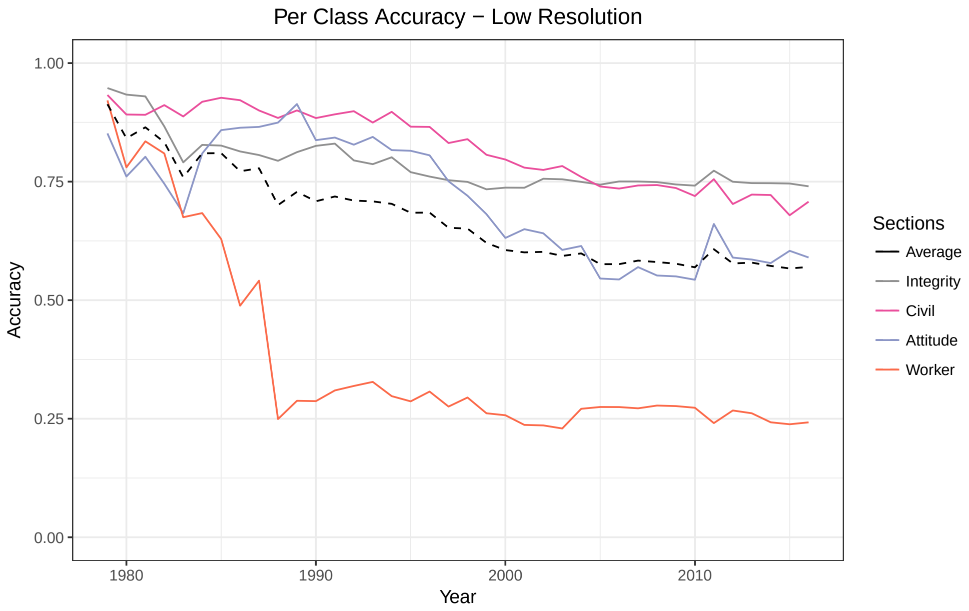


Figure K.3: The accuracy of our low resolution aspect-sentiment labels from 1979 to 2016. The mean is shown as a dotted line, and the accuracies for each of the four individual classes are plotted as solid lines of different colors. The drop in average accuracy masks different patterns across the sections.

would be consistent with this address is parent j. Thus, in these cases, parent j is repeated as a value in the dictionary for that year, where the keys are $2015/2016$ labels. If a label i does not have a matching leaf in year t but does share a parent that was a leaf in t, then our prediction of that i could only be accurate if the actual label j was that parent. If the actual label was not that specific i, then the information would be in an inconsistent place, suggesting a change in aspect-categories.

Thus we use the Best-case Proportional Reduction in Error (PRE). Call the set of 112 key-value pairings for year(t) O_t, with elements O_{it}, accessing the i_{th} set of values, or single value, that define approximately accurate labels i for year t. We define best-case accuracy, as:

$$a_t^{(bc)} = \frac{\sum_1^{D_t} \mathbf{1}[y_{djn,t} \in O_{it} | p_{idf} > p_{kdt} \, \forall \, i \neq k]}{D_t}$$

Thus as long as the actual value is in the set of approximate locations for the prediction, we count the value as accurate, in the best case.

Figure L.1 illustrates the exanple of best case accuracy. The x-axis arrays the 112 classes,

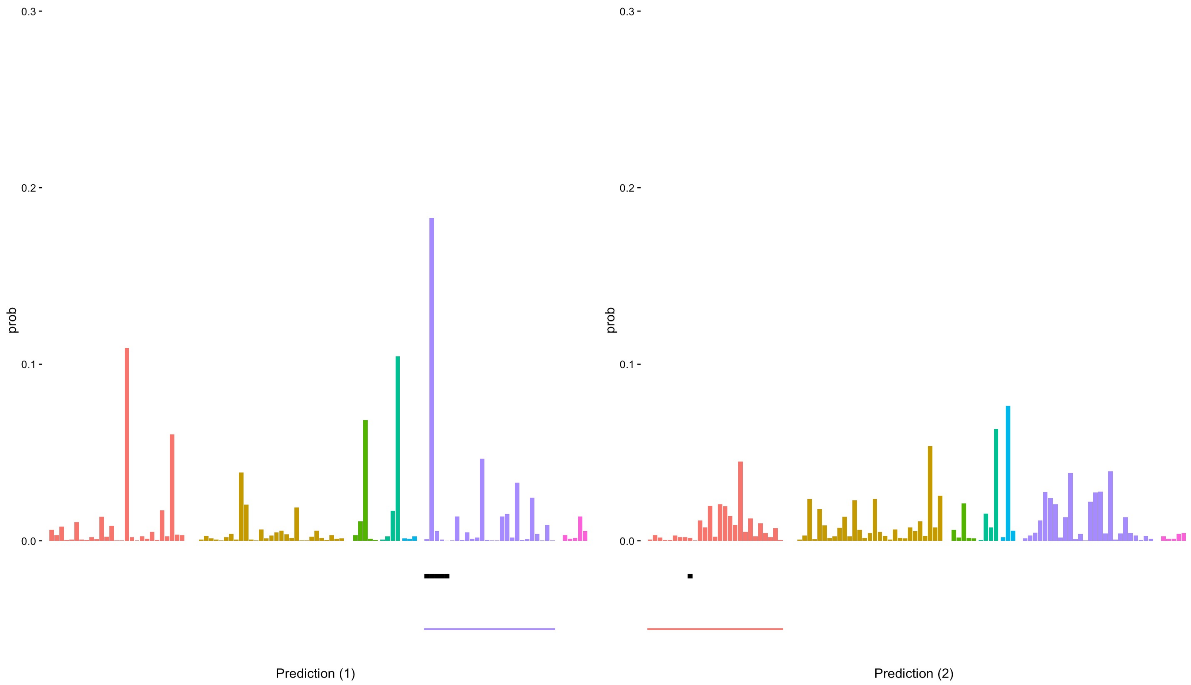


Figure L.1: Example of Best Case Accuracy

the y-axis measures the model's estimated probability that the paragraph belongs in each class. The lines below the x-axis label the actual label on the paragraphs. The lower line is the section (low resolution label) the higher black line is the specific leaf. The bars indicate high resolution predicted probabilities for two paragraphs. Colors correspond to sections. On the left, we only have an approximate location of where the 2015/2016 label would be situated. On the right, there is a one-to-one relationship between the actual label on the paragraph and the 2015/2016 labels. The example on the right would be counted as accurate in this case, because the class with the maximum prediction falls within the set of of approximate labels. In particular, the predicted class has grown from leaf where the paragraph appeared.

However, this measure needs to be corrected for random guesses. The task of predicting the labels is more difficult when there are not a set of labels that would count as accurate, as on the right in Figure L.1.

At the extreme, if we had a new sibling at the root, the set of all leaves at t would be approximately correct and best-case accuracy would be 1 by definition for every prediction. We correct for this by explicitly calculating the best-case accuracy for random predictions. We can define the random probability of $y_{djn,t} \in O_{it} | p_{idt} > p_{kdt} \; \forall \; i \neq k$:

$$r_{dt}^{(bc)} = \frac{|O_{it}|}{C_t}$$

Here $|Oit|$ refers to the length of the list of values that key i refers to. We assume for simplicity of notation that i is the predicted class. C_t again is the total number of leaves in the hierarchy at time t. When $|O_{it}| = C_t$, then the random probability is 1. When $|O_{it}| = 1$, so that there is only 1 value, the probability of being accurate, in the best-case, is $\frac{1}{C_t}$. Because $r_{dt}^{(bc)}$ depends on the predicted label from the model[26], it varies across documents. The average random accuracy across the set of documents is $r_t^{(bc)} = \sum_{d=1}^{D_t} r_{dt}^{(bc)}$

We can then correct $a_t^{(bc)}$ using the idea of the proportional reduction in error (PRE). If we have a forecast with accuracy a and a random baseline with accuracy r, then the PRE is calculated as $1 - \frac{1-a}{1-r}$. Plugging in $a_t^{(bc)}$ and $r_t^{(bc)}$, we have

$$PRE_t^{(bc)} = 1 - \frac{1 - a_t^{(bc)}}{1 - r_t^{(bc)}}$$

If $PRE_t^{(bc)}$ falls as we get to earlier time periods, this suggests that even if there were paragraphs that appeared to refer to aspect's that were consistent with specific 2015/2016 categories, they would be scrambled across inconsistent locations of the tree. This would suggest that the aspect-categories had changed over time. The research design we have summarized will provide evidence about whether a.) general-section aspects of human rights have been consistently discussed across time and b.) more specific aspects have appeared only in recent years, as would be consistent with information effects.

[26] There are different lengths across the items in the dictionary that maps 2015 labels to labels from t.

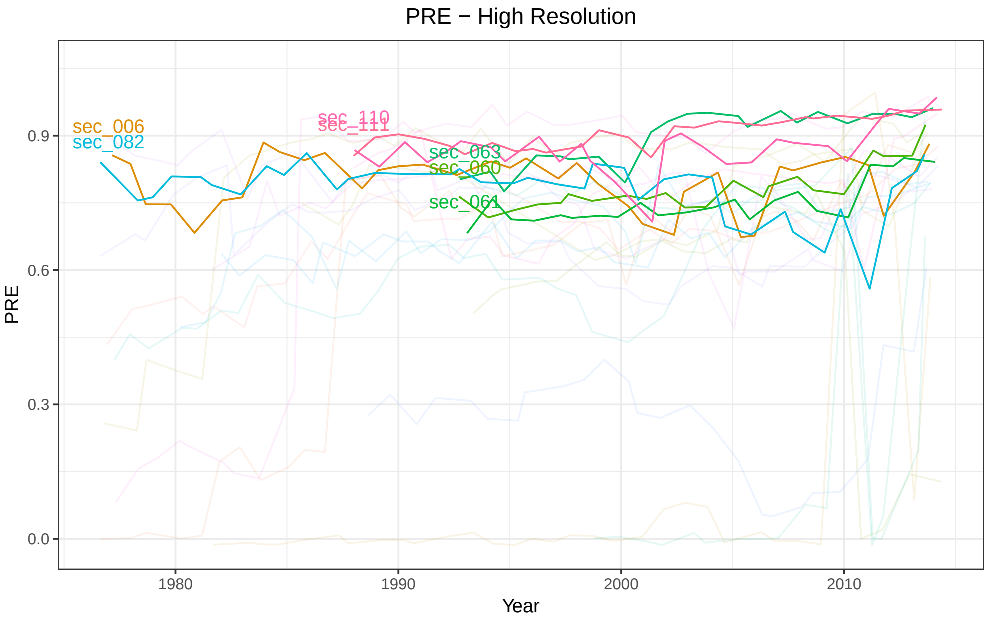


Figure L.2: Best-case proportional reduction in error (BC-PRE) for the high resolution model from 1977 to 2014. The prediction for each paragraph in a given year from the model trained on 2015/2016 leaves is compared to the actual location of the text. If the label is consistent with the location, such that it is equal or they share a parent, then it is counted as accurate. The baseline is calculated as the probability of an accurate prediction by random chance given the predicted value. Each predicted 2015/2016 leaf (112 in total) has a BC-PRE for each year, connected by a line. We highlight the 2015/2016 classes that have above average accuracy and below average variance over the time period.

While we find a mix of consistently discussed and emerging high resolution aspects over time, we find significant changes in the hierarchical placement of the aspects. Figure L.2 plots the best-case PRE values for each 2015/2016 high resolution aspect-category. Four leaves have consistently high relative accuracies over the time period, “Disappearance”, “Torture”, “Arbitrary or Unlawful Deprivation of Life”, and “Denial of Trial”. All four are within the physical integrity rights section. Perhaps even more interestingly, all 4 leaves existed in 1999. While both “Torture” and “Denial of Trial” evolved children, the State Department kept writing general statements in a leaf dedicated to these aspect-categories. For the future, this suggests that we can, in certain instances, locate even semantically general aspect-categories in fine-grained leaves and that these consistent categories could also be useful in providing consistent terms of comparison for countries behaviors over time. Another 2015/2016 high resolution leaf that has relatively consistent best-case PRE is “People with

Disabilities" in the Discrimination section, again, this is a leaf from 2015/2016 that existed going back to 1999. Other than these cases, most predictions for aspect categories perform no better than random guesses going back only a few years in time. This suggest that the even high resolution aspects that have been discussed, and unlabelled in the past, may not have been judged within the section/branch of the hierarchy, in which it eventually appears. This finding lends further credence to the use of the content of paragraphs to learn the aspects being discussed, instead of relying solely on the section and hierarchical address discussed above.

M Additional Results for Reports Released by Amnesty International and Human Rights Watch

Below are the implicate nodes in for the reports released by Amnesty International and Human Rights Watch. For Amnesty we use the same years as the comparison of the State Dept. in the main text. For Human Rights Watch the first set of press releases we have are from 1997, so that is used for the first plot.

M.1 Structure

A similar picture emerges in the change in taxonomic structure for Human Rights Watch and Amnesty International over time. There are few nodes systematically covered in the past, with more in recent years. We define a node as existing in the implicit Amnesty International taxonomy in a given year if that node has more than .3 expected paragraphs per country-report. If we used the same threshold as in the State Department report analysis, then almost no nodes would be present in most years. We get very similar reports when we use a cut-off of 40 paragraphs, un-normalized by the number of countries in the reports over time. For the Human Rights Watch implicit taxonomy over time, we include a node when there are 5 or more press releases expected on that right in a given year. This threshold takes into account that press releases are different forms of monitoring than annual reports.

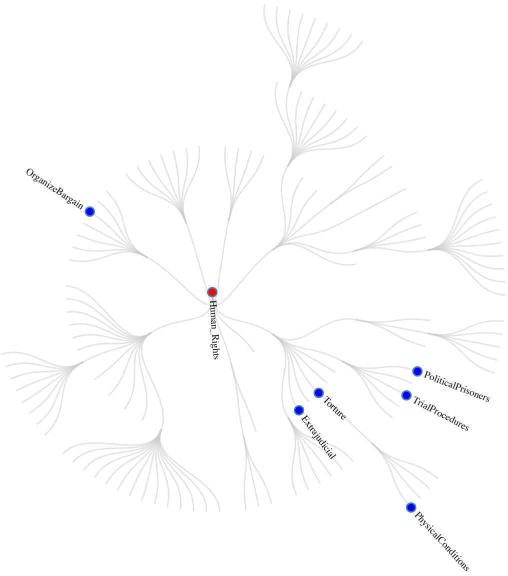


Figure M.1: Amnesty International 1977

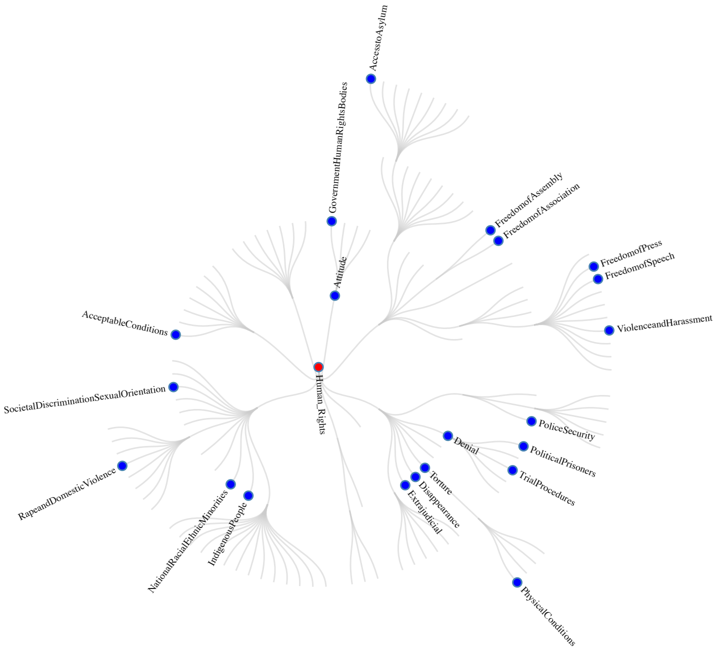


Figure M.2: Amnesty International 2016

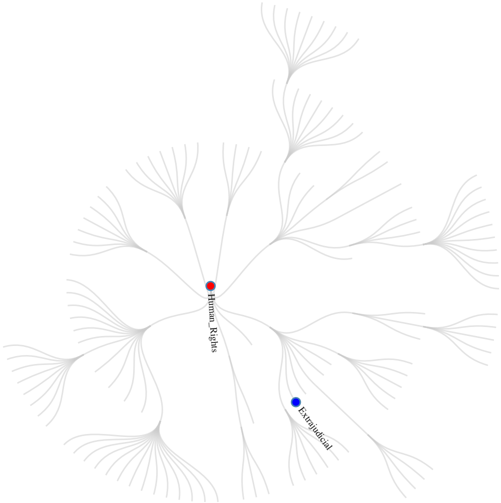


Figure M.3: Human Rights Watch 1997

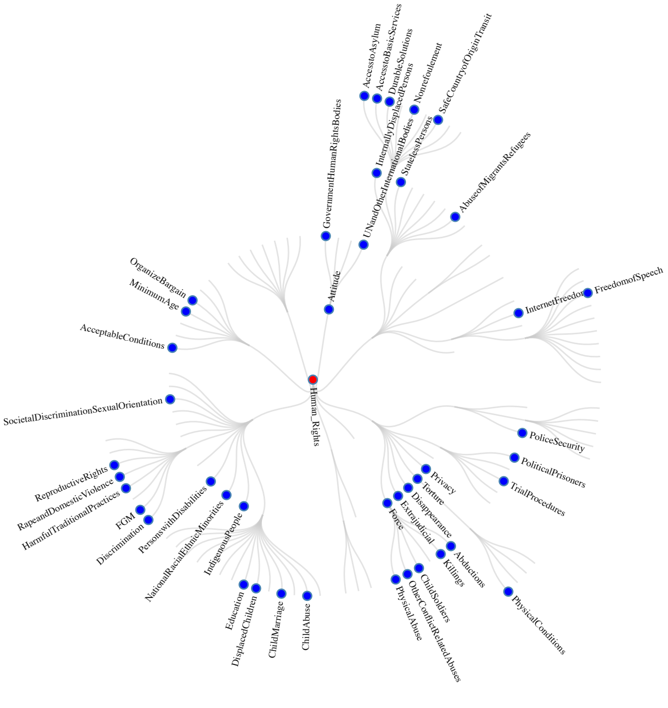


Figure M.4: Human Rights Watch 2016

With multiple implicit taxonomies across monitoring agencies we can now look at difference in coverage of given countries in a specific year. For example, below we illustrate how researchers can use our tools and data to compare State Department (red for more coverage) and Amnesty International (blue for more coverage) coverage of Iran in 2014. We note that the State Department coverage is a super-set of the Amnesty coverage, as there are no blue nodes, pictured, but only red (State Department coverage only) and pink (both agencies covered right) nodes.

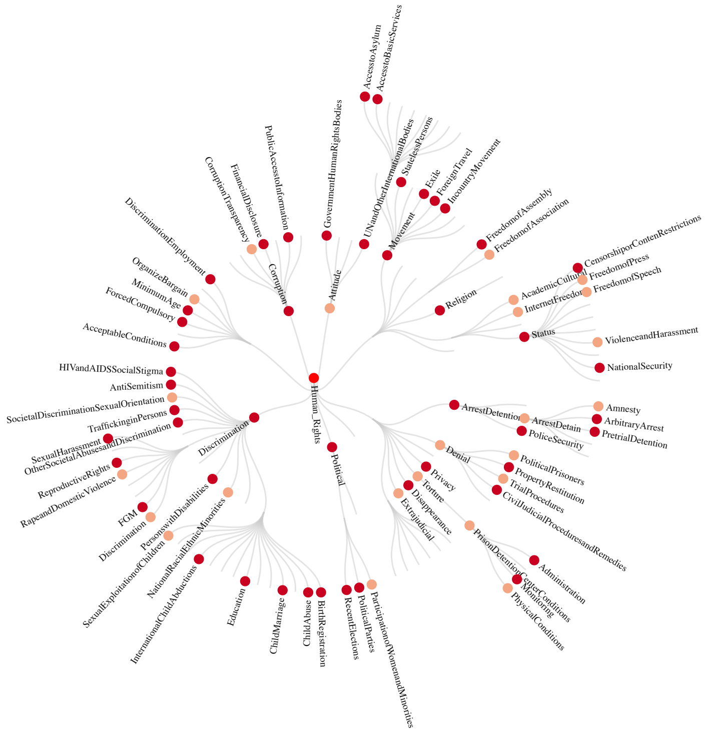


Figure M.5: Implicit Nodes for State Dept. and Amnesty International for Iran in 2014. Red nodes are contained only in the State Dept. report, blue nodes are contained only in Amnesty International reports, and pink nodes are contained in both.

As we discussed in the main paper, we summarize the changes in the average depth of the leaves in the implicit aspect-category hierarchies for each year for Amnesty International (Figure M.6) and Human Rights Watch (Figure M.7). The y-axis plots the average depth (levels down from the root) across the top-level sections of the document in each year. To calculate the average depth of the implicit taxonomy for each year, first, we identify all the nodes where the sum of the (scaled) predicted probabilities for all texts in each year is larger than the respective threshold discussed above. Second, we count the maximum number of all the nodes where they branch out to the next child nodes for each section, then average them by the total number of possible sections. The x-axis represents the number of final nodes/concepts, the most specific concepts labeled in the texts, in a given year. The plot illustrates the path of these implicit taxonomies across time. As in the case of the State Department, there is a clear movement upwards, as the later documents have both increasingly grown rights as distinctions within existing rights, adding complexity and depth to the previous taxonomy. This suggests that the annual reports from Amnesty International and the press releases from Human Rights Watch contain information on violations and protection of a few specific human rights, even if they did not have an explicit section label in that year.

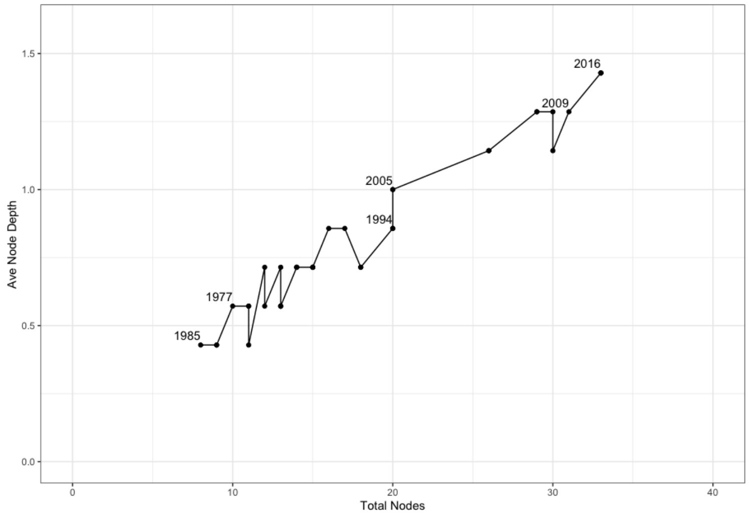


Figure M.6: Amnesty International and Implicit Node Depth Changes: A scatter plot of the total number of leaves in each annual aspect-hierarchy (x-axis) and the average depth of leaves across the section (first level below the root). The points are jittered to avoid over-plotting.

M.2 Attention

Below we present bar plots for the amount of attention for Human Rights Watch and Amnesty International. Colors are keyed to the State Department Attention plots so that a smooth gradient of color in a plot suggests a similar ordering of attention in that section as compared to the State Department, while a jumble of color suggest distinct orderings of coverage. Attention generally

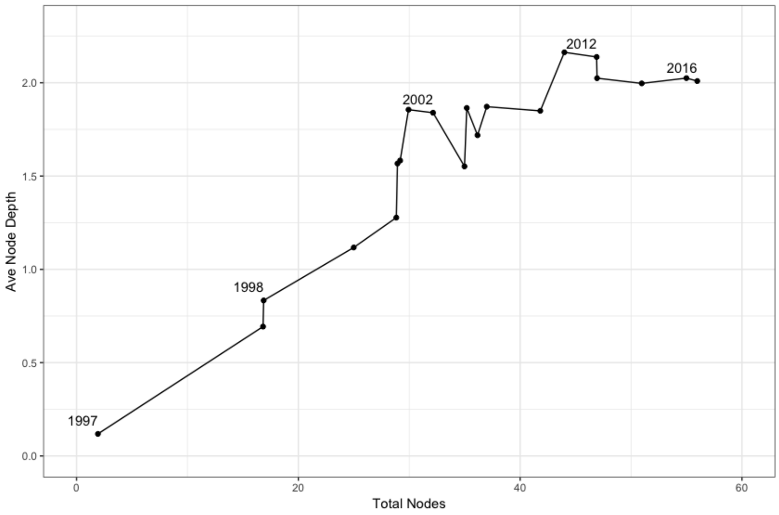


Figure M.7: Human Rights Watch and Implicit Node Depth Changes: A scatter plot of the total number of leaves in each annual aspect-hierarchy (x-axis) and the average depth of leaves across the section (first level below the root). The points are jittered to avoid over-plotting.

rises in most of the plots. The only exceptions are plots where there is very little attention at all, such as Amnesty International Corruption section Political Rights section (y-axis max is very small for these).

M.2.1 Human Rights Watch

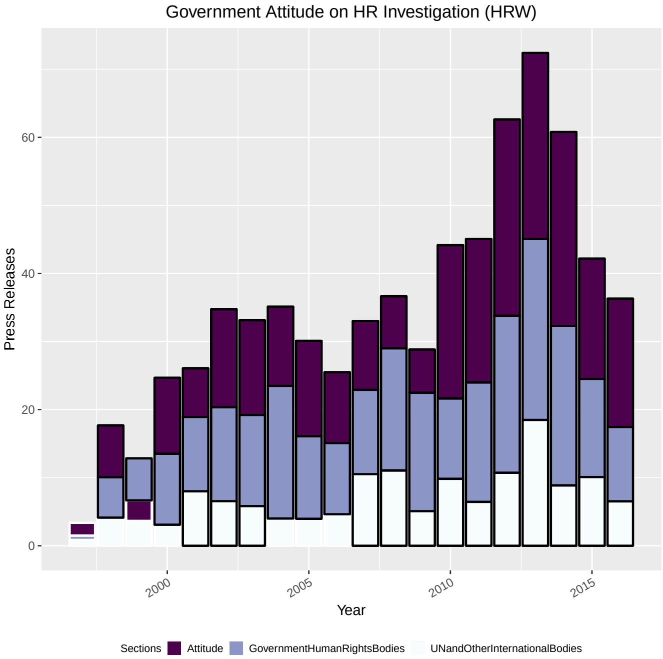


Figure M.8: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

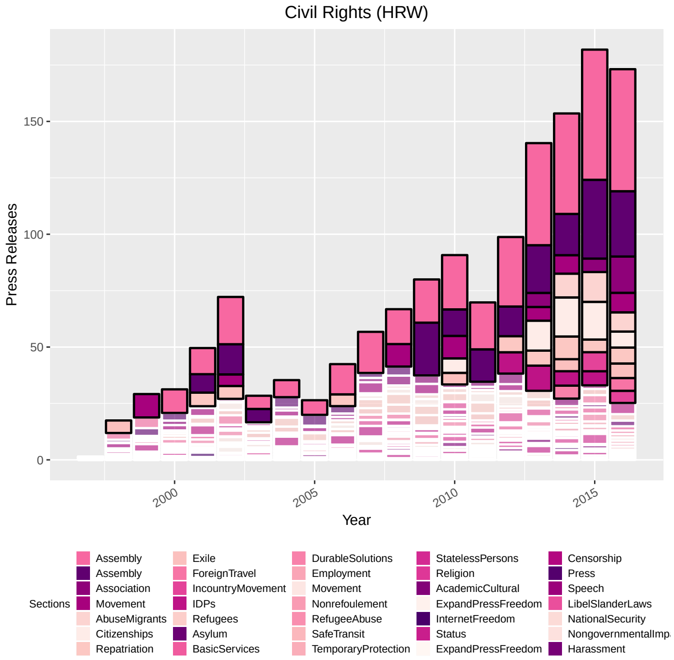


Figure M.9: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

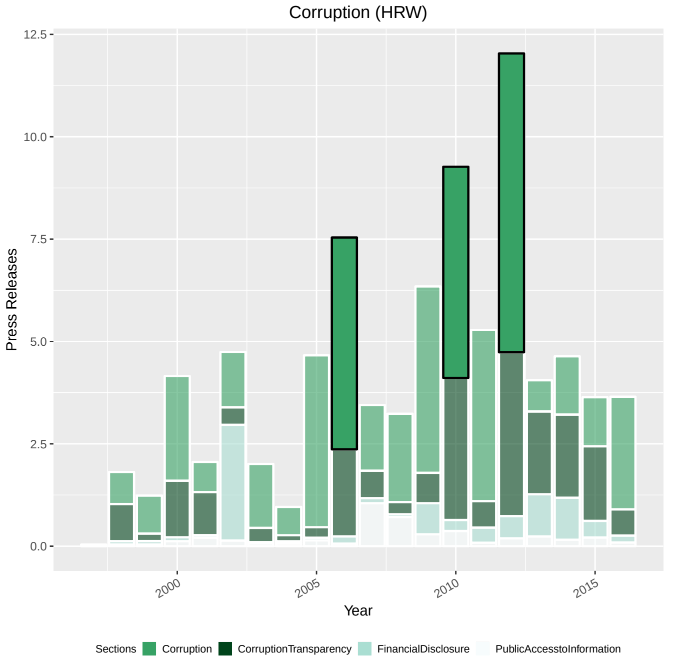


Figure M.10: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

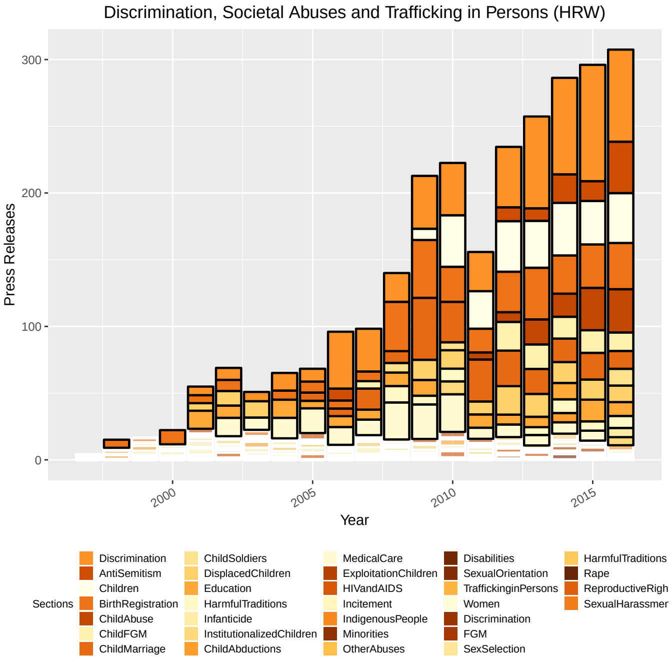


Figure M.11: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

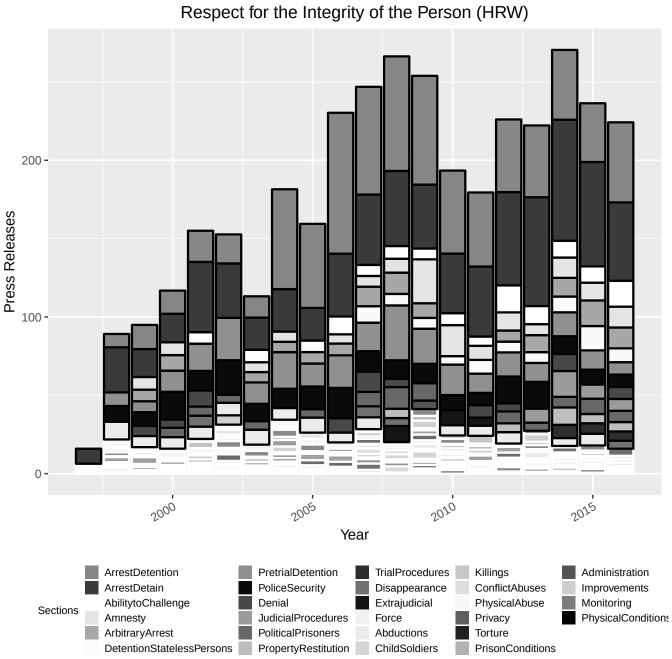


Figure M.12: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

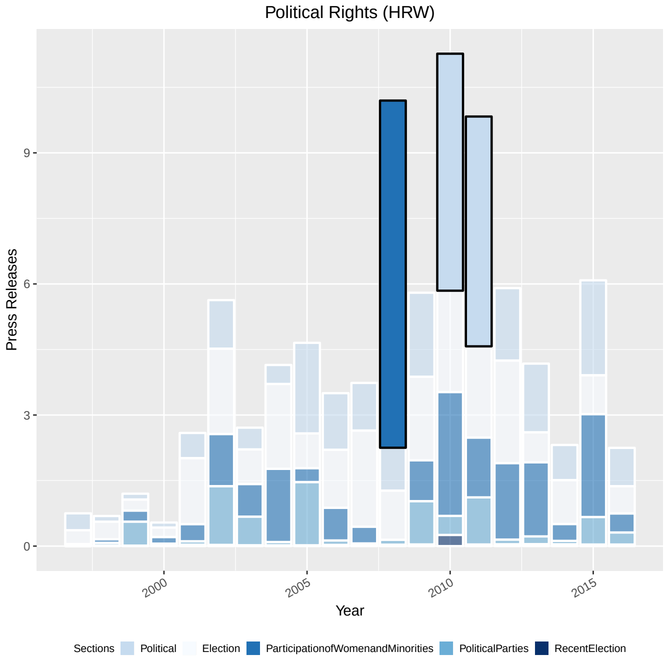


Figure M.13: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

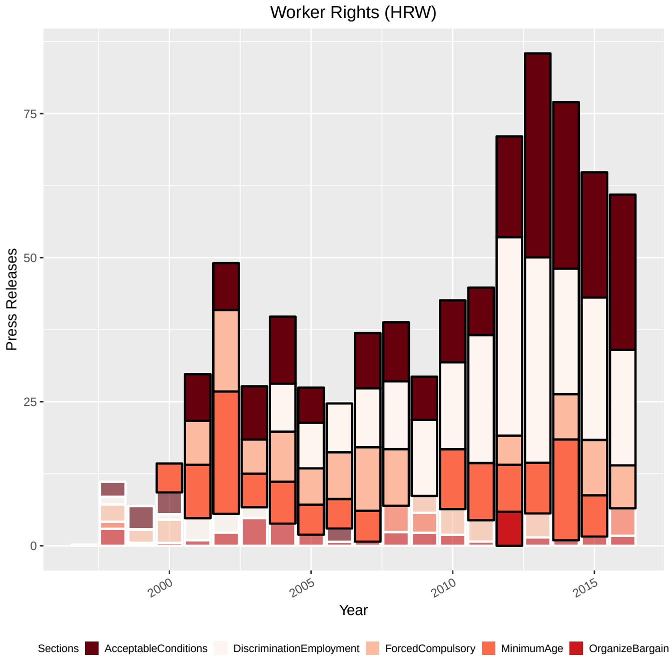


Figure M.14: Expected number of press releases estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than 5 expected releases in that year, and white otherwise

M.2.2 Amnesty International

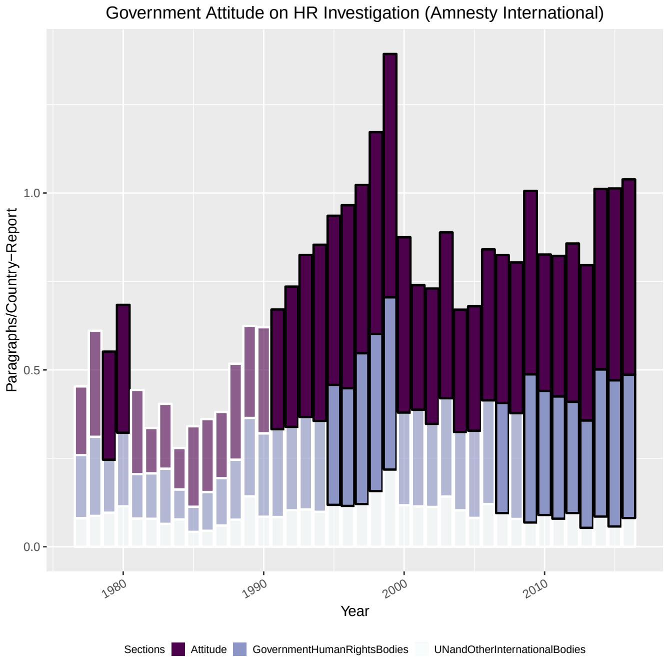


Figure M.15: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

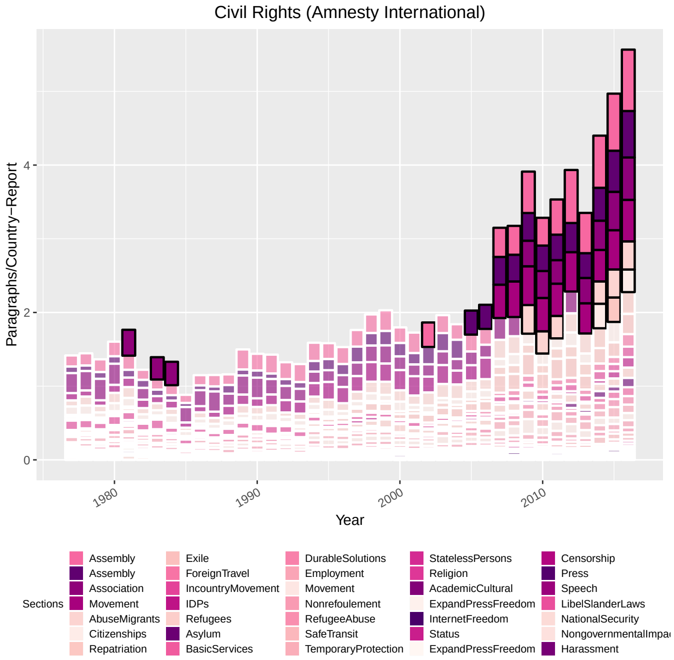


Figure M.16: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

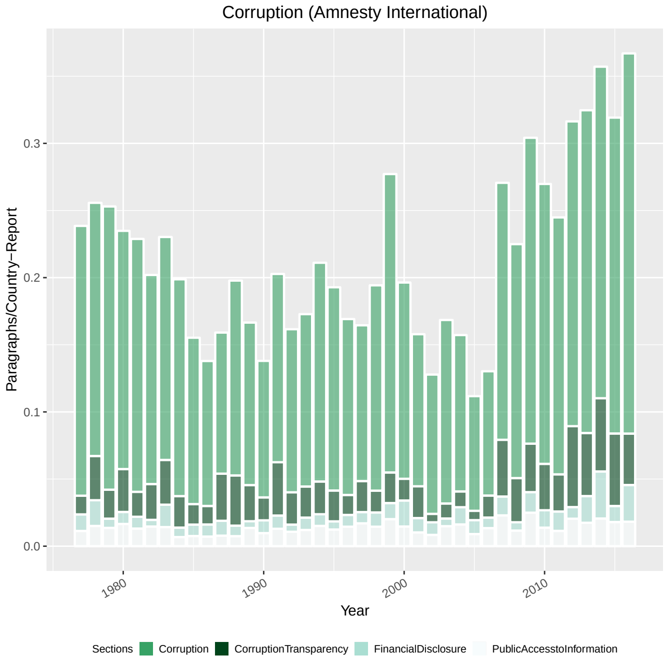


Figure M.17: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

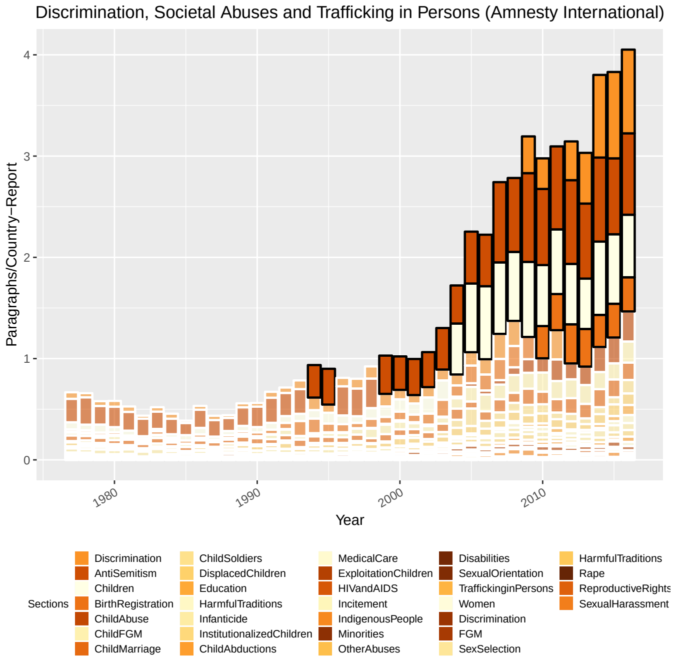


Figure M.18: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

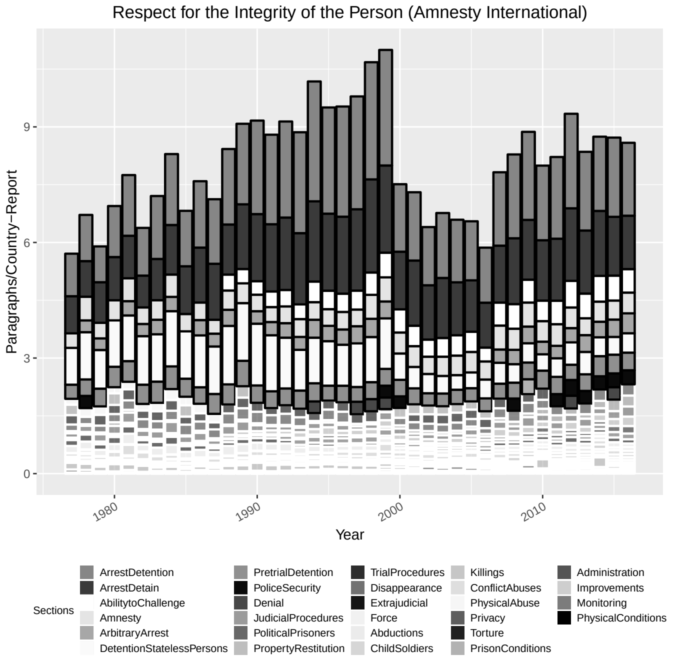


Figure M.19: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

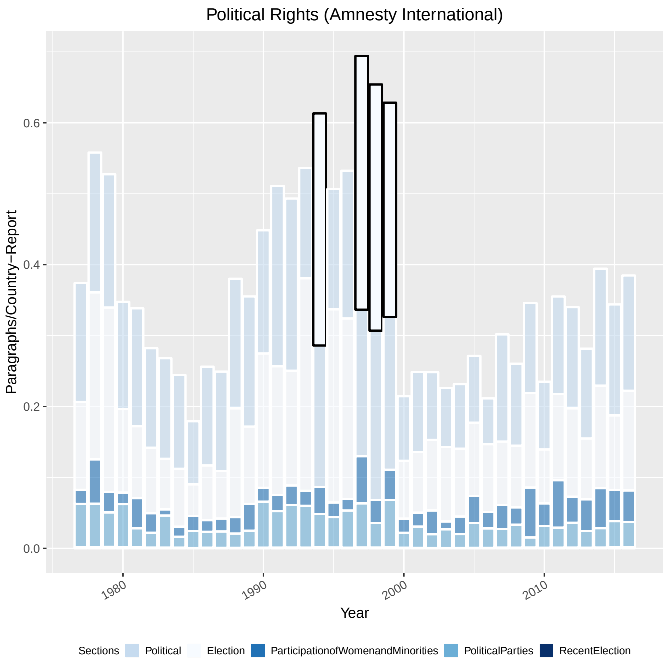


Figure M.20: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

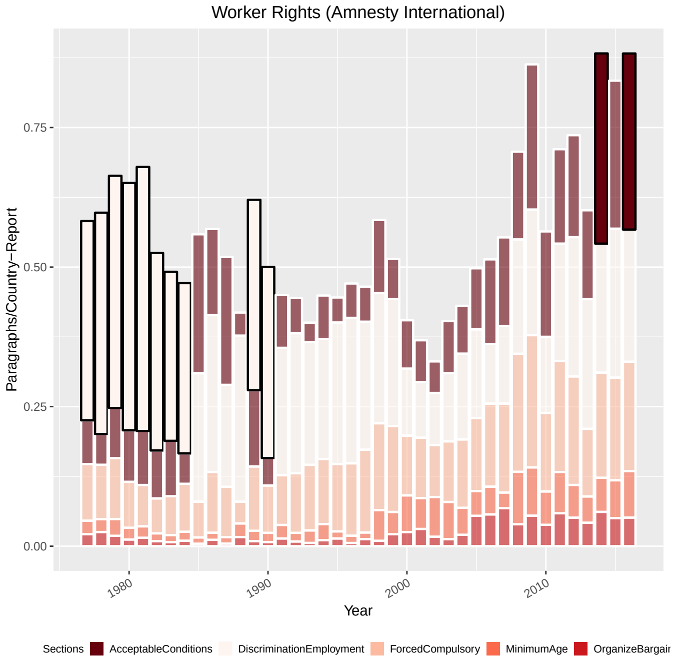


Figure M.21: Expected number of expected paragraphs estimated with the 2015/2016 implicit taxonomy. The outline/border of the bars is black where there are more than .3 expected paragraphs in that year, and white otherwise

M.3 Sharpness

As discussed in the main paper, we present the plot of the average sharpness for Amnesty International (Figure M.22 and Human Rights Watch (Figure M.23 over the Available Information Density (AID). The results we show here for sharpness are conservative. We use the same y-scale as in the State Department plot above. However, given the fewer number of rights being judged by these organizations in their texts (almost 1/3 less), the maximum could be drawn lower and minimum much higher. This would illustrate more dramatic changes over time (and information). However, we present the more conservative plots here because a) they still illustrate the consistent upwards trends, and b) we believe it is better to be conservative with these inferences since the same detail of explicit section/sub-section meta-data was not available for Amnesty International and Human Rights Watch as we had access to for the State Department corpora.

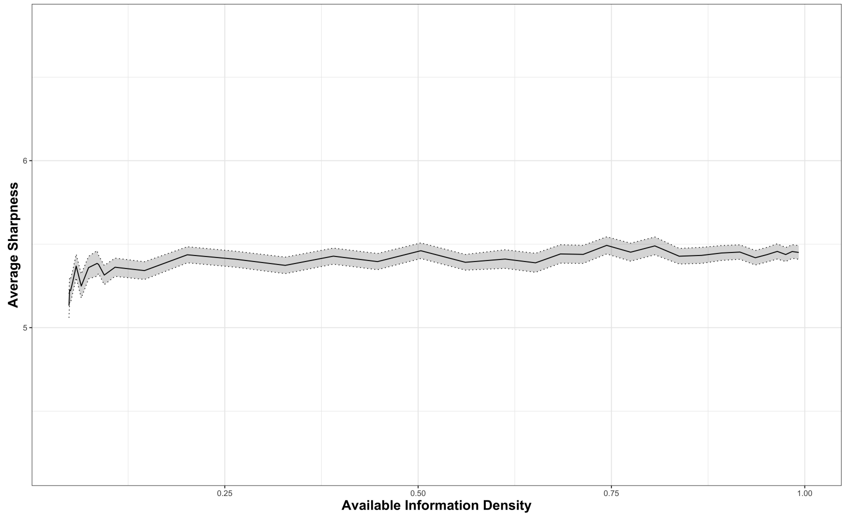


Figure M.22: Amnesty International and Average Sharpness and Available Information Density: The average sharpness of our predictions of the rights in every paragraph in a given year. Higher values reflect more information on distinctions between concepts, and lower values suggest that information on the high resolution human rights taxonomy are missing. The maximum of the y-axis is set to the theoretical maximum average sharpness. The minimum is set to the average sharpness of a classifier that simply randomly assigns a label based on the relative frequency of the locations in the training set. The dotted lines represent standard error.

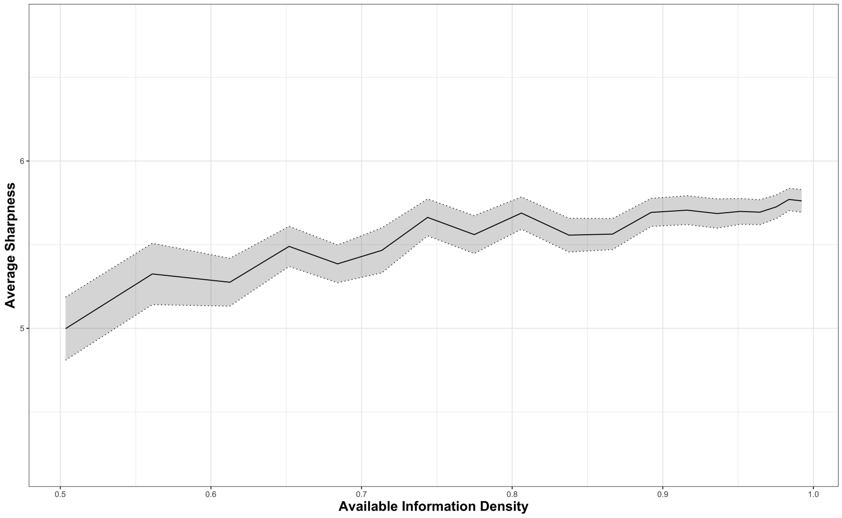


Figure M.23: HRW and Average Sharpness and Available Information Density: : The average sharpness of our predictions of the rights in every paragraph in a given year. Higher values reflect more information on distinctions between concepts, and lower values suggest that information on the high resolution human rights taxonomy are missing. The maximum of the y-axis is set to the theoretical maximum average sharpness. The minimum is set to the average sharpness of a classifier that simply randomly assigns a label based on the relative frequency of the locations in the training set. The dotted lines represent standard error.

N Sources of Human Rights Documents: The Importance of the State Department Reports

As discussed in the main paper, we first use *the Annual Country Reports on Human Rights Practices* from the U.S. State Department to analyze changes in information effects. Because we are interested in identifying and examining the description and classification of new human rights concepts, sub-concepts, and others over time, we need to use the most comprehensive source of documents that describes 1) all the internationally recognized human rights as set forth in the Universal Declaration of Human Rights and other international human rights treaties, not just a few human rights issue areas; 2) a system of classification for different human rights issue areas (i.e. separate sections for each issue); 3) most of the countries in the world, not just a few regions or countries; 4) for the past 30 years at least (for analysis over time). Table N.1 shows other sources of human rights documents.

Amnesty International Amnesty International, one of the most prominent international human rights NGOs, has issued annual reports since 1961. 1) Its main emphasis was individual prisoners of conscience, and it has extended its mandate and covered other human rights issues. 2) Although Amnesty's annual reports cover a variety of human rights issue areas, the structure of the reports is less organized. Until 1999, Amnesty Reports did not have any sections or subsection for different issue areas. Because Amnesty Reports simply list a series of issues for different countries, it often lacks the consistent categories of human rights. Moreover, the reports often pay attention to individual cases and uses proper nouns for categorization (section headings), not more general human rights terms. Thus, it discusses related human rights problems as if they were completely distinct issues, which makes it difficult to identify issue areas of the reports over time. 3) Amnesty Reports do not cover all the countries in the world, particularly in early years. For example, the reports covered 138 countries in 1990 whereas the State Department covered 170 countries.

Source Name		Years	Mandate	State Coverage
Amnesty International		1961-2018	- Armed Conflict - Arms Control - Climate Change - Corporate Responsibility - Death Penalty - Detention - Disappearances - Discrimination - Freedom of Expression - Indigenous Peoples - International Justice - Living in Dignity - Refugees, Asylum-Seekers and Migrants - Sexual and Reproductive Rights - Torture - United Nations	All states
Human Rights Watch		1989-2018	- Arms - Business - Children's Rights - Disability Rights - Environment - Free Speech - Health - International Justice - LGBT Rights - Migrants - Refugee Rights - Terrorism - Torture - United Nations - Women's Rights	All states
UN Commission on Human Rights		1947-2005	- UN Charter - UDHR - Human Rights Treaties	Selected/Targeted states only
UN Human Rights Council		2006-2019	- UN Charter - UDHR - Human Rights Treaties	All states
UN Human Rights Treaty Bodies	CCPR	1966-2019	- Civil and Political Rights	Member states only
	CESCR	1966-2019	- Economic, Social and Cultural Rights	
	CAT	1984-2019	- Torture, Cruel, and Other Inhumane Treatment	
	CRC	1989-2019	- Children's Rights	
	CMW	1990-2019	- Migrant Workers' Rights	
	CED	2006-2019	- Enforced Disappearances	
	CERD	1965-2019	- Racial Discrimination	
	CRPD	2006-2019	- Rights of People with Disabilities	
	CEDAW	1979-2019	- Women's Rights	

Table N.1: Sources of Human Rights Documents

Human Rights Watch Human Rights Watch (HRW) is another well known transnational human rights organization and has published its annual World Reports since 1989. 1) Like Amnesty, HRW also covers a variety of human rights issues. 2) However, until 2005, it did not have any categorization of different human rights issue areas. And it covers only a few selected issues for each country. 3) Although it tries to cover most of the countries in the world recently, it did not document all the countries. For example, in 1995, HRW only covered 62 states whereas the State Department covered over 180 countries.

UN Commission on Human Rights As the charter-based human rights body, the Commission on Human Rights (UNCHR) was the center within the UN for promoting respect for human rights. One of the Commission's main activities was providing information about states' compliance to international human rights standards. By passing official resolutions or providing technical assistance and advisory services, the Commission published session reports since 1947. 1) The UNCHR covered primarily political and civil rights, and very little about economic and social rights (Forsythe, 2009). 2) Moreover, the commission's reports are often vague. For example, the 1503 procedure was a confidential procedure in which states were examined in closed session. Thus, any specific allegations/justification for consideration were not made public. 3) The main problem with the Commission that it did not review all the countries but a few selected states based on member states' political motivations. This was heavily criticized for having unreasonable double standards. For example, countries not renowned for its human rights practices like China, Syria and Libya were elected to the Commission's members whereas the U.S. was dismissed from the seat. Thus, the Commission covered only a few selected (politically motivated) states in their sessions and following reports. 4) Because of this serious problems, it ceased to exist since 2005 and replaced by the Human Rights Council.

UN Human Rights Council In 2006, the Commission was dissolved and replaced by the Human Rights Council (UNHRC). 1) The Council reviews pretty much all the human rights issues and 2) has published its session reports and the Universal Periodic Review (every 4 years for all the countries). 3) Unlike the Commission, the Council reviews all states and provide information on the extent to which states adhere to international human rights standards. 4) But, it has only 13 years of reports.

UN Human Rights Treaty Bodies There are a number of international human rights treaties and monitoring mechanism in the UN. As of 2019, the International Covenant on Civil and Political Rights (1966), Economic, Social and Cultural Rights (1966), Torture (1984), Racial Discrimination (1965), Women (1979), Child (1989), Migrant Workers (1990), Disabilities (2006), Disappearances (2006) monitoring bodies have published reports. These reports could be used for analysis, but there are a few limitations. 1) Because each monitoring committee focuses on different mandates and the UN human rights treaty bodies are not centralized and poorly coordinated (Forsythe 2018), their reports are also decentralized. 3) All these treaty bodies are based on member states that ratified the treaties, state coverage is not consistent. More to the point, all member states have obligation to submit regular reports to monitoring committees on their practices. However, these committees have struggled first with the problem of states failing to submit even initial reports, although legally required. This problem is widespread across the UN system of human rights reporting. Many states' reports, even when submitted, are more designed to meet formal obligations than to give a full and frank picture of the true situation in the country (Forsythe, 2017).

References

Bagozzi, Benjamin and Daniel Berliner. 2016. “The politics of scrutiny in human rights monitoring: Evidence from structural topic models of U.S. State Department Human Rights Reports.”.

Barner, David and Andrew Scott Baron. 2016. *Core Knowledge and Conceptual Change*. New York: Oxford University Press.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li and Allen Riddell. 2017. “Stan: A probabilistic programming language.” *Journal of statistical software* 76(1).

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.

Colaresi, Michael P and Zuhaib Mahmood. Forthcoming. “Lessons from Machine Learning to Improve Conflict Forecast.” *Journal of Peace Research* .

Cordell, Rebecca, K Chad Clay, Christopher J Fariss, Reed M Wood and Thorin M Wright. 2019. “Changing Standards or Political Whim? Evaluating Changes in the Content of the US State Department Human Rights Reports.” *Journal of Human Rights* .

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

Forsythe, David P. 2009. *Encyclopedia of human rights*. Vol. 1 Oxford University Press.

Forsythe, David P. 2017. *Human rights in international relations*. Cambridge University Press.

King, Gary and Will Lowe. 2003. "An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design." *International Organization* 57(3):617–642.

Liu, Bing. 2012. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5(1):1–167.

Nielsen, Finn. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts'*. pp. 93–98.

Pang, Bo and Lillian Lee. 2008. "Opinion mining and sentiment analysis." *Foundations and trends in information retrieval* 2(1-2):1–135.

Park, Baekkwan, Michale Colaresi and Kevin Greene. 2018. "Beyond a Bag of Words: Using PULSAR to Extract Judgments on Specific Human Rights at Scale." *Peace Economics, Peace Science and Public Policy* .

Pontiki, Maria, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. Citeseer pp. 27–35.

Small, Ernest. 1989. "Systematics of Biological Systematics (Or, Taxonomy of Taxonomy)." *Taxon* 38(3):335–356.

Stace, Clive A. 1991. *Plant taxonomy and biosystematics*. Cambridge University Press.

Stan Development Team. 2018. "PyStan: The Python interface to Stan v 2.17.1.0." http://mc-stan.org.