How to Teach Machines to Read Human Rights Reports and Identify Judgments at Scale

Abstract

Human rights organizations communicate their judgments about what countries are violating specific rights over time. The last two decades have seen the rate and scale of these communications increase exponentially; rendering careful human reading, across all potential rights, extremely difficult. The recent digital availability of these texts opens up the possibility of leveraging innovations in machine learning and natural language processing to build systems that can measure and extract systematic judgments from these texts. We introduce PULSAR, a system we have built and tuned to parse human rights reports. The resulting features of the text are sensitive to word order, negation and syntactic relations between words. Using human coded scores on physical integrity rights and women's political rights as targets, we illustrate that PULSAR contributes to more accurate and interpretable automated human rights measurement systems, as compared to contemporary practice.

Bios

Michael Colaresi is William S. Dietrich II Professor of Political Science, University of Pittsburgh. He publishes on topics related to computational social science, changes in human rights concepts over time, high resolution conflict prediction systems, national security secrecy in democracy, and Bayesian inference. His most recent book is Democracy Declassified: The Secrecy Dilemma in National Security (2014).

Baekkwan Park is a postdoctoral research associate at the University of Pittsburgh. His research interests include human rights and political violence.

Kevin Greene is a PhD student at the University of Pittsburgh. His research interests include political violence, human rights and computational social science.

1 Introduction

Many human rights organizations (HROs) around the world dedicate themselves to promoting respect for human rights and abolishing human rights abuses.[1] These organizations marshall human and financial resources to monitor the actions of governments and pressure them to act in accordance with human rights principles. Much of this pressure is exerted indirectly by maintaining extensive websites that publicize and document violations, while often calling for remedial action. This public communication is crucial, human rights organizations are most effective when their calls for reform are backed by strong public advocacy. In fact, because of the strong relationship between HROs and public awareness, a number of human rights scholars have paid attention to HROs and their activities in order to better understand advocacy campaigns as a mechanism for change (Keck & Sikkink 1999, Risse-Kappen et al. 1999, Ron et al. 2005, Lebovic & Voeten 2006, Hafner-Burton 2008, Meernik et al. 2012, Murdie & Davis 2012, Stroup 2012, Hendrix & Wong 2013, Murdie 2014, Kahn-Nisser 2018, Koliev & Lebovic 2018).

However, despite a growing cascade of available information from HROs, only limited signals from this rich information currently informs systematic human rights research directly. Because of its growing scale and linguistic complexity, almost all the research in this area ignores the majority of the content in natural language texts and utilize only simple lower-dimensional observations such as counting and aggregating the number of reports issued by HROs (Ron et al. 2005, Lebovic & Voeten 2006, Meernik et al. 2012, Hendrix & Wong 2013, Koliev & Lebovic 2018) or uses word counts ignoring syntax and word order in the texts (Bagozzi & Berliner 2016, Greene et al. 2018). While these simple approaches may be useful and appropriate for some applications such as event data analysis or topic modeling; they may be less suited to more nuanced tasks, for example identifying opinions and judgments on specific issues or aspects of human rights.[2]

This is not simply a limitation within the human rights literature. The computational analysis of text at scale necessarily involves the process of converting the unstructured text into computer-readable numerical vectors. A commonly used transformation to represent the text of documents as a vector of counts of individual words (or n-grams) is known as the bag of words (BOW) rep-

resentation. Since BOW is simple and intuitive, it makes the quantification of text accessible and has worked surprisingly well for some tasks such as topic modeling (Blei et al. 2003, Quinn et al. 2010, Roberts et al. 2014, Bagozzi & Berliner 2016, Kim 2017). However, it is important for researchers to realize that BOW effectively scrambles other important signals and context from text, rendering them uninterpretable because the transformation ignores word order within a document and because it assumes away grammatical rules or syntactic relations between terms. In particular, as experience with sentiment and opinion-extraction shows, predicting judgments can be significantly improved by moving beyond bag of word representations and using grammatical and syntactic information that is informed by word order in a document (Liu 2015, Lin et al. 2016).

Context and syntax is particularly important for those working with human rights texts. The coding of the presence or absence (and potentially the intensity) of violations and protections of specific rights includes a very different semantic structure compared to the task of topic modeling documents. Most critically, the goal of topic modeling is to extract probability distributions over words (Blei et al. 2003). While each token might be more or less informative for a given topic, all words are assumed to play the same role – they communicate whether or not a topic is being discussed in a document, no other role exists. One can think about this as a word being described by a singleton ontology, where there exists only one possible type of information that a word can convey (topical information).

In contrast, when attempting to extract opinions and judgments from texts, tokens necessarily play distinct but related roles. Some words express what right is being judged while other words communicate the valence of the judgment and still others make clear the connection between the rights and the valence. For example, in the sentence, "*The government systematically abused minorities*", the word "*systematically*" communicates the scale of a judgment, while the terms "*abused*" and "*minorities*" identify that a violation of minority rights is being discussed. There are other relevant roles also, including terms that identify perpetrators of human rights abuses, such as the "*government*" in the previous example. Therefore, to conceptualize judgments-on-rights, we need an ontology of concepts that includes at least these two categories, a) words can convey

judgment and b) terms can communicate the rights that are being judged.[3]

To capture the relevant judgment/opinion expressed in complex texts, like human rights reports at scale; we need systems that can transform natural language into systematic features without ignoring and scrambling the very targets of inference these systems were designed to inform. In this paper, we introduce PULSAR (Parsing Unstructured Language into Sentiment-on-Aspect Representations), a tool that processes large-scale natural language corpora into structured aspect-based sentiment expressions. PULSAR was created specifically for human rights texts, although it can and has been deployed in other domains.

Our work builds upon aspect-based sentiment analysis (ABSA) tasks in the natural language processing (NLP) community. In this vein of successful research, automated systems have been created that can, at their most basic, a) identify the aspect being judged (e.g. an issue, right or product), b) the sentiment/judgment expressed on that aspect (e.g. positive/negative or pro/con, or intensity), and c) link those pairs together. PULSAR takes natural language text as input and uses a series of grammatical and syntactic rules to group terms together, loosely identifying judgments as verb-phrases and aspects as noun-phrases. The output is a series of new tokens, which can be thought of as multi-word expressions (MWE) with a defined role. The simplest ontology of semantic roles we begin with are aspects or judgments. These tokens can then be counted and used as features in supervised and unsupervised learning tasks. Thus, PULSAR transforms documents into bags of aspect-sentiment pairs, in contrast to BOW.

The US State Department reports on Human Rights from 1993 to 2016 provide an important test case to explore the usefulness of PULSAR and syntactic and grammatical information.[4] These reports represent an important source of allegations that have been used to code a variety of the scalar measures of state human rights practices. Additionally, as the reports have grown in length over time they can not be read in their entirety by one researcher or even a small team of researchers in a reasonable amount of time. Thus, we can use human coded data on the severity of human rights abuses in a country in a given year, to test whether PULSAR can extract more meaningful patterns from our textual corpora and help social scientists to move beyond BOW feature representations

of text. Specifically we compare the accuracy and interpretability across algorithms that use BOW versus PULSAR-generated features to predict human-coded scores on two different issues, physical integrity rights (as coded in the Political Terror Scale) and women's political rights (as coded by the CIRI project). We find that, as expected, PULSAR-generated features not only produce modest increases in accuracy, but also allows researchers to identify terms and phrases that are semantically related to the rights being judged and valence of those judgments, as compared to necessarily more vague BOW-generated features. This latter attribute of PULSAR has the potential to connect qualitative and quantitative analyses of human rights texts, as these features can serve as pointers for further human exploration of patterns in the texts.

2 Machine Learning from Human Rights Texts

HROs collect and publish information about human rights conditions and abuses by either implicitly or explicitly judging political actions relative to national and international norms. Descriptions of these judgments are often collected into annual reports that are publicly released, as well as press releases, blog posts, and advocacy messages. The scale of digitally-available text from HROs has been growing significantly over the last decade (Murdie 2014). Other research has made clear that there are benefits to using computational systems to analyze patterns across these texts (Fariss 2014, Bagozzi & Berliner 2016, Greene et al. 2018).

However, we argue that, to date, current computational systems do not attempt to extract crucial information in these human rights documents: the expressed judgments and what aspect of human rights is being judged. Complementing work on topic modeling and text-based prediction systems, we suggest that researchers view human rights reports as communicating judgments (such as identifying systematic violations) of the behavior of an entity (such as Saudi Arabia) on a specific right (such as freedom from the fear of torture).[5] Judgments can take the form of presence or absence of violations/protections (such as stating there were reports of violations) and potentially express an intensity (such as communicating that violations were widespread).

2.1 Previous Work Extracting Systematic Information From Human Rights Text

The human rights research community has a strong history of leveraging human rights texts. However, there is an important lacuna. On the one side, while many influential teams of researchers have read texts to measure judgments/opinion, few automated systems attempt to extract similar information. On the other side, the automated systems that have been build to leverage the mass of information in human rights texts do not attempt to measure judgment/opinion directly. Despite the vast quantity of information in HRO releases, human rights scholars have not fully leveraged the rich information on the expressed judgments on specific rights contained in these human rights documents.

Some teams, following a long tradition of qualitative research in political science, have manually coded judgments/opinions in background reports and press releases found in the *Amnesty International Cumulative Guide 1962-2000* (Ron et al. 2005), *Urgent Action (UA)* (Hendrix & Wong 2013, Meernik et al. 2012) or annual session reports from *the United Nations Commission on Human Rights (UNCHR)* (Lebovic & Voeten 2006) or the session reports from *the UN Committee Against Torture* (Kahn-Nisser 2018). Influential data collection efforts by the PTS (Gibney et al. 2015) and CIRI (Cingranelli et al. 2014) teams can also be seen as efforts to extract general judgments of countries on specific aspects over time. Yet, qualitative human reading does not scale to the accelerating mass of information available currently and suffers from potential complaints about reliability (King & Lowe 2003).

There are a few notable exceptions to the absence of large-scale automated use of text in human rights research. On the basis of 432 HROs listed in the *Yearbook of International Organizations*, Murdie & Davis (2012) created a dataset of human rights events in *Reuters Global News* reports concerning HROs from 1992-2007. This innovative work built on an event data coding ontology, measuring who did what to whom, when (Schrodt et al. 2014), but not what judgment was expressed on government actions. Bagozzi & Berliner (2016) analyze the *U.S. State Department Human Rights Reports* using a structural topic model to understand the changes of underlying themes

in the reports. As noted earlier, however, topic coding attempts to extract probability distributions over words, not the valence of terms (Blei et al. 2003, Quinn et al. 2010).

2.2 General Approaches to Extracting Judgments from Texts at Scale

Despite the absence of automated systems to extract human rights judgments from texts, currently, other domains have made progress on related problems. Across the social sciences, there have been influential efforts to capture sentiment from text. Dictionary-based sentiment analysis is the most common in the social sciences. In this approach, text is fed into a simple computer program which, in essence, collapses the high dimensional expressions in language into a vocabulary of, what is conventionally, just three signals ("positive", "neutral", and "negative"). Crucially for our research, this approach does not attempt to connect sentiment with the aspects or entities being judged. Functionally, every word in a vocabulary is matched to one of these three sets, representing one of three possible valences. Words in a given set are often given a corresponding value (i.e. 1, 0 ,-1, respectively). A specific sentence is then represented by the sum, or other aggregation function, of the scores for its constituent words. Note that dictionary-based sentiment analysis relies on an implicit BOW model, as it does not take word order, grammar or syntax into account. Two improvements have been suggested to dictionary-based sentiment analysis, using supervised learning approaches to map words to sentiment categories (Pang et al. 2002) and using rule-based classifiers to carry out aspect-based sentiment analysis. We return to these innovations below.

Perhaps less obviously, word scaling methods (Slapin & Proksch 2008, Laver et al. 2003, Monroe & Maeda 2004, Lowe 2013), several of which were explicitly inspired by ideal point models, can be cast as models that are estimating relative judgments from texts. Scaling methods effectively assume that sentiments/opinions are expressed as words (usually unigrams) regardless of the aspects and entities being discussed in a given document. The innovation of word scaling methods is that they do not assume that the translation of words into positions is known a priori. Instead, identifying restrictions are used to learn the set of vectors, also known as loadings, that transform the latent positions (e.g. how negative or positive, left or right, liberal or illiberal) of actors into

observed behaviors, represented as counts of term frequencies.

It is important to emphasize that these scaling methods also rely on an explicit BOW assumption to compress language to a smaller fixed dimension. Further, the BOW assumption has a particular and perverse consequence for word scaling models. Because words are assumed to transmit the same information across documents, there is no analog to bills or questions in ideal point and item response models that inspired these approaches. Technically, this is an identification problem that can be solved by assuming that item/bill parameters do not vary. Conceptually, this constraint directly implies that comparisons are being communicated across the same dimension, regardless of the words being expressed. Monroe & Maeda (2004) discuss this in more detail but for our purposes, this assumption is akin to assuming that the same human right is being judged across all words.

Thus, the state of the art approaches to extracting judgments from texts, across the social sciences, either fall back on human-reading or assume away different rights that could be judged.[6] We argue below that these limitations largely spring from the constraints imposed by BOW representations of text. Using these existing approaches within a BOW framework would not allow us to extract the judgments-on-rights that we know are densely expressed in human rights texts.

3 Potential Problems with BOW Representations for Judgment Extraction

BOW representations necessarily remove the syntactic, and thus semantic, dependencies between words across a document. For example, take two sentences with very different meanings in the context of a human rights document, Sentence 1: *"The government agents attacked the protesters."* and Sentence 2: *"The protesters attacked the government agents."*. These sentences are identical in terms of word counts, as calculated with unigram BOW statistics, presented in Table 1. But, of course, they have very different semantic meanings when read, particularly when attempting to judge the behavior or the government. Word order, the grammatical role a word plays, and syntax

impact our ability to decode the overall sentiment/judgment expressed.

Negation, and in particular, what is being negated, also influences the quality of sentiment extraction on specific rights/aspects. A BOW model only counts negation terms, it cannot represent what term is being negated; since this is function of the syntax of the sentence. Thus, Sentence 3: *"The government protected human rights abusers, not victims."* and Sentence 4: *"The government protected victims, not human rights abusers."* again have identical BOW vectors, but the meanings are completely different. Table 1 makes clear that BOW vectors can differentiate distinct topics (as the language across sentences 1/2 ("*agents, protesters, attacked*") and 3/4 overlaps little ("*abusers, victims, protected*"), but word counts aggregate over and ignore the nuance of judgments and what is being judged (which are quite different in 1 versus 2 and 3 versus 4). Notice that sentence 1 and 2, or 3 and 4, will receive the same dictionary-based sentiment score and word scaling, since these use the BOW representation.

	abusers	**agents**	**attacked**	**government**	**human**	**not**	**protected**	**protesters**	**rights**	**the**	**victims**
Sentence 1	0	1	1	1	0	0	0	1	0	2	0
Sentence 2	0	1	1	1	0	0	0	1	0	2	0
Sentence 3	1	0	0	1	1	1	1	0	1	1	1
Sentence 4	1	0	0	1	1	1	1	0	1	1	1

Table 1: BOW Representation with Word Count Vector for the Sentences. Sentence 1: *"The government agents attacked the protesters."* and Sentence 2: *"The protesters attacked the government agents."* have identical BOW vectors (along their respective rows) but very different semantic judgments of government behavior. Likewise, Sentence 3: *"The government protected human rights abusers, not victims."* and Sentence 4: *"The government protected victims, not human rights abusers."* illustrate a similar pattern with negation.

A growing literature on supervised [7] and rule-based learning[8] approaches to sentiment analysis have also reached the conclusion that word context and syntax are crucial to understanding the valence of sentences. In their seminal work on sentiment analysis of movie reviews, Pang et al. (2002) introduced the task of capturing sentiment by using a BOW representation to engineer features in a supervised learning approach. Many have followed up on this work, adding non-BOW features such as negation, intensifiers, grammar parsing and syntactic dependencies to substantially improve the accuracy and usefulness of these more recent systems (Pang & Lee 2008, Liu 2015,

Feldman 2013).[9]

4 Using PULSAR for Aspect-Based Sentiment Detection

Instead of detecting the overall polarity of a sentence or a document, ignoring the target entities and their aspect, aspect-based sentiment analysis (ABSA) is an approach that is interested in identifying the aspects/facets of given target entities and estimating the sentiment polarity for each mentioned aspect (Liu 2015). ABSA specifically introduces an ontology of differential semantic roles for words to play. Thus, instead of just counting words, we need to match words/phrases to their relevant semantic functions. BOW representations miss the differential meanings within the pair of sentences above because they, by design, do not recognize that terms have distinct functions for sentiment extraction. In sentences 3 and 4 above, the negation of the targets of the violation ("*protected human rights abusers*, *not victims*") is very different than the negation of a protection ("*protected victims, not human rights abusers*"). ABSA attempts to detect whether words are communicating aspects being judged, or the sentiment/judgment on those aspects (or something else).

By focusing on the aspect-sentiment representations of text introduced in ABSA, we recover the potential to systematically extract what people are judging and talking about, simultaneously. Yet, this means moving away from BOW representations. The question then becomes, how do we quantify text without using BOW assumptions?

Parsing Unstructured Language into Sentiment-Aspect Representations (PULSAR)

Towards answering this question, we have built a tool, PULSAR, that can extract/tag phrases as expressing either aspect or judgment functions, as well as link these phrases together when the judgment is offered on that identified aspect. Our approach is distinct from previous ABSA research in at least two respects. First, PULSAR is optimized for the language used in nuanced human rights

texts. Traditionally, ABSA research has focused on product and movie review corpora. However, human rights reports include language that is often indirect and includes the phrasing dealing with "reports" that are not present in conventional reviews. Second, we set out to improve the interpretability of automatic systems so qualitative and quantitative researchers interested in these issues can use PULSAR to learn more about the underlying process of human rights judgments. Rather than focus solely on improving the accuracy of our model, it is our hope that PULSAR facilitates the interpretation of what signals our models are learning and what they are missing (Colaresi & Mahmood 2017).

PULSAR utilizes, instead of discarding, the syntax in documents. We have developed predefined rules about grammar dependency relations between opinion-oriented words and aspects (Liu et al. 2015, Wu et al. 2009, Liu et al. 2013, Qiu et al. 2011) that we detail below. Recently several contributions have shown that syntactic rule-based methods easily outperform statistical methods in some benchmarks. Further, syntactic rule-based methods like ours have been shown to be effective with only small sets of labeled aspects (Liu et al. 2015, Poria et al. 2014).

Our system also uses a small set of domain-specific expressions to guide the parsing of sentences into aspects and judgments. In order for sentences to be parsed and tagged with their roles (including a null value) accurately and clearly, PULSAR includes domain-specific multi-word expressions (MWE) as input, but can also return frequent MWE expressions and what role they are used in (aspect or expression). A MWE is a sequence of words that acts as a single unit at some level of linguistic analysis (Calzolari et al. 2002). According to current linguistic theories, language users generate and understand sentences without necessarily fully breaking them into individual words, instead, units of multi-word function as the basic building blocks (Goldberg 1995, Kay & Fillmore 1999, Stefanowitsch & Gries 2003). Thus, the importance of MWE in the description and processing of natural language has been long recognized. In the social sciences, MWE have generally been underutilized in analyzing text corpora. Of course, uni-gram BOW representations also discard MWE (Handler et al. 2016). Moreover, bigrams and trigram BOW models do not necessarily solve this problem, as they simultaneously inflate the vocabulary with meaningless

contiguous unigrams (e.g. “from the”) and miss MWE that are longer than the preset limit.

Rights/Aspects We begin with the observation that aspects are usually based around noun phrases. To define the root of a potential aspect phrase, we combine part-of-speech (POS) tagging with a set of rules defining aspect MWEs around nouns. We rely on Justeson & Katz (1995)’s set of POS patterns to find and extract noun phrases, as MWEs. Our rule-based approach has two parameters (G_A, M_A). G_A, our aspect grammar, denotes a non-recursive regular expression that defines a set of POS tag sequences. These sequences are outputs produced from running sentences through the Stanford CoreNLP suite (Manning et al. 2014). Especially, and in contrast to previous work using n-grams, we do not specify n-grams of fixed length. For example, a noun preceded by a several adjectives would fit our rule (described in more detail below). M_A, our aspect matching strategy, defines our choice for how to scan documents to apply the patterns in G_A. It is possible to extract different patterns when you begin searching from the beginning, end or alternative initial location in the document. For this work, we begin scanning from the end of a sentence and identify noun phrases including one or more adverbs (POS tags R: RB, RBR, RBS), adjectives (POS tags A: JJ, JJR, JJS), nouns(POS tags N: NN, NNS) without including any proper nouns or prepositions in our MWE. In addition, we include a special noun phrase MWE dictionary to capture important aspects that have prepositions such as “rule of law”, “freedom of speech”. The following regular expression summarizes our set of patterns:

$$(R|A)|((A|N)*N)$$

For a concrete example, take the sentence (1) and (2), *“The government agents attacked the protesters.” “The protesters attacked the government agents.”* PULSAR would first use a POS tagger to identify the grammatical role each token plays in the sentence.

The POS tags are from Toutanova et al. (2003) and PULSAR currently uses the *Stanford CoreNLP* toolchain (Manning et al. 2014). From this POS tagging, our aspect-extraction regular expression matches the phrase [*protesters*] and [*government_agents*] respectively, since the second

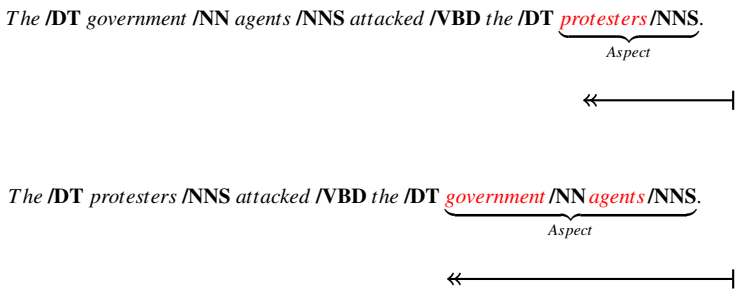

Figure 1: Extracting Aspect with MWE from Sentence (1) and (2)

sentence is preceded by a noun (NN + NNS). Since the noun (NNS) is the root of this phrase it is classified as an aspect MWE of the sentence (See Figure 1).

Judgment/Sentiment For our next task, identifying judgment MWEs, we note that judgments are often communicated in verbs. Reporting agencies note the presence, absence, giving, or taking of rights. Violations and protections are often actions or related to actions. These actions often also have an intensity provided, such as being widespread or systematic. We again use a rule-based method to identify these sequences of terms, with two parameters (G_S, M_S). G_S is now our sentiment grammar, defining a set of *the Penn Treebank* syntactic tagsets (Taylor et al. 2003). These are symbols that represent types of edges or connections between words. Our sentiment matching strategy M_S defines how we scan a sentence to apply G_S. For the sentiment expression, we take advantage of the *Penn Treebank* syntactic tree structure and develop a second grammar to capture them. After parsing a sentence for its syntactic structure presented in *CoNLL-X* format of tree bank, we start searching syntactic tagsets from the beginning of the sentence and capture all verb phrase syntactic tagsets (Syntactic tag VP) with negation *no* (POS tag ND: DT (no), RB (not)), if any. As negation changes the direction (meaning) of sentiment, this is emphasized by adding a *NEG* tag to the sentiment expression. If a sentence does not contain negation we define the verb-phrases through a set of regular expression patterns, similar to the aspect regular expressions. PULSAR attempts to keep intensifiers with the judgment expressions, using a set of finite rules

applied to the POS tagged sentence.

(ND|VP)*

Take the same sentences above Figure 2 shows the *Penn Treebank Tree* structure of the sentence.

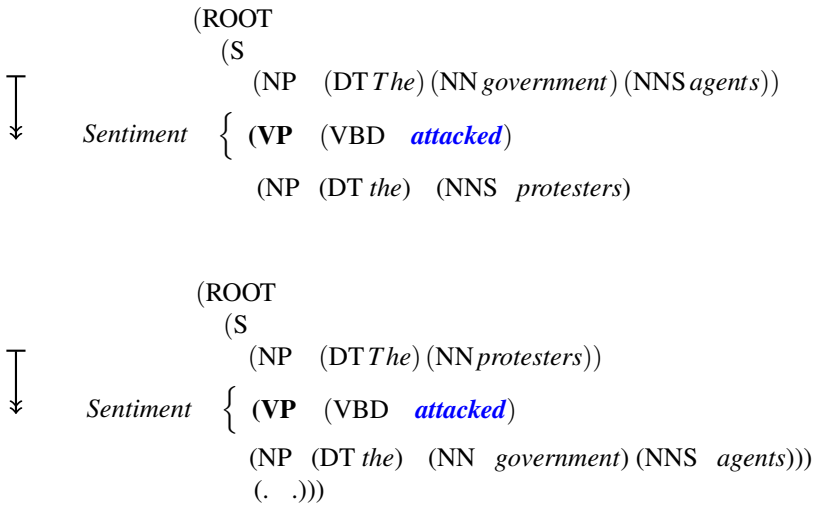

Figure 2: Penn Tree for Extracting Sentiment for Sentence (1) and (2)

From this pass over the text we capture [*attacked*] as the sentiment of the sentence by scanning from the top root to the end of each sentence.

Pairing the Right/Aspect with the Judgment/Sentiment Once we have identified potential aspects and sentiments separately, the next task is to identify whether they can be paired. Specifically, we need to identify whether the parsed sentiment/judgment is being offered on the tagged right/aspect. This can be visualized in Figure 3. In PULSAR 1.0, we again use output from the open-source Stanford CoreNLP software (Manning et al. 2014) for this task. The left figure is the original dependency parsing results, without PULSAR. It does not recognize the extracted MWE aspect above. The right figure shows the dependency relations between the extracted aspect and the corresponding sentiment in the sentence. PULSAR searches for syntactic dependencies between

an aspect MWE and sentiment MWE in the same sentence. If a specific dependency is found, a pair is formed. In our example, PULSAR outputs the sentiment-aspect pair [*attacked*, *protesters*], [*attacked*, *government_agents*].

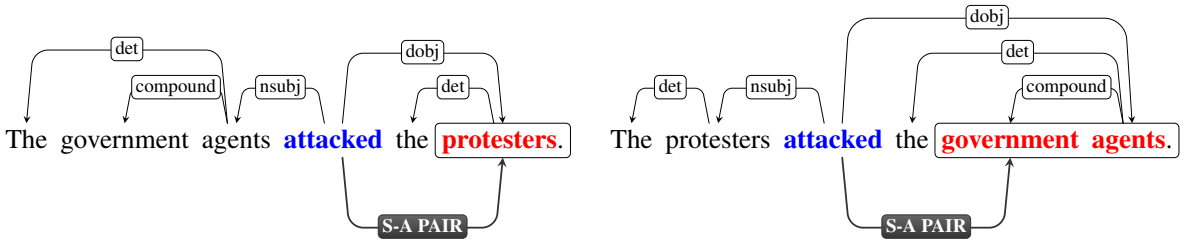

Figure 3: Dependency between Aspect and Sentiment for Sentence (1) and (2)

The sentiment-aspect pairs identified by PULSAR can then be used as new tokens and quantified for supervised or unsupervised learning tasks. The judgments can also be processed further to define positive or negative sentiment on given aspects. Moreover, aspects can themselves be topic coding or placed within a more general taxonomy (Park, Greene & Colaresi 2018). The terms within the phrases can also be stemmed, lemmatized and combined as is needed by researchers.

	[attacked, protesters]	[attacked, government_agents]	[protected, human_rights_abusers_not_victims]	[protected, victims_not_human_rights_abusers]
Sentence 1	1	0	0	0
Sentence 2	0	1	0	0
Sentence 3	0	0	1	0
Sentence 4	0	0	0	1

Table 2: The PULSAR-generated representation of the same sentences from table /refDocTermMatrix, counting the MWEs that comprise the aspect-judgment pairs. As is evident here, the vector of counts is distinct between 1 and 2 and between 3 and 4, suggesting that PULSAR can translate different semantics into counts that are missed by BOW representations.

As can be seen from these examples in Table 2, The PULSAR output is sensitive to word order, because POS tagging and syntactic parsing depends on word context. Compared to BOW representations (See Table 1), PULSAR-generated output helps us differentiate not only Sentence 1 and 2 from Sentence 3 and 4, but also Sentence 1 from Sentence 2, and Sentence 3 from Sentence 4. The question remains how the output of PULSAR compares to BOW representations of the same

texts on a well-defined social science problem.

5 Experiments: Predicting Human Coded Judgment Scores Across Physical Integrity Rights and Women's Political Rights Using PULSAR versus BOW textual features

With PULSAR, we recover what judgments composers are making on expressed rights/aspects. Here we use the tagged judgment expressions themselves instead of simplifying further towards positive versus negative versus neutral classes. We apply PULSAR to *the annual US State Department Country Reports on Human Rights Practices* in order to highlight not only how aspect-based sentiment features represent human rights over time, but also to explore whether we can more accurately forecast the human coded scores with aspect-based sentiment features as compared to BOW-generated features. The ability to accurately forecast human rights scores may provide a means of automatically coding them from the text, which should require less human hours to produce and may be more consistent. However, as noted earlier we also want to ensure that we have a good understanding of why our model makes its predictions, that is the model's features should be interpretable. One of the great benefits of using text as data is that words can be placed within context.

Our research design has two related components. First, within a set of rights, we want to know whether PULSAR-generated features are useful, as compared to BOW-generated features, in extracting the valence of judgments/sentiment across documents. Second, we want to know whether PULSAR-generated features carry information on the aspects/rights being judged. To accomplish both of these goals we need human-coded labels that measure relative judgment across issues. We choose two sets of labels that provide this information: the Political Terror Scale (PTS), numeric (1-5) ordered scale provided in Gibney et al. (2015); and the Women's Political Rights (WOPOL) scale from *Cingranelli-Richards Human Rights Dataset* (Cingranelli et al. 2014), a four

category ordered scale.[10]

If we only used one set of labels, we could compare features across different judgments (positive versus more negative), but not issue. With these two sets of labels, we can build models that are shown PULSAR or BOW features, and then compare the accuracy of the trained models in each case, as well as the explore the interpretability of the features from high performing models across labeled judgments and distinct aspects (physical integrity and women's political participation).

We use the text-based features from the *Physical Integrity Rights* section in the US State Department Human Rights country reports to predict the Political Terror Scale (PTS) numeric labels provided in Gibney et al. (2015). Similarly, since women's rights falls within the *Discrimination, Societal Abuses, and Trafficking in Persons* section of the State Department reports, we encode the text from that section to predict the WOPOL labels. Both labels are coded at the country-year level, so this defines the documents of interest, where counts of PULSAR-features and words are aggregated to the country-year report level.

Organizing the texts in this way provides a difficult task for PULSAR relative to BOW representations. By only using terms in one section of the reports, Physical Integrity Rights and then Discrimination respectively, we are already subsetting context and meaning by a given general aspect. Other work (Park, Colaresi & Greene 2018), has shown that when including the text from all the sections of a human rights reports, PULSAR-features produce a substantively important improvement over BOW representations based on accuracy and interpretability. It remains to be seen whether this is the case when applied specifically to the text of the Physical Integrity section and Discrimination. Additionally, since these labels are only provided at the country-year report level, nuance at the sentence level may be lost.

We compare the out-of-sample accuracy of bag-of-word features to sentiment-aspect features, using only Physical Integrity Rights documents in Table 3. We train each model on the text of the data from 1999-2013, and test the accuracy on data from 2014 to 2016.

We use four different algorithms in these sets of experiments to learn the mapping from the text-generated features to the labels, naive Bayes (NB), logistic regression (LR), a support vector

machine with a radial basis function (SVM), and a random forest (RF). Each algorithm is first trained only using BOW features and next using PULSAR-generated features. Note that the PTS labels have 5 classes so a random uniform baseline accuracy is .20.

We find that using sentiment-aspect pairs boosts accuracy, relative to BOW features, by approximately 2 points for the best performing algorithm (SVM). This translates into approximately 51 additional countries that would be expected to be accurately coded with the correct PTS label, using PULSAR-features, as compared to BOW features, for country reports from 1999 to 2016. This is despite the fact that the labels in this case are applied to the document as a whole, and not specific sentences. Our results provide some evidence that by accounting for syntax and grammar researchers may be better able to extract opinions and judgments, allowing for more accurate predictions.

We also find that the features that are most helpful for predicting state PTS scores are much more interpretable across the classes for the sentiment-aspect pairs that PULSAR creates, as compared to bag-of-words. Looking at both the aspects and judgments are helpful in identifying the signals that these algorithms have identified. While the bag-of-words features in Table 4 provide some face validity, some proper nouns are included in the top words. This fact suggests that these models are overlearning names, and not generalizable features that would apply to new countries (like South Sudan). Thus, as countries human rights performance changes, these models performance will necessarily degrade. There are also some curious words, such as "south", being a top feature for the worst category (5).

By comparison, the sentiment-aspect features in Table 4 are much clearer. The parser has cleared most specific named entities, to focus on abstract aspects, such as "independent judiciary" and the treatment of "civilians", "women", and "children". The sentiments are also able to capture whether there were "reports" or not, with "*neg*" representing negation. The top features also progress from "respect,govern" (1), to "killed, civilians" and other groups (5). The features in the middle categories capture new nuances, that would be impossible with only single words. If there are cases in which "mistreated, police" but countries "allow" "family_members" to visit them, and

there were no ("neg_were") reported "political_killings" and "politically_motivated_disappearances", the score is likely to be a two. More to the point, simply looking at the BOW top features such as "independent" in PTS 1, it is unclear what it refers to, but the sentiment-aspect features help us understand that "independent" is most likely to referring to "independent_judiciary" and it is "generally_enforced". Similarly, the BOW feature "mistreatment" alone does not give us much information about human rights conditions. Sentiment-aspect features specify that it most likely refers to mistreatment by the police or while in police custody ("mistreated, police").

BOW Features	Accuracy	S-A Features	Accuracy
NB	0.70	NB	0.71
LR	0.71	LR	0.72
SVM	0.71	SVM	0.72
RF	0.67	RF	0.64
Base	0.20	Base	0.20

Table 3: Out of sample accuracy of bag-of-words (BoW) features and sentiment-aspect pairs (S-A pairs) across four models, Naive Bayes, Logistic, SVM, and Random Forest.

To demonstrate that the PULSAR approach can be applied to additional contexts, and to explore whether PULSAR is learning both sentiment and aspect information, we next use the Women's Political Rights (WOPOL) from *Cingranelli-Richards Human Rights Dataset* (Cingranelli et al. 2014).

PTS 1: Best (BOW)	PTS 1: Best (S-A Features)	PTS 2: Moderate (BOW)	PTS 2: Moderate (S-A Features)	PTS 5: Worst (BOW)	PTS 5: Worst (S-A Features)
independent	[employed, government_officials]	mistreatment	[generally_respected, government]	reported	[killed, civilians]
judge	[neg_were, other_extrajudicial_killings]	trials	[has, authority]	south	[killed, soldiers]
fair	[generally_observed, prohibitions]	la	[prohibits, constitution]	forces	[raped, women]
generally	[neg_were, reports]	constitutional	[neg_were, political_killings]	numerous	[took, action]
provides	[generally_enforced, independent_judiciary]	disciplinary	[neg_were, politically_motivated_disappearances]	paramilitary	[killed, persons]
appeal	[generally_respected, practice]	accused	[visited, prisons]	children	[resulted_in, deaths]
arbitrary	[generally_observed, government]	overcrowded	[received, complaints]	facilities	[tortured, security_forces]
visits	[permitted, independent_human_rights_observers]	media	[mistreated, police]	killed	[including, children]
effective	[are_subject_to, effective_legal_sanction]	delays	[were_allowed, family_members]	baghdad	[accused_of, crimes]
act	[generally_met, conditions]	defendants	[respected_generally, government]	000	[destroy, emergency_regulations_authorities]

Table 4: Comparing BOW and ABSA Top Features based on feature weight, predicting labels 1 (best) and 5 (worst), versus all other categories (See Appendix for the full result table)

As in the case of Physical Integrity Rights, we train 4 different models on data from 1993-2008, and test the performance on data from 2009 to 2011. We find that sentiment-aspect pair features increase the out-of-sample accuracy, again, by around 2 percent, now for all the models.

BOW Features	Accuracy	S-A Features	Accuracy
NB	0.80	NB	0.82
LR	0.80	LR	0.82
SVM	0.79	SVM	0.81
RF	0.80	RF	0.82
Base	0.25	Base	0.25

Table 5: Women's Political Rights: Out of sample accuracy of bag-of-words (BoW) features and sentiment-aspect pairs (S-A pairs) across four models, Naive Bayes, Logistic, SVM, and Random Forest.

WOPOL 1: Low (BOW)	WOPOL 1: Low (S-A PAIR)	WOPOL 3: Highest (BOW)	WOPOL 3: Highest (S-A PAIR)
leveled	[provides_for, constitution]	military	[are_subject_to, military_draft]
governmental	[received, police]	receive	[pos_improve_of, women]
inhibited	[are, victims]	male	[pos_improve_of, status]
1995	[protect, victims]	religious	[often_control, family_finances]
tried	[make_up, women]	specifically	[keep, marriage]
underreported	[is, criminal_offense]	lack	[exercising, men]
common	[paying_lower_jobs, jobs]	remained	[keep, own_names]
seven	[prohibits, spousal_rape]	custody	[have_traditionally_enjoyed, active_role]
prosecute	[was_common, violence]	cited	[have_traditionally_enjoyed, high_status]
minimum	[enjoy, women]	employers	[often_seek, family_intervention]

Table 6: Women's Political Rights: Comparing BOW and ABSA Top Features based on feature weight, predicting labels 1 (low) and 3 (highest), versus all other categories

Table 6 compares the top features between a BOW approach and an ABSA approach. Investigating the features from the BOW approach provides little indication that the specific object of interest is women's political rights. We see general features such as "governmental", along with features that convey little conceptual information on rights (women's or otherwise), such as "1995" and "seven". On the other hand PULSAR-features provide more meaningful information to better understand the underlying issues within the Women's sub-section. From the Low category, it indicates that violence directed toward women is a common problem (was_common, violence) and domestic violence/sexual violence is also a critical issue (prohibits, spousal_rape), and informs us that women are often the victims of these issues. The fourth column (Highest) particularly enlightens us about the levels of women's political rights. "military" from BOW top features alone is not very informative but Sentiment-Aspect features (are_subject_to, military_draft) tells us that women also serve in the military equally to men. We also identify that women's political status has been improved by looking at (pos_improve_of, women)[11] From the feature (have_traditionally_enjoyed,

active_role), we are able to understand women's active participation in political activities in a given country.

More to the point, comparing these top features from Physical Integrity Rights to those from Women's Political Rights, it is much clearer what specific issues (aspects) are being judged. BOW features do not give us much information about the context, whereas Sentiment-Aspect features provide us with clues. For example, (generally_enforced, independent_judiciary), (killed, civilians), (neg_were, politically_motivated_disappearances), are specific to Physical Integrity Rights, thus, are unlikely to appear in Women's Political Rights. Similarly, features like (prohibits, spousal_rape), (often_control, family_finance), (have_traditionally_enjoyed, active_role) are specific to Women's rights and are unlikely to appear in Physical Integrity Rights. Thus, PULSAR can be applied to find the information that differentiates the reporting on one aspect from another, as well as usefully differentiating positive and negative judgments.

6 Conclusion

In this article, we attempted to make the case that extracting judgments on specific rights from human rights reports is an important task for researchers. Further, we argued that BOW representations are unlikely to be useful for this task because it implicitly assumes, by ignoring context and syntax, that all words have the same semantic role. Instead of giving up on the machine-aided extraction of judgments, we introduce PULSAR 1.0, a user-friendly tool that can help human rights researchers convert large-scale natural language corpora such as HROs reports into structured aspect-sentiment expressions. PULSAR has the potential to empower social scientists to move beyond topic coding and dictionary-based sentiment analysis and towards more accurately and clearly extracting aspects and judgments on aspects from speech.

PULSAR representations provide crucial new leverage for estimating the evolution of latent positions of speakers (or HROs) where votes and other sparse signals are unavailable. For example, in political science, it has been argued that decision makers often control the substance of what

is voted upon. In part, these decisions are conditioned on the preferences of the governing and opposition parties. This means that politics, as observed in votes, is only a select projection of preferences. Aspect-sentiment representations of text allows researchers to see potentially richer, less constrained, projections of political preferences, values and opinions on specific issues and topics. In this vein, Monroe & Maeda (2004) make clear that they are unable, using bag of words assumptions, to simultaneously estimate the positions of individuals and words. This limitation arises because word-scaling models treat every situation that gives rise to a word's utterance, as identical. Quite reasonably, this is why researchers often subset their analysis by a given topic (Monroe et al. 2008). Our approach to parsing text into aspect-sentiment expressions recovers both expressive phrases (akin to categorical votes), as well as the aspects that are being judged (akin to bills). Thus, PULSAR or future systems like it, open up new avenues for the systematic analysis of high-dimensional opinions and judgments at scale. PULSAR is open source software that is available for researchers to use, and we are building research guides to help ease the path to adoption.

We are currently working on augmenting PULSAR with more detailed ontologies, such as identifying the perpetrator and victim, as well as locations of violations or protections. We are also working to implement PULSAR on Amnesty International and Human Rights Watch reports, as well as blog and press release formats.[12] Further, we are creating a semi-supervised system where judgments and aspects are organized into finer-grained categories. The goal of PULSAR now and in the future is not to be a substitute for human reading and judgment, but to allow for computers to enhance our ability to track and understand human rights processes. The amount of human-rights relevant texts are expanding, and no research team can consume and analyze it all. PULSAR and systems like it can help guide human reading and also substitute for simpler systems, such as those based on BOW, that might distort the content of reports across countries and time.

Notes

[1]We define HROs broadly including human rights non-governmental organizations such as Amnesty International or Human Rights Watch, and inter-governmental organizations such as the UN Human Rights Committee, the UN Committee Against Torture, and government agencies such as the Bureau of Democracy, Human Rights, and Labor in the U.S. State Department.

[2]It is important to remember that these judgments are allegations of rights and the accuracy and usefulness of the texts depend on the source material.

[3]There are of course other concepts that can be included in this ontology, as we discuss below. Importantly, words function to link aspects to rights, identify perpetrators and denote victims.

[4]In one of our experiments, we analyze Women's rights from the State Department Reports. The reports started to discuss women's rights in a separate section in 1993. We are working to expand PULSAR to additional reporting agencies.

[5]In addition, we can think about who is communicating the judgments (such as the US State Department or Amnesty International).

[6]Another approach taken outside the human rights literature is to simply measure non-textual information such as votes instead.

[7]In supervised learning, an algorithm is trained on examples that contain observed labels or scores and then deployed on new data to predict unlabeled or unscored examples (Pang et al. 2002, Wilson et al. 2004)

[8]In rule-based learning, a set of rules are deployed to label or score new examples (Tong 2001, Turney 2002)

[9]With the growing interest in neural networks (deep learning), scholars have begun to incorporate continuous vector representation of words as features and shown improvement in capturing sentiment signals (Bespalov et al. 2011, Yessenalina & Cardie 2011, Socher et al. 2011, 2012, 2013).

[10]WOPOL includes the right to vote, the right to run for political office, the right to hold elected and appointed government positions, the right to join political parties, and the right to petition government officials (Cingranelli et al. 2014)

[11]In PULSAR 1.0, similar to "*neg*" is used to indicate the absence of aspects, "*pos*" refers to the presence of the aspects.

[12]We include some preliminary results using PULSAR on Amnesty International Reports in an online-appendix accompanying this article.

References

Bagozzi, B. E. & Berliner, D. (2016), 'The politics of scrutiny in human rights monitoring: evidence from structural topic models of us state department human rights reports', *Political Science Research and Methods* pp. 1–17.

Bespalov, D., Bai, B., Qi, Y. & Shokoufandeh, A. (2011), 'Sentiment classification based on supervised latent n-gram analysis', *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* .

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C. & Zampolli, A. (2002), Towards best practice for multiword expressions in computational lexicons., *in* 'LREC'.

Cingranelli, D. L., Richards, D. L. & Clay, K. C. (2014), The ciri human rights dataset. v.2014.04.14.

Colaresi, M. & Mahmood, Z. (2017), 'Do the robot: Lessons from machine learning to improve conflict forecasting', *Journal of Peace Research* **54**(2), 193–214.

Fariss, C. J. (2014), 'Respect for humn rights has improved over time: Modeling the changing standard of accountability', *American Political Science Review* **108**(2), 297–318.

Feldman, R. (2013), 'Techniques and applications for sentiment analysis', *Communications of the ACM* **56**(4), 82.

Gibney, M., Cornett, L., Wood, R., Haschke, P. & Arnon, D. (2015), The political terror scale 1976-2015.

Goldberg, A. E. (1995), *Constructions: A construction grammar approach to argument structure*, University of Chicago Press.

Greene, K., Park, B. & Colaresi, M. (2018), 'Machine learning human rights and wrongs: How the successes and failures of supervised learning algorithms can inform the debate about information effects', *Political Analysis* .

Hafner-Burton, E. M. (2008), 'Sticks and stones: Naming and shaming the human rights enforcement problem', *International Organization* **62**(4), 689–716.

Handler, A., Denny, M., Wallach, H. & O'Connor, B. (2016), Bag of what? simple noun phrase extraction for text analysis, *in* 'Proceedings of the First Workshop on NLP and Computational Social Science', pp. 114–124.

Hendrix, C. S. & Wong, W. H. (2013), 'When is the pen truly mighty? regime type and the efficacy of naming and shaming in curbing human rights abuses', *British Journal of Political Science* **43**(3), 651–672.

Justeson, J. S. & Katz, S. M. (1995), 'Technical terminology: some linguistic properties and an algorithm for identification in text', *Natural language engineering* **1**(1), 9–27.

Kahn-Nisser, S. (2018), 'When the targets are members and donors: Analyzing inter-governmental organizations' human rights shaming', *The Review of International Organizations* pp. 1–21.

Kay, P. & Fillmore, C. J. (1999), 'Grammatical constructions and linguistic generalizations: the what's x doing y? construction', *Language* pp. 1–33.

Keck, M. E. & Sikkink, K. (1999), 'Transnational advocacy networks in international and regional politics', *International social science journal* **51**(159), 89–101.

Kim, I. S. (2017), 'Political cleavages within industry: firm-level lobbying for trade liberalization', *American Political Science Review* **111**(1), 1–20.

King, G. & Lowe, W. (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.

Koliev, F. & Lebovic, J. H. (2018), 'Selecting for shame: The monitoring of workers' rights by the ilo, 1989 to 2011', *International Studies Quarterly* .

Laver, M., Benoit, K. & Garry, J. (2003), 'Extracting policy positions from political texts using words as data', *American Political Science Review* **97**(1), 311–331.

Lebovic, J. H. & Voeten, E. (2006), 'The politics of shame: the condemnation of country human rights practices in the unchr', *International Studies Quarterly* **50**(4), 861–888.

Lin, Y.-R., Margolin, D. & Lazer, D. (2016), 'Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of us members of congress', *Journal of the Association for Information Science and Technology* **67**(9), 2072–2089.

Liu, B. (2015), *Sentiment Analysis: Mining Opinions, Sentiments and Emotions*, Cambridge University Press, New York.

Liu, Q., Gao, Z., Liu, B. & Zhang, Y. (2013), A logic programming approach to aspect extraction in opinion mining, *in* 'Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences', Vol. 1, IEEE, pp. pages 276–283.

Liu, Q., Gao, Z., Liu, B. & Zhang, Y. (2015), Automated rule selection for aspect extraction in opinion mining, *in* 'International Joint Conference on Artificial Intelligence (IJCAI)'.

Lowe, W. (2013), There's (basically) only one way to do it. SSRN.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014), The Stanford CoreNLP natural language processing toolkit, *in* 'Association for Computational Linguistics (ACL) System Demonstrations', pp. 55–60.

Meernik, J., Aloisi, R., Sowell, M. & Nichols, A. (2012), 'The impact of human rights organizations on naming and shaming campaigns', *Journal of Conflict Resolution* **56**(2), 233–256.

Monroe, B. L., Colaresi, M. P. & Quinn, K. M. (2008), 'Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict', *Political Analysis* **16**(4), 372–403.

Monroe, B. L. & Maeda, K. (2004), Talk's cheap: Text-based ideal point estimation, *in* 'presented to the Political Methodology Society', Palo Alto, CA.

Murdie, A. (2014), 'The ties that bind: A network analysis of human rights international non-governmental organizations', *British Journal of Political Science* **44**(1), 1–27.

Murdie, A. M. & 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.

Pang, B. & Lee, L. (2008), 'Opinion mining and sentiment analysis', *Foundations and Trends in Informational Retrieval* .

Pang, B., Lee, L. & Vaithyanathan, S. (2002), 'Thumbs up?', *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP '02* .

Park, B., Colaresi, M. & Greene, K. (2018), 'Beyond a bag of words: Using pulsar to extract judgments on specific human rights at scale', *Peace Economics, Peace Science and Public Policy* **24**(4).

Park, B., Greene, K. & Colaresi, M. (2018), 'Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects', *University of Pittsburgh Working Paper* .

Poria, S., Ofek, N., Gelbukh, A., Hussain, A. & Rokach, L. (2014), Dependency tree-based rules for concept-level aspect-based sentiment analysis, *in* 'Semantic Web Evaluation Challenge', pp. 41–47.

Qiu, G., Liu, B., Bu, J. & Chen, C. (2011), 'Opinion word expansion and target extraction through double propagation', *Computational linguistics* **37**(1), 9–27.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. & Radev, D. R. (2010), 'How to analyze political attention with minimal assumptions and costs', *American Journal of Political Science* **54**(1), 209–228.

Risse-Kappen, T., Risse, T., Ropp, S. C. & Sikkink, K. (1999), *The power of human rights: International norms and domestic change*, Vol. 66, Cambridge University Press.

Roberts, M. E., Stewart, B. M. & Airoldi, E. M. (2014), 'Structural topic models', *Retrieved May* **30**, 2014.

Ron, J., Ramos, H. & Rodgers, K. (2005), 'Transnational information politics: Ngo human rights reporting, 1986–2000', *International Studies Quarterly* **29**(3), 557–587.

Schrodt, P. A., Beieler, J. & Idris, M. (2014), Three'sa charm?: Open event data coding with el: Diablo, petrarch, and the open event data alliance, *in* 'ISA Annual Convention'.

Slapin, J. B. & Proksch, S.-O. (2008), 'A scaling model for estimating time-series party positions from texts', *American Journal of Political Science* **52**(3), 705–722.

Socher, R., Huval, B., Manning, C. D. & Ng, A. Y. (2012), Semantic compositionality through recursive matrix-vector spaces, *in* 'Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning', Association for Computational Linguistics, pp. 1201–1211.

Socher, R., Lin, C. C., Manning, C. & Ng, A. Y. (2011), Parsing natural scenes and natural language with recursive neural networks, *in* 'Proceedings of the 28th international conference on machine learning (ICML-11)', pp. 129–136.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. & Potts, C. (2013), Recursive deep models for semantic compositionality over a sentiment treebank, *in* 'Proceedings of the 2013 conference on empirical methods in natural language processing', pp. 1631–1642.

Stefanowitsch, A. & Gries, S. T. (2003), 'Collostructions: Investigating the interaction of words and constructions', *International journal of corpus linguistics* **8**(2), 209–243.

Stroup, S. S. (2012), *Borders among activists: international NGOs in the United States, Britain, and France*, Cornell University Press.

Taylor, A., Marcus, M. & Santorini, B. (2003), The penn treebank: an overview, *in* 'Treebanks', Springer, pp. 5–22.

Tong, R. M. (2001), An operational system for detecting and tracking opinions in on-line discussion, *in* 'Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification', Vol. 1.

Toutanova, K., Klein, D., Manning, C. D. & Singer, Y. (2003), Feature-rich part-of-speech tagging with a cyclic dependency network, *in* 'Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1', Association for Computational Linguistics, pp. 173–180.

Turney, P. D. (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *in* 'Proceedings of the 40th annual meeting on association for computational linguistics', Association for Computational Linguistics, pp. 417–424.

Wilson, T., Wiebe, J. & Hwa, R. (2004), Just how mad are you? finding strong and weak opinion clauses, *in* 'aaai', Vol. 4, pp. 761–769.

Wu, Y., Zhang, Q., Huang, X. & Wu, L. (2009), Phrase dependency parsing for opinion mining, *in* 'Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing', Vol. 3, Association for Computational Linguistics, pp. 1533–1541.

Yessenalina, A. & Cardie, C. (2011), Compositional matrix-space models for sentiment analysis, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', EMNLP '11, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 172–182.