Machine Learning Human Rights and Wrongs: How the Successes and Failures of Supervised Learning Algorithms Can Inform the Debate About Information Effects

Kevin T. Greene
Department of Political Science

Baekkwan Park
Department of Political Science

Michael Colaresi
Department of Political Science
mcolaresi@pitt.edu

Abstract

There is an ongoing debate about whether human rights standards have changed over the last 30 years. The evidence for or against this shift relies upon indicators created by human coders reading the texts of human rights reports. To help resolve this debate, we suggest translating the question of changing standards into a supervised learning problem. From this perspective, the application of consistent standards over time implies a time-constant mapping from the textual features in reports to the human coded scores. Alternatively, if the meaning of abuses have evolved over time, then the same textual features will be labeled with different numerical scores at distinct times. Of course, while the mapping from natural language to numerical human rights score is a highly complicated function, we show that these two distinct data generation processes imply divergent overall patterns of accuracy when we train a wide variety of algorithms on older versus newer sets of observations to learn how to automatically label texts with scores. Our results are consistent with the expectation that standards of human rights have changed over time.

We apply a text-as-data approach to provide new insights into whether the standards of judging human rights behaviors may be evolving over time. Using the framework for supervised learning, we find that using older human rights reports to automatically learn the rules by which teams have coded human rights scores lowers the accuracy of predicted scores generated from recent reports. This pattern provides new evidence of meaningful underlying changes in the coding of quantitative human rights measures over time. Human rights measures may need to be calibrated for these dynamic information effects to increase their comparability over time and space. More broadly, these changes may not simply represent linearly increasing standards of human rights, but involve more complicated patterns of distinct new aspects being judged in more recent reports.

1 Motivation

A debate has emerged around the question of whether changing standards of human rights have polluted highly influential quantitative measures of state practices. Clark and Sikkink (2013), building on work by Keck and Sikkink (1998), argue that the greater availability of information on abuses around the world over time has led to differences in the newer versus older textual reports that are used to code the Political Terror Scale (PTS) and the Cingranelli-Richards Human Rights Data (CIRI). One piece of evidence offered to support this conjecture is that the reports have become longer in more recent years. Fariss (2014) extends these arguments to suggest that numerical scores may indicate deteriorating human rights practices over time, not because actual human rights violations are increasing, but because later violations are more likely to be recorded in the underlying texts and in the human coded scales.[1]

There are several potential mechanisms by which information effects could alter human rights scores. Coding information effects could occur when human coders read texts differently across years. For example, it is likely that coders have access to more information about countries in recent times, and so might inadvertently augment their scores in recent years with information that is not in the text, while they rely more closely on the text in earlier years. Compositional information effects, alternatively, can occur when different aspects of human rights violations are recorded in the texts over time. It could be the case that details of violations, such as sexual violence, are discussed systematically in later but not earlier reports. When human coders read these recent reports, seeing more evidence of violations, a worse numerical score may be recorded. Fariss (2014) uses word counts of the country-report texts as evidence of dynamic instability in the composition of the documents by agencies such as the US State Department and Amnesty International, while suggesting other potential measurement concerns. More recently, Bagozzi and Berliner (forthcoming) discover evidence of changes in the underlying distribution of words in the country-reports over time. Together, this research suggests that at least compositional information effects might bias human rights data over time.

In a response to these arguments, Richards (2015) finds little evidence of the "information effects" noted by Clark and Sikkink (2013). He argues that the aggregate human rights scales are stable, with few exceptions. Richards also notes that it is not the length of the documents used to code the scores that matters, as suggested by previous work, but the specific words used in the reports. Terms such as "widespread", "systematic", "extensive" cause a state to be coded as having poor respect for an aspect of human rights, regardless of the length of the report, as is consistent with the CIRI codebook. Thus, Richards (2015) is largely arguing that coding informational effects are negligible, even if compositional effects, leading to longer reports, are present, since there should be a constant mapping from key words that appear in the text to human rights scores.

[1]These authors also address other potential issues with the creation of scores over time such as possible ceiling effects. We focus on their arguments relating to changing standards over time for this letter. See Clark and Sikkink (2013); Fariss (2014).

If we are to assess the impact key concepts such as regime type and political contention have on human rights practices, we must understand whether the human scoring of texts is consistent over time. Moreover, The quantitative study of human rights is likely to continue its rise in prominence in the future as concepts such as Responsibility to Protect and international humanitarian interventions in regional crises continue to draw significant global attention.[2]

2 Approach

We contribute to this debate by applying a research design strategy inspired by work in machine learning to compute how the underlying texts have been translated into the Political Terror Scale (PTS). We treat the conversion of the natural language in the human rights reports into quantitative scores of human rights practices as a supervised learning problem (Kotsiantis, 2007; Mitchell, 1998). Supervised learning is a branch of Machine Learning whereby researchers attempt to compute a mathematical representation of the unknown process by which input features, such as word counts from texts, predict continuous or discrete response values, such as quantitative human rights scores. Crucially, since the goal of machine learning is to build models that generalize to unseen data, researchers in this area have emphasized strategies that can identify overfitting the sample data while avoiding limitations that lead to underperformance (Flach, 2012).

Our aim in this letter is to use supervised learning algorithms to evaluate competing arguments in an important theoretical debate. The two sides of the information effects debate can be represented by two distinct sets of heuristics for building an algorithm that attempts to learn how the lexical features of the human rights texts are mapped into quantitative scores by human coders. We define this map as a function $y_{it} = f_t(x_{it})$, where y_{it} is a known scalar human rights score for country $i \in (1 \dots N)$ at time $t \in (1 \dots T)$, x_{it} is a known vector of textual features such as term frequencies, and $f_t(\cdot)$ is a an unknown function projecting x_{it} to y_{it}, that can depend on time. If the informational content within the text that is relevant for coding a particular human rights score has changed, either due to coding or compositional effects as suggested by Clark and Sikkink (2013), then $f_t(\cdot) \neq f(\cdot), \forall t \in (1, \dots, T)$. Therefore, an algorithm that attempted to learn the function that created the examples in a recent training set, $f_t(x_{it})$ by training a model on older instances would instead be fitting a representation of $f_{t-i}(x_{i,t-j}),\ t > j > 0$. This algorithm would be learning an older version of the function, as opposed to the process that generated the test set. While this overfitting is not observable within the training window, models trained on stale instances would not generalize to recent out-of-sample cases, degrading out-of-window performance relative to models trained on instances that were in closer temporal proximity to the test set.[3] Conversely, if the information in the texts that is relevant for creating quantitative scores has been stable over time, even if texts are longer or more embellished, as suggested by Richards (2012), then there should be a consistent translation from the texts to the measures across years, and thus $f_t(\cdot) = f(\cdot), \forall t \in (1, \dots, T)$. In the appendix, we provide simulation evidence that we can differentiate static and changing patterns of coding inputs into a score, utilizing this research design.

Because we are using the underlying texts, we are also able to examine which lexical features have changed in importance over time (Grimmer and Stewart, 2013; Quinn et al., 2010; Monroe, Colaresi and Quinn, 2008).

[2] See also Fariss (2014) and Bagozzi and Berliner (forthcoming).

[3] We use the terms in-window and out-of-window to differentiate our sampling strategy from those that measure accuracy on the observations that are specifically used to train fitted models (in-sample accuracy).

3 Empirics

To assess the arguments above, we use the text from the State Department Human Rights Reports for the years 1978 through 2010.[4] Each document (yearly country-report) is represented by a feature count vector, modeled as a bag-of-words.[5] Using these word features we train a number of machine learning algorithms: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF), as well as a majority vote ensemble classifier[6] to predict a given state's PTS score in a particular year.[7]

For each classifier we divide the dataset into two periods: (1) 1978-2005 and (2) 2006-2010 and use (1) to draw the training set and measure in-window performance and (2) for out-of-window testing. All the models are computed and evaluated first within a specific training window with 5-fold [8] cross validation, to check the performance of the model on temporally proximate data, then out-of-window using set (2) (2006-2010), allowing for meaningful comparison of accuracy for models trained on different epochs.[9]

The top graph in Figure 1 displays the out-of-window prediction accuracy[10] for each of our algorithms trained on unique 10 year rolling windows.[11] If there were no information effects, and scores were coded consistently from the text, we would expect out-of-window accuracy to be similar across sampling windows, as the features that predicted particular PTS scores in the early period would predict the same in the later period. However, we find that as the reports used to fit the model become more distant from the test set (2006-2010), the accuracy falls. This is a pattern that is consistent with information effects and the time-varying generative model from the simulations in the appendix. Since the algorithm is perfectly reliable in how it maps words into predicted PTS scores[12], then there is something else that is unique to the human coded scores that is reducing the accuracy.

The constant in-window accuracy, computed using hold-out evaluation set within the training window, presented in the middle plot in Figure 1 suggests that the later scores are not more difficult to learn from the texts than earlier periods across a number of different algorithms.[13] Within a given temporal range, an algorithm can learn how to produce PTS scores with an average accuracy of approximately .75. The set of patterns we observe – changing out-of-window accuracy with consistent in-window accuracy – points to a dynamic data generation process, but also supports Richards (2015)'s contention that the coding process was not necessarily noisier or more difficult in earlier periods.[14]

[4] While other reports, notably Amnesty International, may communicate different facets of human rights, investigating the systematic differences between reports is beyond the scope of this letter.

[5] We also explore the role of higher order n-grams, as it is possible that these contain additional information.

[6] We thank the editor for this suggestion.

[7] We run a total of 96 models, not counting the ensemble voting classifier. A more in-depth review of the data collection process and the algorithms used can be found in the supplemental appendix.

[8] We also tried 10-fold cross validation and there is no meaningful difference in the results.

[9] We choose this split to mirror the arguments made by Fariss (2014) and Clark and Sikkink (2013) as they relate to changes in human rights over time. Holding out more recent data will allow us to successively draw older and older training windows. Accuracy is presented in the main text and Precision, recall, and F1 were also computed and available in the supplemental appendix.

[10] Out-of-window accuracy is the proportion of cases that a classifier correctly predicts. A mistake of one level (predicting 3 when the actual value was a 4), is treated identically as missing by 4 levels. We present more complete confusion matrices in the appendix.

[11] Since PTS has 5 categories, the baseline accuracy for an algorithm that randomly guesses is .2.

[12] With a given model, the identical text will always map to the same score.

[13] We also explore, using our random forest ensembles, the in-window accuracy using 5-fold cross-validation. The same flat within-window pattern holds.

[14] As we discuss further below, another interesting test of these propositions would be to analyze the patterns of inter-coder reliability over time.

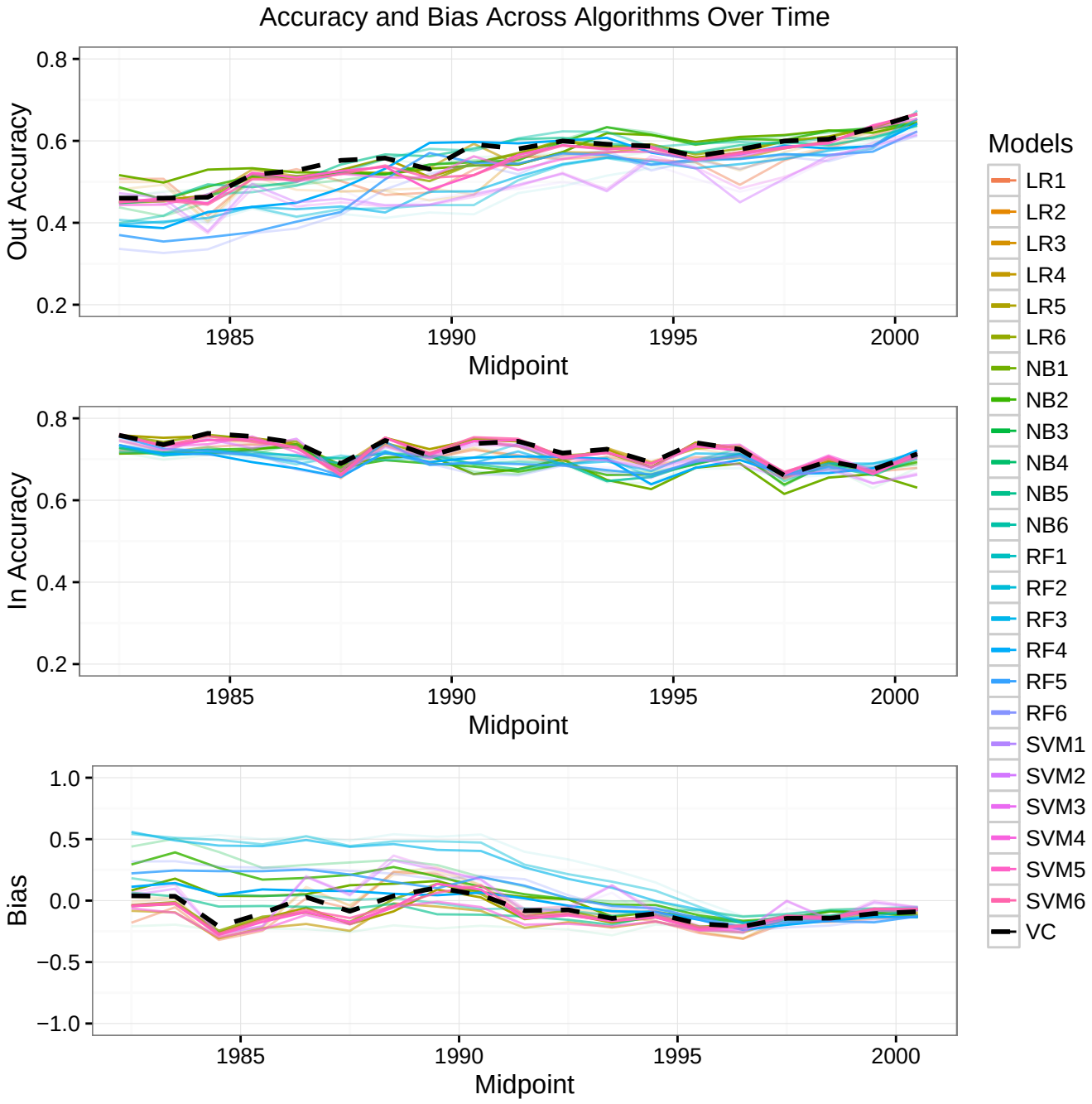

Figure 1: The out-of-window accuracy for algorithms trained on overlapping/rolling 10 year windows of the lexical report features in the top plot illustrates an upwards trend, consistent with a dynamic human rights score generation process. The in-window accuracy, calculated on a held-out evaluation set from within the time period used to train the models, is plotted in the middle window, and by comparison is flat, suggesting that the PTS data has not grown more difficult to learn over time. The lower plot illustrates the average bias in the out-of-window period (predicted value - observed value). There is not a consistently negative bias for models using older data. The black dashed line in each plot represents an ensemble majority vote classifier that takes the predictions of each of the other algorithms as input, and predicts the class that has the plurality of votes. The keys for the algorithms are defined in the appendix.

Another piece of evidence can be found in the direction of the mistakes that our algorithms are making. Fariss (2014) suggests that standards for human rights have grown more stringent over time, leading to higher scores in later years even if state behavior has remained consistent. If the human composers of the text applied these more stringent standards in later years utilizing new, more extreme language, then our algorithms that learned the mapping between the text and the score in earlier years would be systematically lower (indicating less egregious violations) than the actual scores in later years; since the more stringent standard would have been applied by humans and encoded into terms that were not available to our algorithms.[15] The bottom plot of Figure 1 presents the bias, our forecast minus the actual value, across our classifiers for different rolling windows on the out-of-window test set.[16] We do not see a clear pattern of a negative bias across the algorithms. In fact, several of the algorithms have a positive bias, suggesting that the actual human scores were lower (fewer abuses) than expected from the text alone in the later reports.

Next, we turn to what is changing within the algorithms that allows them to be more accurate when trained on later, as compared to earlier, human coded reports. We present a selection of terms that either are in the top 25 highest ranked relative feature importances (first six rows) drawn from the Random Forests algorithm or are flagged in the PTS codebook or by Richards (2015) as keywords that should consistently imply a specific score (last two rows). The rank across each of our fitted windows is then logged and displayed in figure 2.[17]

The plot reveals several interesting patterns that provide some insights into the nature of the changes in the State Department text over time. In the top two rows, we present several of the top terms that illustrate how information for violations themselves are discussed. The term "according" is often used in recent texts to reference specific sources of information and has increased (particularly for 4s), while the less specific term "reported", has declined in importance. Interestingly, two terms that signal a discussion of meaning and reliability of information, the terms "interpretation" and "reliable" have declined in importance for a score of 5. Likewise, in the second row, we see that "committed" signaling a discussion of who perpetrated an abuse in this context has increased while the more general "scale" has declined; instead of talking of overall victims, there is a shift to specific "populations" that might be effected. There is also a shift in the actors being referenced. There is an increase in the importance of "armed" groups, as opposed to formal "armies" and "regimes", as well as talking about "commanders" and how fighters are "conscripted". Similarly, there is an increasing focus on "civilians".

One of the more striking changes over time is the number of different types of abuses that are increasingly influential in the random forest predictions over time. There is an increase in discussion of "internally" "displaced" people, as well as "humanitarian" concerns such as "food". The use of "mines" and events where victims were "raped" are identified as crucial in more recent, but not past, scoring of human rights texts.

On the other hand, there is not clear evidence that terms that would encode the severity of violations have shifted. There were changes in the importance of terms such as "frequently", "numerous", and "widespread", particularly for scoring a 4. However, words noted as important in the PTS codebook such as "regularly" and "routinely" provide somewhat less predictive power overall, but are consistent over time (see the last two rows). The use of the term "common" is consistently important. Of course, as

[15] Another possibility that our research design helps to alleviate is whether the actual human coders themselves are seeing the same language but coding different scores at distinct times.

[16] If a classifier's best guess for a countries score in a year, given the report, is a 3, but the actual human coded score was a 4, the bias is -1.

[17] The top terms are selected based on the highest feature importance averaged over the first and last three windows of the 10 year rolling models. The full output for these top features can be found in the supplemental appendix. The terms are unstemmed in our analysis, so we include several tokens that represent potentially important stems.

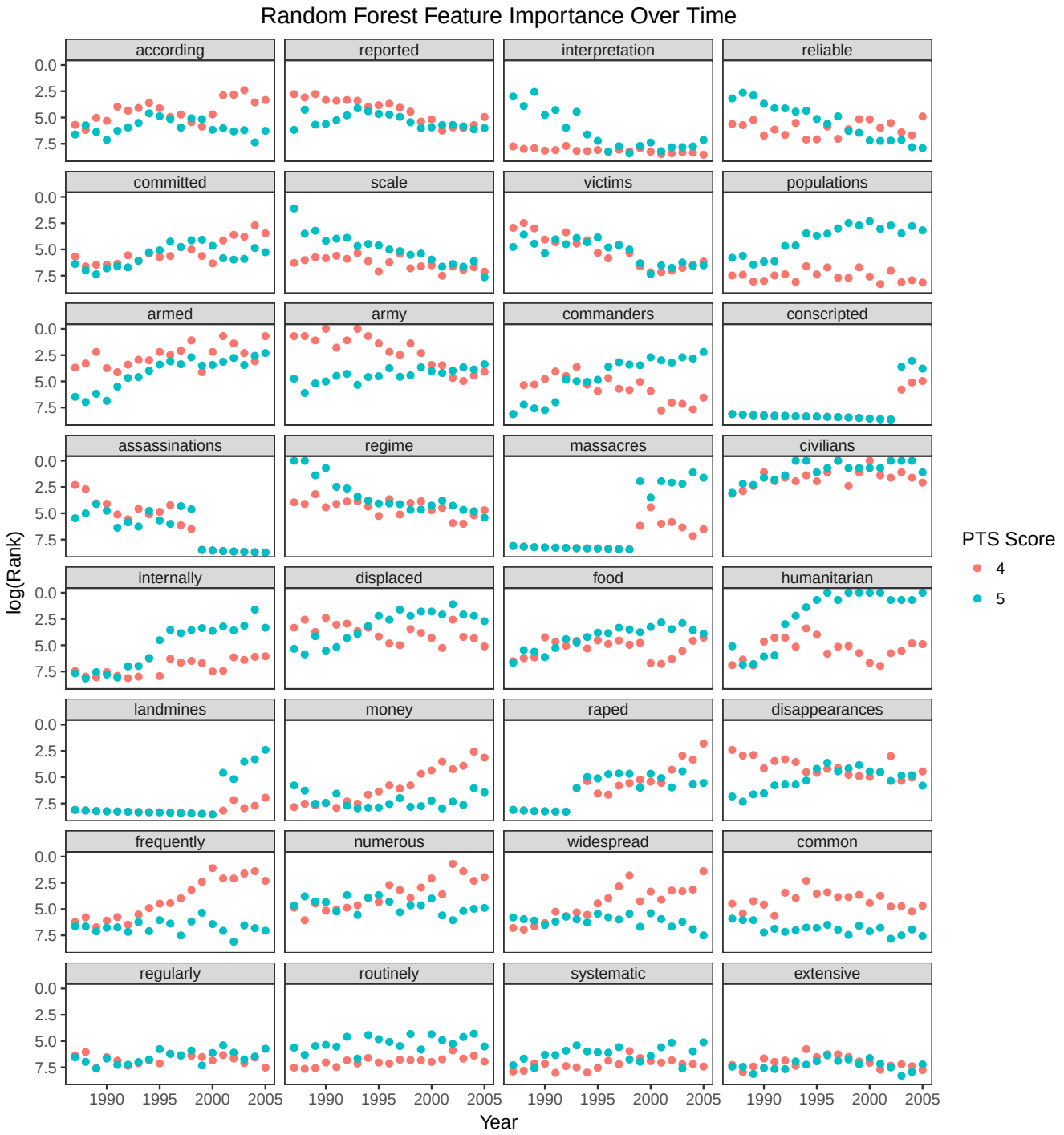

Figure 2: Each subplot presents the descending rank (logged) of select terms based on their feature importance for predicting PTS label 4 (red) and 5 (blue) using the unigram Random Forests Model and GINI impurity. The last year of the 10 year rolling training windows are presented on the x-axis. The y-axis is reversed so that greater importance values (lower ranks) are higher on the plot. Feature importance is computed on decision trees using unigram features within a random forest. The trees in the forest are trained to classify a report as either a 4 or not, and then as either a 5 or not. We use the change in gini-impurity from parent to child nodes, weighted by the number of observations that reach each node as our feature importance metric. These values are averaged across the trees. A flat line represents consistent importance in predicting recent scores over the past training windows. The first six rows of plots are all within the top 25 ranked features for either the earliest of the latest windows. The last two rows include words that we pulled from the PTS codebook and Richards discussion.

these words are generally used to modify other important human rights features, we might find them to be more important if they were linked to the aspects they modify, such as "routinely killed" or "regularly engaged in killings".

Our results provide some support for Clark and Sikkink's claim that aspects of human rights may receive greater focus in later reports. However, these changes could be due to coder effects, where the humans scoring the reports are reading the reports from earlier years differently than later years, or because of specific compositional effects where terms that coders use to assign specific scores appear or disappear from texts over time. Either way, we have provided evidence that there are more than superficial changes in the underlying process that generates human rights scores.

4 Discussion

While we do not claim to have settled the debate over information effects in human rights scores over time, we do believe our approach has provided new insights into whether the standards of judging state behavior may be evolving, while also providing a framework for using supervised learning to inform theoretical debates. We find that the older the human rights reports used to learn the rules by which teams have coded human rights scores, the lower the accuracy of predicted scores generated from recent reports; suggesting there is some underlying change in the coding of the human rights measures from the texts over time. Next, research teams should begin building models with time-varying parameters that can attempt to learn the dimensionality and evolving structure of human rights scores. There are already several useful projects ongoing in this area (Bagozzi and Berliner, forthcoming; Fariss et al., 2015). In parallel, it would be useful to return to analyzing the relative patterns of human coders. Our argument suggests that inter-coder reliability might fall as new documents, draw from a different time period than the set used to train coders, are scored. In this future exercise, humans would take the place of our algorithms, and would be trained on documents from a specific time-window.[18]

More generally, as the role of prediction in political science continues to grow (Ward, 2016), we hope that our letter makes the case that machine learning approaches can be deployed to create new knowledge on core political science questions. This approach can still be heavily grounded in theory and domain expertise, while allowing researchers to address questions from new angles. In particular as many studies implicitly or explicitly refer to components of text, such as word counts, or lexical features, the use of computational text analysis, coupled with machine learning, may be particularly fruitful.

Finally, using bag of word assumptions we were able to develop an automated coding system for PTS that achieves out of sample accuracy well above 70%, a large improvement over random chance (20%). By accounting for syntactic features of the text this accuracy could likely be improved. Such a system could provide a means to develop a quicker and more consistent coding of human rights data over time, possibly helping to alleviate some of the concerns posed by Clark and Sikkink (2013).

5 Bibliography

References

Bagozzi, Benjamin and Daniel Berliner. forthcoming. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of U.S. State Department Human Rights Reports." *Political Science and Research Methods* .

[18] We thank the editor for this suggestion.

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

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

Fariss, Christopher J., Fridolin J. Linder, Zachary M. Jones, Charles D. Crabtree, Megan A. Biek, Ana-Sophia M. Ross, Taranamol Kaur and Michael Tsai. 2015. "Human Rights Texts: Converting Human Rights Primary Source Documents into Data." *PLOS ONE* 10(9).

Flach, Peter. 2012. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. New York: Cambridge.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21(3):267–297.

Keck, Margaret and Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Itaca, NY: Cornell University Press.

Kotsiantis, S.B. 2007. "Supervised Machine Learning: A Review of Classification Techniques." *Informatica* 31(1):249–268.

Mitchell, Tom. 1998. *Machine Learning*. New York: McGraw-Hill.

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

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Richards, David L. 2012. "Rhetoric and Reality' Revisited." *Journal of Human Rights* 11(3):337–343.

Richards, David L. 2015. The Myth of Information Effects: A Reply to Clark and Sikkink. Working paper University of Connecticut.

Ward, Michael. 2016. "Can we predict politics? Toward what end?" *Journal of Global Security Studies* 1(1):80–91.