Using Machine Learning to Assess Amnesty International Human Rights Reporting

Sarah Hunter
University of Georgia

Lucas Nussbaumer
University of Georgia

Introduction

Over the last several decades, human rights have been brought to the forefront of global politics. Much of this is the work of several non-governmental organizations (hereafter NGOs). NGOs, such as Amnesty International (hereafter AI) or Human Rights Watch, are instrumental in the reporting and providing the basis of enforcement for human rights around the globe. In fact, many people from politicians to academics to activists regularly use the human rights reports generated by NGOs to make policy, produce research, and campaign. However, are these reports to be trusted for such endeavors? Some scholars have said "yes" (Hill et al., 2013; Cingranelli et al., 2014), while others remain skeptical. Ron et al. (2005) find that the content of reports, while remaining accurate, is shaped by the political environment in which they are given. Many of these studies use sophisticated regression analysis or qualitative techniques to determine the levels accuracy or bias in Amnesty International's reports.

However, states are not the only actors to commit human rights abuses. Many non-state actors are also human rights abusers. Much research has been conducted on the reporting of state human rights abuse, but this is not the complete picture. There is a gap in the literature on NGO reporting on non-state actors compared to state actors. Also, machine-learning approaches have yet to be widely applied for this area of research. Therefore, we will use both topic models and sentiment analysis to assess the level of bias, if any, that can be found in Amnesty International?s reporting on non-state actors. This paper seeks to fill both gaps in the literature. We use machine learning to examine the content of Amnesty International's reporting on the human rights practices of a state in the presence of a violence perpetrated by non-state actors.

NGOs in International Relations

Nongovernmental organizations (hereafter, NGOs) have shown to have an effect on state behavior. NGOs are non-profit "legally constituted organization created by private persons or organizations without participation or representation of any government" (UIA, 2015). They differ dramatically in size, budget, membership, and scope. Often times, the power in the NGOs comes when they act through Transnational Advocacy Networks (TANs) (Keck and Sikkink, 1998). TANs are a network of states, local NGOs, international NGOs, and intergovernmental organizations that all work toward a similar end. TANs are especially prevalent in the areas of human rights and environmental protection. In this model, local NGOs are blocked from making changes by the regime. In response, they reach out to international NGOs who encourage third party states or intergovernmental organizations to pressure the targeted state's regime. This is called the "boomerang model" of international relations (Keck and Sikkink, 1998). The series of repeated iterations

of the boomerang model required for targeted states to internalize a norm is the "spiral model" (Risse and Sikkink, 1999).

Information politics is an NGO's most used and most powerful tool. Keck and Sikkink (1998) define information politics as "the ability to quickly and credibly generate politically usable information and move it to where it will have the most impact"(16). International NGOs (hereafter INGOs) play two roles in information politics: "local population empowerment" and "information production" (Murdie and Peksen, 2013). International NGOs work with local NGOs to support them with resources and training to more effectively pressure the regime at the grassroots level as well as distributing information for the global community to provide pressure on the regime by third party states and international organizations. Usually, the NGOs turn to "naming and shaming" or "naming and blaming" to force a targeted regime to change policy (Murdie and Davis, 2012). There is much empirical evidence to suggest that NGOs are successful in this endeavor. For example, shaming by NGOs have led to the punishment of states with bad human rights records by the imposition of sanctions (Murdie and Peksen, 2013) or lower foreign direct investment (Barry, Clay, and Flynn, 2012). NGO shaming has even led to an increased likelihood of humanitarian intervention (Murdie and Peksen, 2014). Not all shaming events are equal, however. In the next section, we look at the literature on the relationship between NGOs and domestic protest.

NGOs and Domestic Protest

Reporting human rights abuses are not the only dimension of advocacy for NGOs. Violent mass protests and dissent has exhaustively been dissected (Feaver, 2003). The literature on revolutions, rebellions, military dissent and coups is vast. As is the literature on the impact of NGO advocacy. Murdie and Bhasin (2011) find that human rights international NGO activity actually increases the likelihood of domestic violent and non-violent protest. This occurs as a result of the domestic population's commitment to human rights. However, does this research also have implications for human rights reporting? This article seeks to expand on former academic work and expand on a new dimension: NGOs reporting of domestic protest.

Autocracies are widely recognized to be the worst perpetrators of human rights abuses (Davenport, 2007; Mason, 2004). However, autocracies have to be distinguished from one another. Rather than treating them like equal entities, autocracies have different features that may or may not influence the way they utilize repression against their population (Geddes et al., 2014). The underlying logic of why and how political leaders allocate their resources and use different methods to maintain their power is linked to their desire of political survival (Bueno de Mesquita et al., 2004). Gandhi (2008) provides up until today one of the most comprehensive frameworks analyzing institutions in autocracies and reveals that legislative and

partisan institutions play a major role for how dictators rule their country. Davenport (2007) lays out a comprehensive approach how repression and political order are connected and what the current puzzles are.

The literature on mass protest has expanded rapidly over the last decade. More and more different types and structures within the protest movements have been analyzed and linked to their evolution, duration and characteristics (Schock, 2004). Chenoweth and Stephan (2011) brought forward the argument that non-violent conflicts are by far underrepresented in the analysis of civil resistance and follow it up with their own dataset to draw a more comprehensive picture of violent and non-violent conflicts (Chenoweth and Lewis, 2013). Furthermore, Nepstad (2011) underpins this endeavor with empirical analysis on how and why non-violent struggles succeed or fail.

The core of this article is human rights abuses or, precisely, physical integrity rights and how Amnesty International reports on them amidst violent and nonviolent protest. These rights have often been the center of academic work in international law, political science and beyond. Several datasets exist on the multitude of abuses (Cingranelli and Richards, 2010; Wood and Gibney, 2010). The link between mass protests and repression or human rights abuses appears to be persistent (Carey, 2010; Ritter and Conrad, 2016). Scholars have pointed out the different ways dictators utilize repression and violations of physical integrity rights to consolidate their rule in times of uprisings and protests (Conrad, 2014; Gaub, 2014; Mason, 2004). The core argument of this article, however, is that different types of civil resistance movements produce a different style of reporting by the NGO Amnesty International.

Information Politics and the Power of NGOs

However, some scholars have suggested that NGOs are not as altruistic as earlier authors have suggested. These skeptical authors argue that NGOs are strategic actors with their own set of preferences (Cooley and Ron, 2002; Bob, 2002). In this context, NGOs choose certain projects or locations based on what benefits the NGO the most. Bob (2002) argues that NGOs gravitate toward issues and places that are already getting a lot of attention in order to attract funds. Other scholars regard NGOs as interest groups that seek to efficiently use scarce resources and political opportunities to achieve social change (Sell and Prakash, 2004; Bloodgood, 2011). Regardless, NGO information is still considered to be accurate (Edelman 2003), and is widely used to monitor issues such as human rights (Poe, Carey, and Vazquez, 2001). Amnesty International, in particular, is a popular source of global human rights information (Eldelman, 2003; Poe, Carey, and Vazquez, 2001).

Amnesty International's brand and prestige in the human rights sector gives it a lot of power in information politics. However, does Amnesty International use this to its advantage or do they maintain

their integrity, which is an important part of their mission statement? Amnesty International's website claims: "Our experts do accurate, cross-checked research into human rights violations by governments and others worldwide" (Amnesty International, 2017). Accuracy plays an important role in AI's reporting. However, what about their organizational incentives? Organizations like Amnesty rely on donations to continue their work. According to the theory of material motivation of NGOs, Amnesty International has an incentive to exaggerate of over report on certain countries to get more media attention and therefore gather more donations (Cooley and Ron, 2002). Yet, others maintain that Amnesty International maintains carefully accurate reports, not influenced by organizational incentives (Ron et al., 2005; Hill et al., 2013).

While Amnesty International is accurate in their reporting, the volume of reporting is another issue altogether. Amnesty International carefully vets all information it releases, but how does it decide how much information to release on certain countries? Ron et al. (2005) asked this question as well. They found several factors, outside of human rights practices, that impacted the volume of reporting per country. Factors that increase Amnesty International's reporting include previous reporting, media profile, United States military aid, and size of the military of the targeted state. They argue that Amnesty chooses these countries to raise its own media profile in the Global North and therefore attract more donors (Ron et al., 2005). In fact the country targeted the most of Amnesty?s press releases is the United States, despite the fact that they are not even close to the top ten human rights abusers (Ron et al., 2005).

The previously cited literature establishes that, while accurate, Amnesty International is still a strategic international actor (Cooley and Ron, 2002; Ron et al, 2005; Hill et al, 2013; Keck and Sikkink, 1998). While NGOs publicize human rights abuses, they can also control the advocacy agenda by vetting issues (Carpenter, 2014). Perhaps they use this same power make strategic decisions about human rights reporting. Many of the studies use quantitative techniques to assess factual accuracy or quantity of Amnesty International?s reporting. There is a gap in the literature when it comes to quantitative analysis of the actual content of those reports. This paper will use machine-learning techniques to fill this gap and assess the strategic choice of actual topics covered in Amnesty International's reports.

Theory

As an advocacy NGO, Amnesty International uses information politics through TANs to achieve their normative goal of better global human rights practices. In this approach, information flows up from local NGOs to the international NGO (i.e. AI), who then quickly vet and publicize the information (Keck and Sikkink 1998). Human rights NGOs like AI also work to influence third parties into punishing the targeted

state for human rights abuses (Murdie and Davis 2012). International NGOS like AI are especially suited for this mission based on their many contacts and centrality in the global human rights network (Carptenter, 2014).

In the sector of transnational action, NGOs find themselves in a quandary. On one hand, NGOs have a normative goal that they would like to achieve. For Amnesty International, that goal is the global respect for human rights. However, in order to achieve that goal, AI needs material resources and media attention (in order to engage in information politics). Therefore, NGOs must simultaneously satisfy their normative mission, their donors, maintain access to target states, and also continue to survive as an organization. Sometimes these preferences work in harmony; however there can be discrepancies. Hill et al (2013) call this "conflicting incentives." Not all human rights reporting brings equal attention to AI's mission. This paper will use the conflicting incentives theoretic framework used by Hill et al (2013) to explain the difference in AI reporting on violence by non-state actors.

AI releases information about human rights abuses in a variety of ways. First, AI releases annual country reports, in which they document the state of human rights practices in every country in the world. In addition to the regular, annual reports, AI releases background reports, press releases and urgent action reports. The latter three are usually in response to new information or circumstances that demands AI's attention. In order to create these reports, AI must rely on other sources: domestic NGOs and reporters on the ground. After the information is gathered, AI must then decide how much credibility to give the information. This is where AI faces its conflicting incentives. AI needs to vet the information to maintain their authority and creditability as an international actor. However, AI also needs material things in order to pursue their normative goals, giving AI the incentive to exaggerate claims of human rights abuses to gain attention and therefore donations/volunteers/political authority.

The biggest obstacle to human rights reporting is the unstable information environment. How can AI effective vet and release information in places where quality information is nearly impossible to find? AI has two choices: release unverified information to raise its own media profile or choose not to release any unverified information. Their choice to verify information relies on several factors identified by Hill et al. (2013). They found that the increased probability of AI exaggerating allegations of torture raise in response to media attention, a small number of domestic NGOs, and a small winning coalition to selectorate ratio (Hill et al. 2013).

While the conflicting incentives that drive probability of exaggerating, this is not the complete picture. AI also makes strategic decisions about what to write and how to write background reports and press releases. The conflict again rests with the need for a media presence and satisfied donors versus the normative goal.

When non-state actors are committing acts of violence, AI finds itself in an impossible situation. On the one hand, AI must report on violence. On the other hand, AI's influence is directed toward states, which have more authority to direct their agents to respect human rights. The tools AI has developed to address human rights abuses (naming and shaming, for example) were developed to put pressure on states. Furthermore, AI tends to encourage, if not deliberately, domestic human rights protest (Murdie and Bhasin, 2011). The TAN approach is also created to change state behavior, not non-state. All of these conditions combine to lead AI to change the way in which it discussion countries with anti-government protest, especially if that government if repressive. Therefore, we hypothesize:

H1: The presence of a violent anti-government protest will be a significant indicator of topic choice in Amnesty International reporting on human rights.

Non-state actors also have an advantage of being able to frame their own narrative Bob (2005) finds that non-state actors that present themselves as a victim are more likely to get NGO support abroad. Non-state groups protesting a repressive or authoritarian regime are the ideal candidates for sympathy. Hill et al (2014) argue the size of the wining coalition in relation to the selectorate. This makes sense with Bob's (2005) theory of rebel marketing. When groups are using violent protest to confront an authoritarian regime, they can market themselves a victim of the regime. Also, when citizens of states with bad human rights practices are ready for change, there can lead to an increase in violent protests (Murdie and Bhasin 2011).

H2: Amnesty International will use different topics when discussing the human rights of democracies versus autocracies.

Research Design

The goal of this research is to understand the strategic choices AI makes when reporting about human rights. Do they stray from certain topics based on the type of violence or country circumstances? Are they likely to report on democracies in the same manner as autocracies? In order to answer these questions, this project requires a two-stage methodological approach. The first stage requires the use of machine learning techniques to sort Amnesty International reports into topics based on the text of the reports. From there, we use logistic regression models to predict the likelihood of each country year reports falling under a specific topic. Each stage of the research design is described in more detail below.

Topic Models

For this project, we estimated topic models over the set of AI background reports from 2000-2017 for African countries. We chose Africa to be able to use the Social Conflict Analysis Database (hereafter SCAD). Africa also presents a large amount of variance in political institutions, human rights practices, the state of political dissent, and AI reporting. We used Latent Dirichlet Allocation in the "topicmodels"(Grun and Hornik, 2011) package in R. Latent Dirichlet Allocation (hereafter LDA) is a process that seeks to estimate a probabilistic model of topics in a corpus (or collection of documents) based on the frequency and distribution of words. Therefore, in LDA the outcome variable is a topic and the inputs are the words. The goal is to estimate the (unobserved) underlying structure of the corpus using the distribution of words in each document (the observed indicators). The LDA process then sorts documents into these estimated topics, again, based on probabilities. This process is similar to that of Bayesian hierarchical modeling (Blei, Ng, and Jordan, 2002). The levels here are words nested in documents, which are then nested in topics in a corpus.

We estimated a three-topic model, based on theoretic and empirical considerations. Empirically, the perplexity (a measure of fit for topic models) scores of the three-topic model was statistically insignificant from the two-topic model and better than any models with more topics. Theoretically, the topics returned from the three-topic models were more distinct and allow for more variance than the two-topic model. These topics, the frequent words, and our interpretation of the topics are included in the results section below.

Logistic Regression

After fitting the topic models, we then extracted the predicted probabilities of each document being sorted into each topic. From there, we created three new variables: "topic 1", "topic 2", and "topic 3". Theses variables were coded one if the document had a better than chance probability of being in that topic, and zero otherwise. These three dichotomous variables became our final dependent variables. We estimated three logistic regression models, one for each topic. In these logit models, we used several country level variables as well as controls. Our main independent variables of interests included the human rights practices, regime type, the type of protest activity, number of deaths caused by protest activity, and the presences of a violent protest group. We also controlled for the natural log of population.

$$Pr(Document = TopicX) = \beta_0 + \beta_1(PHYSINT) + \beta_2(polity2) + \beta_3(riot) + \beta_4(demo) + \beta_5(strike) + \beta_6(pro\_gov) + \beta_7(anti\_gov) + \beta_8(ndeath) + \beta_9(log\_pop) + \beta_{10}violence$$

With these logistic regression models, we can determine the validity of our hypotheses about the impact of country level circumstances on AI's human rights reporting. We have learned from Hill et al. (2013) that AI does not inflate allegations of torture, but that does not mean that they are not biased in their reporting. AI makes deliberate decisions on which topics to report. Our research design is adjusted accordingly to test this hypothesis.

Data and Variables

In order to examine the validity of our hypotheses we created a new dataset on the reporting of Amnesty International on human rights abuses in African countries during the time period from 2000-2017. This data set is the first data set using the summary of Amnesty International reports (not to be confused with country reports). On the one hand, these reports pin-point injustice, human rights abuses and maltreatment on the lowest level. On the other hand, these reports also share information about successful events and meetings fighting human rights abuses or achievements in promoting rights and justice. For now the data set entails a total of 401 report summaries. The period from 1990 to 1999 will be coded in the near future to generate more data. All of the summaries are available in Amnesty International's archives . Below we discuss how we coded the dataset and which criteria were used to create our variables.

Amnesty International Reports

In order to capture the reporting of Amnesty International on human rights matters in the world we identified all reports from 2000-2017 in Africa. Unlike country reports, background reports are only released on topics AI perceives as important. Every country will have one country report per year and considering similar countries and regions they will not vary largely. In contrast to that, AI background reports, our baseline measure, are released on pressing issues or events which AI deems important to highlight and inform the public about. These types of reports are expected to reflect AI's attitude towards the issues more drastically than an overall yearly report.

We created a dataset with the summaries of all the reports uniquely identifying a single country. Reports discussing multiple countries, countries outside of Africa or organizations, like the African Union, were excluded, because they do not match our theoretical interest. Using summaries instead of the whole reports came down to two benefits. First, some of the reports in AI's archive are up to 50 pages long containing a wide range of topics whereas others are only a few sentences reporting on an event or the

condition of a prisoner. Treating both types of documents equally, even though they are reports, would have proven futile. Especially, for machine learning techniques and automated classifiers documents with long background reports would have run a higher risk of being falsely classified because the algorithm would have picked up redundant information that does not reflect AI's stance on the topics. Second, the summary as opposed to the actual report is expected to highlight what AI deems important in the report and emphasizes their attitude towards the topic(s) at hand. These reports are the underlying basis for our empirical evaluation on the influence of protest groups on AI's reporting on human rights issues in a country. The following sections lay out our variables.

Dependent Variable

Our dependent variable is a dichotomous variable indicting the presence (1) of the document being in each topic, or the absence (0) directly derived from the LDA topic models used on our AI reports data set. The outcome variable predicts the likelihood of one of the reports reporting on one of the three topics we received from our topic model analysis. Bearing our hypotheses in mind, we are interested in the effect of a protest movement's presence on AI's reporting. Our research design lets us identify the three major topics underlying AI's human rights reports and, furthermore, establish a correlation between the presence of a protest movement and AI's style of reporting.

Independent Variables

Our independent variables are comprised of a variety of measures concerning protest movements. We use the SCAD data set to identify the presence of protest movements and their behavior. However, since SCAD is an events data set on a daily basis and our analysis is conducted on the country-year unit we had to collapse SCAD's information to a yearly analysis level. The origin of our analysis is SCAD?s variable "etype" indicating of which type a protest movement is. There are 10 different types in the original SCAD dataset: Organized Demonstration, Spontaneous Demonstration, Organized Violent Riot, Spontaneous Violent Riot, General Strike, Limited Strike, Pro-Government Violence (Repression), Anti-Government Violence, Extra-Government Violence, and Intra-Government Violence. We recoded and merged some of the above and collapsed them as binary variables on a country-year level if at least one protest of a certain type was present in a country-year.

The variable *demo* indicates the presence of a distinct, continuous and largely peaceful action directed toward members of a distinct "other" group or government authorities. This variable captures both organized and spontaneous demonstration. The difference lies within clear organization and leadership. However, we

do not suspect to see a different type of reporting based on this difference, but rather on the fact that such a movement was present or not.

The variable *riot* identifies the presence of a distinct, continuous and violent action directed toward members of a distinct "other" group or government authorities. The participants intend to cause physical injury and/or property damage. Again, SCAD's original variables distinguished between the presence or absence of a clear leadership. We do not see a need for this distinction and therefore combine both types of riot.

Whether members of an organization or union engaged in a total abandonment of workplaces in limited sectors, industries or public facilities is captures with our binary variable *strike*. Unlike SCAD we do not distinguish between the public and private sector. Strikes are actions to voice dissent. For our argument, it is not necessary to differentiate who the target of the strike was, but rather the fact that a strike was present or not.

The variables *pro_gov, anti_gov, extra_gov,* and *intra_gov* capture the direction of violence. *Pro_gov* is coded as a "1" for distinct violent events waged primarily by government authorities, or by groups acting in explicit support of government authority, targeting individual, or "collective individual" members of an alleged opposition group or movement. These types of events are always initiated by the government or pro-government actors. *Anti_gov* describes distinct violent events waged primarily by a non-state group against government authorities or symbols of government authorities. It is distinguished from riots by having a semi-permanent or permanent militant wing or organization.*Extra_gov* captures violence waged primarily by a non-state group targeting individual, or "collective individual" members of an alleged oppositional group or movement. Opposed to that, *intra_gov* captures violence between two armed factions associated with different elements within the government. This includes violence between two legally constituted armed units or between unofficial militias associated with particular governmental leaders.

As an additional variable of interest, we included *ndeath* in our data set. This variable captures the total number of deaths occurring within a country- year due to protest movements or government action related to dissent. The actual count of deaths will help us to determine the severity of protest movements and, ultimately, distinguish between AI's reporting human rights abuses in countries with more violent and less violent protest movements. To further distinguish between violent and nonviolent forms of protest we created the binary variable *violence*. It is coded as a "1" for a country-year with more than 10 protest-related deaths and otherwise "0". This measure will further elaborate on AI's reporting style not only determined by the severity of protest, but also by the primary form of violent or nonviolent conflict.

Control variables

This study controls for three factors: population, regime type and physical integrity rights. The measure for the total national population is derived from the UNESCO data set. We are using a logarithmic version of the variable, because we expect the effect to gradually flatten with rising numbers of population. Population overall is expected to influence AI's reporting in two ways. First, the larger the population the higher the tendency for human rights abuses. Second, the larger the population the more likely AI is to pick up on abuses and report on them. Therefore, we hypothesize that with an increase in population AI will report more frequently on countries and therefore might over represent reports on a country for a certain category.

Our second control variable, regime type, is derived from the Polity IV data set. Autocracies are widely recognized to be the worst perpetrators of human rights abuses (Davenport 2007; Mason 2004). However, autocracies have to be distinguished from one another. Rather than treating them like equal entities, autocracies have different features that may or may not influence the way they utilize repression against their population (Geddes et al. 2014). The underlying logic of why and how political leaders allocate their resources and use different methods to maintain their power is linked to their desire of political survival (Bueno de Mesquita et al. 2004). Therefore, we hypothesize that a lower Polity score is connected to harsher critique and more human rights abuses.

The last control variable for our study is human rights abuses captured by Cingranelli and Richards Physical Integrity Rights Index (CIRI). Higher CIRI scores are an indicator of better human rights practices in a country, whereas lower CIRI scores indicate lower human rights practices. Even though this measure is to some degree endogenous to our dissent variables, it is crucial to also control for the overall effect of human rights abuses in a country beyond protest movements and public dissent.

Results

Topic Models

The first step in our empirical analysis was to identify the underlying topics in the summaries of AI's reports. Using the perplexity measure for the LDA algorithm we were able to identify two and three topics with the highest perplexity. Given the indifference between both scores and our theoretical endeavor to distinguish between as many topics as possible, we decided to settle for three methods. It is important to note, that the perplexity got significantly worse for adding more topics. The three topics we were able to

identify are as follows:

Table 1: Results of Topic Models

Words Associated with Topics		
Topic 1	Topic 2	Topic 3
right	right	government
Sudan	government	force
protect	women	Darfur
government	violate	detention
commit	recommend	torture
justice	elect	secure
violate	Sudan	include
crime	international	attack
Darfur	use	Sierra
force	Court	right
civilian	Nigeria	trial
impunity	peace	concern
letter	organization	end
concern	call	nation
arm	Leon	victim
year	kill	arm
state	arrest	civilian
prison	law	crime
abuse	republic	Liberia
international	concern	recommend
also	countries	ensure
violence	violence	express
make	opposition	displaced
secure	abuse	special
law	arm	people

All three topic have a common basis of recurrent words. These include "right[s]", "arm[s]" and "government". This is to no surprise, since those words pick up exactly what the reports are about: human

rights abuses. However, we did intentionally not exclude those words because we had not theoretical reason to do so. For instance, some of the reports did not deal with rights and the government in detail. Therefore, the algorithm was not necessarily to be expected to pick up on those. Looking at the results it becomes clear, however, that this was the case and the major topics are indeed governments and rights. What is of more interest, are the words that are unique to the topics and thus define it.

The first topic, which we labeled "Government abuse of human rights", entails the words "commit", "justice", "violate", "crime", "force", "civilian", "impunity", "concern", "prison", "abuse", "violence", "secure" and "law". As we can deduct from the words and patterns in this topic, there was a large emphasis on the violation of legal rights, justice, crimes as well as security. Especially, the words "abuse" and "prison" lead to the conclusion, that this topic deals primarily with human rights violations of prisoners and the suppression of justice and all government abuses of human rights.

The second topic, "Public Dissent", is made up of the words: "women", "violate", "recommend", "elect[ion]", "Court", "peace", "organization", "kill", "arrest", "law", "concern", "peace", "violence", and "opposition". In this topic, we can identify the targets of abuses. "Women" and "opposition" are clearly in the center. Furthermore, "election", "arrest", and "kill" lead to the conclusion that the matter of this topic is violent dissent against the regime. The opposition during elections is either killed or arrested. The word "peace" strikes out in this row of words. It might be a stemmed form of the original word "peace[ful]" which describes the modus operandi of the opposition movement and the civilian protestors.

We define the last topic as "Post-conflict struggles". This topic is primarily concerned with "force", "secure", "Sierra[-Leone]", "Liberia", "end", "trial", and "attacks". Besides these indicators of violence and human rights abuses, we find the words "victim" and "civilian" as well as "displaced" and "people". From the connotation of these words we derive at the conclusion that civilians are struggling to rebuild society after conflict. The word "nations" gives another hint. This can be traced back to the full phrase "United Nations" which often gets involved in post-conflict rebuilding. Also, the inclusion of Sierra Leone and Liberia, countries recovering from civil war in the is evidence for this topic dealing with post-conflict countries and societies.

Logistic Regression

Identifying the three underlying topics was the first step. We used logistic regression to test those topics on our hypotheses. It is important to note that every single country-year we registered a report in, the shamed country experienced a violent form of protest. Not even one country-year in our data set was without a violent protest. This is no surprise though. Most of the countries in our data set are highly

autocratic and known for their repression as well as human rights abuses. That is why AI is reporting on them in the first place. This inherent selection bias is at no time a limitation of this study, since we are interested in predicting different styles of reporting, not the absence of reporting. Our findings indicate that AI uses three distinct topics to report on human rights abuses in those countries, as shown in Table 2.

Model 1 describes the likelihood of an AI report on a specific country using topic 1, "Government abuse of human rights". Reports on states with poor physical integrity rights practices are more likely to be in this category, obviously. Countries being criticized in this department are also likely to experience anti-government violence with a, comparably, lower number of deaths. This makes sense because states with poor human rights practices facing anti-government protests are likely to respond with repression. However, the overall experience of violence increases the likelihood of a report being classified in topic 1. Interestingly enough, the more deaths experienced as a result of social movements per country year, the less likely the report will fall into this category. This makes sense, especially considering the second category of public dissent. More deaths can imply a revolution or rebellion of some kind, which would be more suited to talking about civil war rather than physical integrity rights abuses. Regime type is not a statistically significant indicator of topic 1.

Model 2 in contrast to that is the exact opposite of model 1. The more autocratic a country is the higher its likelihood to receive a reporting on topic 2, "Public Dissent". Especially, strikes and a large number of protest-related deaths increase the likelihood of AI reporting on topic 2 in that country. Interestingly, countries in topic 2 on average experience less anti-government violence. This can be explained by the statistical significance of strikes. The form of public dissent found in this topic is not necessarily related to the government itself or expressed towards it, but rather in violent riots, or a struggle for power between factions. Civilians in more autocratic regimes are powerless to openly challenge the regime. The fact, that the government is so much more repressive then leads to the unbalanced distribution of power. In times of unrest and protest the government cracks back harder on the civilian population, thus killing more of the peaceful opposition. The number of death is positively associated with the category of public dissent, indicating the possible presence of outright rebellion against the government or violence between competing factions during civil war. Also, in this we see that states with better human rights practices are more likely to be in this category. So when dissidents have opportunities to express dissent, they take advantage of them, which makes that state more likely to be in the reporting category of "Public Dissent".

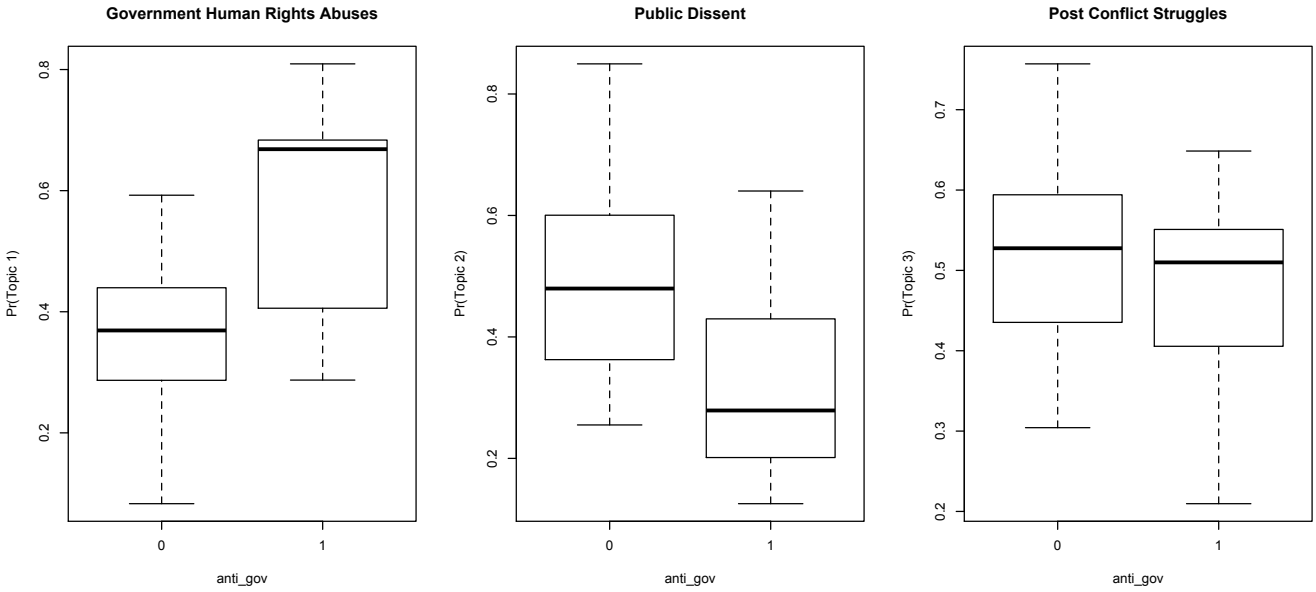
The last model, "Post-Conflict Struggles", is very unique compared to the other two in the form that only strike among the forms of protest is statistically significant. Overall, we see very little evidence for human rights abuses in this topic. However, and this is important, the polity score is significant and

Table 2: Results from Logistic Regression

	Dependent variable:		
	topic1	topic2	topic3
	(1)	(2)	(3)
PHYSINT	-0.158^{*}	0.141^{*}	0.091
	(0.088)	(0.085)	(0.083)
polity2	0.020	-0.058^{*}	-0.071^{**}
	(0.033)	(0.034)	(0.031)
riot	0.334	−0.305	−0.119
	(0.320)	(0.317)	(0.307)
demonstration	−0.592	0.462	0.318
	(0.406)	(0.409)	(0.394)
strike	−0.236	0.681^{**}	0.530^{*}
	(0.298)	(0.303)	(0.289)
pro_gov	0.066	0.202	0.104
	(0.320)	(0.329)	(0.307)
anti_gov	1.020^{***}	-1.056^{***}	−0.364
	(0.297)	(0.313)	(0.283)
extra_gov	0.288	−0.335	−0.148
	(0.315)	(0.322)	(0.308)
intra_gov	0.449	0.165	0.643
	(0.565)	(0.560)	(0.520)
ndeath	-0.039^{***}	0.037^{***}	−0.003
	(0.012)	(0.012)	(0.011)
log(population)	0.034	−0.098	0.262
	(0.183)	(0.186)	(0.177)
violence	0.784^{**}	−0.652	−0.014
	(0.391)	(0.401)	(0.378)
Constant	−0.186	0.380	-4.937^{*}
	(2.955)	(2.976)	(2.860)
Observations	334	334	334
Log Likelihood	−209.573	−205.409	−223.654
Akaike Inf. Crit.	445.145	436.817	473.308
Note:			*p<0.1; **p<0.05; ***p<0.01

negative. This means that countries with higher polity scores (and therefore more democratic) are less likely to be in this category. This makes sense because states that are doing well post-conflict are more likely to make successful democratic transitions, and therefore do not need to be reported on by AI.

Figure 1: Anti-government Protests and Topics



We find significant support for our first hypotheses on AI?s reporting on countries with anti-government protest movements. In countries with such movements, AI specifically criticizes the ongoing violence (especially by the government). Reports that appear in the "public dissent" category are less likely to be about countries with anti-government activity. This finding is very crucial in pointing out the importance of AI background reports. Our models confirm that AI is indeed reporting correctly on the issues at hand and that there is indeed a difference in reporting on countries with and without anti-government protest movements. AI prefers to talk about anti-government activity in the context of human rights violations (usually repression), while the topic of public dissent is less likely to discuss anti-government protest. Models 2 and 3 partially support our second hypothesis. Overall, we do not find a statistically significant effect of different types of protest on document topic, with the exception of strikes. We do find, that AI indeed reports differently on more autocratic regimes as compared to more democratic regimes. While our data are limited to Africa, we do have enough variance in polity scores (the range is -9 to 9). From this sample, we see that reports on more democratic states are less likely to fall under topic 2 (public dissent) or topic 3 (post conflict struffles). Democracies in general have a higher respect for physical integrity rights, however, the polity variable is not significant in topic1 (human rights abuses by governments). This points to support

Figure 2: Physical Integrity Rights Respect and Topics

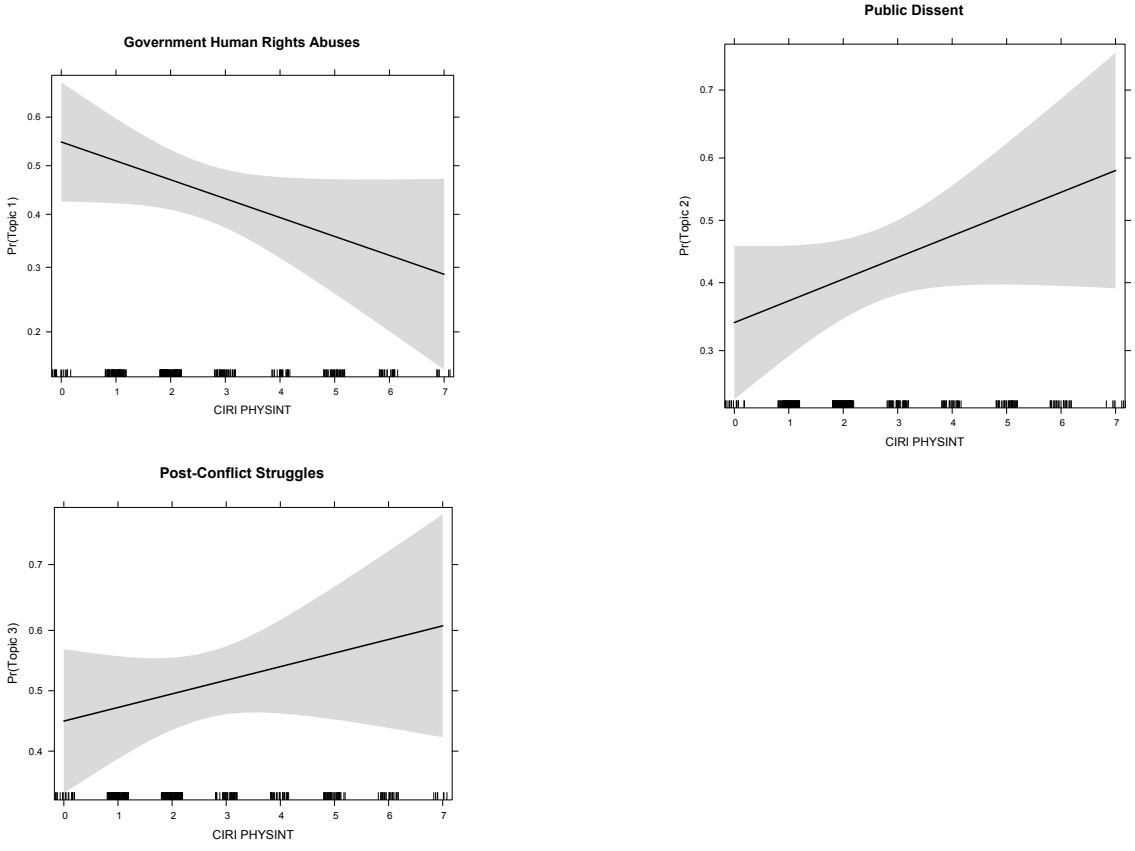
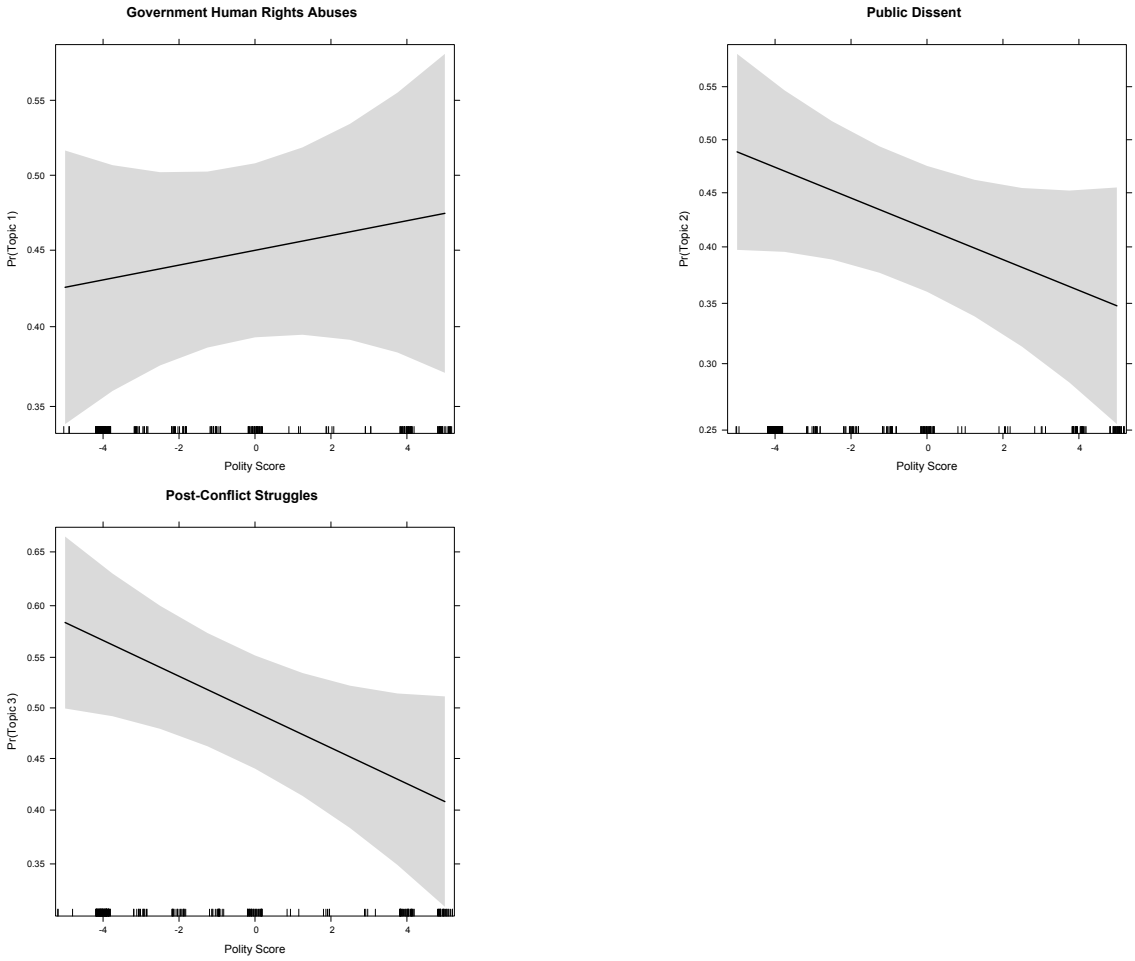


Figure 3: Polity IV Score and Topics



that regime type does matter in AI's reporting. Further data points and observations will allow us in the future to elaborate on this hypothesis.

Conclusion

This preliminary draft of our research finds that there is indeed a difference of AI reporting on countries with and without anti-government protest movements and further that AI distinguishes between more and less autocratic regimes. Using machine learning techniques like topic models and the LDA we were able to identify the different topics dictating AI's background reports. Although we were able to provide preliminary findings this draft will need improvements in several ways which will be briefly pointed out here.

First, our data set thus far consists of 401 documents. About 70 of those were excluded in our predictions since we lacked information, such as control variables, to back them up. Having more data for

this endeavor is crucial. One the one hand, this can be achieved by coding the remaining ten years from 1990-1999 to increase the amount of observations and, on top of that, have a complete sample of all AI background reports. On the other hand, better and more precise control variables are necessary. Many of the countries we are investigating are either too small to have their population, polity scores or CIRI scores reported or there is no institution to receive those data from, as for example in Somalia. Several statistical methods might be of use to decrease the bias in our data due to missing data points.

Second, we will have to include a control variable for the number of NGOs present in a country-year. This variable could influence drastically the outcome of the reporting. Since countries without an NGO or only a few are very likely to lack sufficient reports on abuses and will might therefore lead to a bias in reporting. We expect this measure to be even more valuable than counting the number of reports, since we theorize that the bias does not lie within the amount of reports, but rather within the number of NGOs able to report.

Third, we initially examined the difference in various time spans, however, we did not find any major differences in our results. This was mainly driven by the accusation of the NGO Watchdog towards Amnesty on biased reporting in 2006. We controlled for this effect by splitting our data in two time periods 2000-2006 and 2007-2017. We did not find any bias in the form of different results in terms of our topics. In fact, our results were consistent over both time spans. Although this makes us more confident in our preliminary findings, there might well be a difference in reporting before 2000 we did not capture in our data set so far. Increasing the time span of our data set in the future will hopefully shed more light on this issue.

Fourth, for now the use of Latent Dirichlet Allocation (LDA) performed well and provided us with robust results. However, we are aware of some of the shortcomings of this algorithm and will explore different machine learning methods to identify AI?s style of reporting and thus increase the precision of our findings.

Fifth, for now most of the different types of protest did not have a significant effect on the style of reporting. Only anti-government violence and strike yielded results. Given the fact that every single country-year in our data set experienced at least one form of violent protest this is surprising to us, but it might indicate that some forms are more important for media attention than others. We will further investigate how media attention and NGO reporting reflects on these types of protest.

There is no doubt that there is much more work to be done on this topic that is outside the scope of this paper. The previous paragraphs give some direction for what we would like to see in the future of the topic. However, we do believe that this paper provides a foundation upon which many scholars can build. Not only do we provide a methodological diverse and innovative tool to study human rights reporting, but also we uncover a potential source of bias in NGO human rights reporting. This does not mean that AI

reports are incorrect: the information contained in AI's reports is factually correct and rigorously verified. However, AI remains a strategic actor and can make a decision on what to emphasize. Future studies should keep this in mind when exploring the mixed motives of NGOs that conduct human rights, or indeed any issue, advocacy. We hope our work sheds some light onto the complex preference formation of advocacy NGOs.

References

[1] Barry, C.M., Clay, K.C., and Flynn, M.E. (2012). "Avoiding the Spotlight: Human Rights Shaming and Foreign Direct Investment." *International Studies Quarterly*. 57(3): 532-544.

[2] Blei, D.M., Ng, A. Y., and Jordan, M.I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(1); 993-1022.

[3] Bloodgood, E. A. (2011). "The Interest Group Analogy: International Non-Governmental Advocacy Organizations and International Politics." *Review of International Studies* 37(1): 93-127.

[4] Bob, C. (2005). *Marketing Rebellion*. Cambridge: Cambridge University Press.

[5] Bueno de Mesquita, B. et al. (2004). *The Logic of Political Survival*. Cambridge: MIT Press.

[6] Carey, S. (2010). "The Use of Repression as a Response to Domestic Dissent." *Political Studies* 58: 167-186.

[7] Chenoweth, E. and Lewis, O. (2013). "Unpacking Nonviolent Campaigns Introducing the Navco 2.0 Dataset." *Journal of Peace Research* 50 (3): 415-23.

[8] Chenoweth, E. and Stephan, M. (2011). *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict*. New York: Columbia University Press.

[9] Cingranelli, D./Richards, D. (2010). "The Cingranelli and Richards (CIRI) Human Rights Data Project." *Human Right Quarterly* 32 (2): 401-24.

[10] Cingranelli, D. L., Richards, D. L., and Clay, K. C. (2014). "The CIRI Human Rights Dataset." http://www.humanrightsdata.com.

[11] Carpenter, C. (2014). *Lost Causes: Agenda Vetting in Global Issue Networks and the Shaping of Human Security*. Ithaca, N.Y.: Cornell University Press.

[12] Conrad, C. R. (2014). "Divergent Incentives for Dictators: Domestic Institutions and (International Promises Not to) Torture." *Journal of Conflict Resolution* 58 (1): 34-67.

[13] Cooley, A. and Ron, J. (2002). "The NGO Scramble: Organizational Insecurity and the Political Economy of Transnational Action." *International Security* 27(1): 5-39.

[14] Davenport, C. (2007). "State Repression and Political Order." *Annual Review of Political Science* 10 (1): 1-23.

[15] Edelman, R. (2003). "Building Trust: A Special Report." Available online: (https://www.edelman.com/assets/uploads/2014/01/2003-Trust-Barometer-Global-Results.pdf) 11 April 2017.

[16] Feaver, P. (2003). *Armed Servants: Agency, Oversight, and Civil-Military Relations*. Cambridge, MA: Harvard University Press.

[17] Gandhi, J. 2008. *Political Institutions under Dictatorship*. Cambridge:Cambridge University Press .

[18] Gaub, F. (2013). "The Libyan Armed Forces between Coup-Proofing and Repression." *Journal of Strategic Studies* 36 (2): 221-44.

[19] Geddes, B. et al. (2014). "Global Political Regimes Data Set." http://dictators.la.psu.edu/.

[20] Grun, B. and Hornik, K. (2011). "**topicmodels**: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13): 1-30.

[21] Hill, D. W., Moore, W. H. and Mukherjee, B. (2013). "Information Politics Versus Organizational Incentive: When Are Amnesty International's Naming and Shaming Reports Biased?" *International Studies Quarterly* 57(2): 219-232.

[22] Keck, M. E., and Sikkink, K. (1998). *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

[23] Mason, T. D. (2004). *Caught in the Crossfire: Revolutions, Repression, and the Rational Peasant.* Oxford: Roman and Littlefield Publishers.

[24] Murdie, A. and Davis, D. R. (2012). "Shaming and Blaming: Using Events Data to Assess the Impact of Human Rights INGOs." *International Studies Quarterly* 56(1): 1-16.

[25] Murdie, A. and Peksen, D. (2013). "The Impact of Human Rights INGO Activities on Economic Sanctions." *Review of International Organizations* 8(1): 33-53.

[26] Murdie, A. and Peksen, D. (2014). "The Impact of Human Rights INGO Shaming on Humanitarian Interventions." *Journal of Politics* 76(1): 215-228.

[27] Nepstad, S. (2011). *Nonviolent Revolutions: Civil Resistance in the Late Twentieth Century.*New York: Oxford University Press.

[28] Poe, S.C., Carey, S.C., and Vazquez, T.C. (2001). "How are These Pictures Different? A Quantitative Comparison of the US State Department and Amnesty International Human Rights Reports, 1976-1995." *Human Rights Quarterly* 23(3): 650-677.

[29] Risse, T. and Sikkink, K. (1999). "The Socialization of International Human Rights Norms into Domestic Practices: Introduction" in Risse, T., Ropp, S. C., and Sikkink, K. (Eds). *The Power of Human Rights: International Norms and Domestic Change*. Cambridge, UK: Cambridge University Press.

[30] Ritter, E. H. and Conrad C.R. (2016). "Preventing and Responding to Dissent: The Observational Challenges of Explaining Strategic Repression." *American Political Science Review* 110 (1): 85-99.

[31] Ron, J., Ramos, H., and Rodgers, K. (2005). "Transnational Information Politics: NGO Human Rights Reporting, 1986-2000." *International Studies Quarterly* 49(3): 557-588.

[32] Schock, K. (2004). *Unarmed Insurrections: People Power Movements in Nondemocracies*. Minneapolis: University of Minnesota Press.

[33] Sell, S.K. and Prakash, A. (2004). "Using Ideas Strategically: The Contest Between Business and NGO Networks in Intellectual Property Rights." International Studies Quarterly 48(1): 143-175.

[34] Union of International Organizations. (2015). *The Yearbook of International Organizations*. Accessed Online: $https://www.uia.org/sites/uia.org/files/misc_pdfs/Types_of_organization.pdf$

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

WORD COUNT: 7656