MINING TEXTS TO EFFICIENTLY GENERATE GLOBAL DATA ON POLITICAL REGIME TYPES

JAY ULFELDER, SHAHRYAR MINHAS, AND MICHAEL D. WARD

ABSTRACT. We describe the design and results of an experiment in using text-mining and machine-learning techniques to generate annual measures of national political regime types. Valid and reliable measures of country's forms of national government are essential to cross-national and dynamic analysis of many phenomena of great interest to political scientists, including civil war, interstate war, democratization, and coups d'état. Unfortunately, traditional measures of regime type are very expensive to produce, and observations for ambiguous cases are often sharply contested. In this project, we train a series of Support Vector Machine (SVM) classifiers to infer regime type from textual data sources. To train the classifiers we used vectorized textual reports from Freedom House and the State Department as features for a training set of prelabeled regime type data. To validate our SVM classifiers, we compare their predictions in an out-of-sample context and the performance results across a variety of metrics (accuracy, precision, recall, F-1 score) are very high. The results of this project highlight for us the ability of these techniques to contribute to producing real time data sources for use in police science that can also and that can be routinely updated at much lower cost than human-coded data.

Date: Presented at the Annual Meeting of the American Political Science Association in Washington, DC. This research was funded in part by NSF Award 1259190, Collaborative Research: Automated Real-time Production of Political Indicators. We thank Philip A. Schrodt for helping to develop this research design and Alex Hanna for comments on an earlier version of this paper. All errors are our own.

1. INTRODUCTION

Time-series cross-sectional data on countries' national political regime types feature prominently in virtually all research on the antecedents and causes of political development and conflict. In the past 25 years, conceptualization and measurement of democracy has become one of a few dominant themes in the subfield of comparative politics, and recent work on variations in forms of autocracy is growing in prominence as well. What's more, these debates over concepts and measures are not simply theoretical; they can have significant effects on research that shapes policy and advocacy as well. For example, myriad studies have shown that national regime type shapes countries' propensity to go to war with each other or their own citizens Hegre (2014); work by the Political Instability Task Force has shown regime type to be the single most-powerful predictor of onsets of national political crises such as civil wars, coups, and state collapse Goldstone et al. (2010); and work by numerous scholars suggests that the effects of events like sanctions Geddes (2002), Escribá-Foch and Wright (2010) and social unrest Ulfelder (2005) on the likelihood of authoritarian regime breakdown are conditional on formal and informal institutional features of those regimes.

Of practical importance, the measures of regime type that are widely used today are expensive to produce. Polity IV is coded by hand, so the data set's producers must spend many hours locating and reviewing relevant documents for upwards of 160 countries (Marshall and Jaggers 2002). Freedom House's Freedom in the World depends on repeated surveys of a large number of subject-matter experts. The newly launched Varieties of Democracy project, or V-Dem, does the same (Coppedge et al. 2013). Other widely cited data sets on national political regimes–including the Democracy and Dictatorship Dataset (Cheibub, Gandhi and Vreeland 2010) and the Autocratic Regimes Dataset (Geddes, Wright and Frantz 2014) – are not routinely updated, and the high cost of doing so appears to be one reason why. The Unified Democracy Scores (UDS) data set thoughtfully addresses the problem of uncertainty about one crucial regime concept (Pemstein, Meserve and Melton 2010), but it is derived from, and therefore wholly dependent on, the human-coded data sets that are so labor intensive to produce.

This paper explores ways to use text mining in combination with natural language processing techniques to generate measures of regime type. By relying on texts that are routinely produced and publicly available, we expect to be able to generate measures of regime type at much lower cost and at much greater frequency. Tremendous growth in global reportage on democracy and human rights in the past 20 years has produced a massive corpus of texts on these topics that is ripe for content analysis. Concurrent improvements in computer software and hardware have made the process of analyzing large bodies of text much more efficient, and the field has matured with the development of a common set of methodologies with well-tested characteristics. Automated coding is completely transparent, so it avoids the unreproducible subjective elements of human coding. Once the source texts have been prepared, recoding to account for new theoretical or technological components can be done relatively quickly and efficiently. Last but not least, innovations in methods for content analysis are helping researchers use those tools to produce more interesting and more useful results, including a number of applications to political texts. These approaches are increasingly being used to generate political event data D'Orazio et al. (2014), King and Lowe (2003), O'Connor, Stewart and Smith (2013) and to measure things like political tensions Chadefaux (2014), partisan affiliation (Slapin and Proksch 2010, Yu, Kaufmann and Diermeier 2008), and legislative agendas (Grimmer 2010). To our knowledge, though, these techniques have not previously been applied to the task of measuring political regime types over time across many countries.

We start by selecting a corpus of familiar and well-structured texts describing politics and human-rights practices each year in all countries worldwide: the Country Reports on Human Rights Practices published by the U.S. Department of State, and Freedom House's Freedom in the World reports. After pre-processing those texts, we merged the reports for each country-year into bags of n-grams and used text mining to extract features from those bags in the form of vectorized tokens. Next, we used those vectorized tokens as inputs to a series of binary classification models representing a few different ideal-typical regime types – democracy, military rule, one-party rule, and monarchy – as observed in few widely used, human-coded data sets. Finally, we applied those classification models to a test set of country-years held out at the start to assess the models' ability

to classify regime types in cases they had not previously "seen." Our results demonstrate that this strategy works. Our classifiers perform well out of sample, achieving high or very high precision and recall scores in cross-validation on all four of the regime types we have tried to measure so far.

2. Raw Data

For this project's initial iteration, we limited the corpus to annual reports available online from the U.S. Department of State and Freedom House. Both of these series of reports have structures that are explicitly similar and use equivalent concepts and language to describe political practices in the referent countries.

As the State Department notes on its website, its human rights reports "cover internationally recognized individual, civil, political, and worker rights, as set forth in the Universal Declaration of Human Rights and other international agreements." The annual production of State's human rights reports is mandated by law under the Foreign Assistance Act of 1961 and the Trade Act of 1974. All U.N. member states and other countries receiving U.S. assistance must be covered.

In its Freedom in the World reports, the U.S.-based non-governmental organization Freedom House provides summary descriptions of all countries of the world and some disputed territories. Per its website, these descriptions are intended to capture the advance or decline of "freedom" in each polity, where the idea of freedom "is grounded in basic standards of political rights and civil liberties, derived in large measure from relevant portions of the Universal Declaration of Human Rights." Freedom House is a non-governmental advocacy group that receives a substantial share of its funding from the U.S. government.

We scraped the State Department's annual reports on countries' human rights practices from 1999 to 2013 and Freedom House's Freedom in the World reports from 1999 to 2014, which cover the period 1998-2013. These are all of the reports in the two series that are currently archived online on the source organizations' web sites.

3. Preparing Texts

Once those documents had been ingested, we cleaned the texts and did some simple feature extraction. Specifically, we:

- Removed numbers, punctuation, and stop words;
- Removed proper names and acronyms;
- Tokenized the results; and
- Lemmatized the tokens to group together inflected forms.

Next, we calculated the tf-idf weight for each token and used these values as inputs to a series of binary classification models designed to identify features associated with each of a few regime types suggested by prior theory. Employing tf-idf weights is common practice in information retrieval from textual data and has been successful in document classification.

4. RESEARCH DESIGN

We focus on four regime types: (1) Democracy, (2) Military rule, (3) One-party rule, and (4) Monarchy. In principle, these categories are mutually exclusive. In practice, however, regime type is inherently uncertain, and any given regime may exhibit features of more than one of these ideal types at any time. Consistent with that idea, we decided to treat each of these archetypes not as different sections of a single plane but rather as distinct dimensions in a multi-dimensional regime space, and therefore train separate models for each category.

To train our regime classification models, we used pre-existing, human-coded data on national political regimes from a few widely-used sources: Polity; Geddes, Wright, and Franz (2014); and Hadenius and Teorell (2007; see also Wahman Teorell and Hadenius 2013). For our democracy model, we created a four level ordinal variable that equals one if a country's Polity rating is between 10 to 6, 5 to 1, 0 to -5, and -6 to -10. For our models of military rule, one-party rule, and monarchy, we coded our binary indicators as one if both GWF and HT identified a country-year as belonging to that category and only that category and zero otherwise. In other words, we only considered a regime to be sufficiently representative of each of those authoritarian archetypes if both sources agreed that it belonged in that category and only that category. Conceptually, we want to learn from the cases of whose status we are confident and then leave it to the models to tell us how closely the more ambiguous cases approximate those archetypes.

To objectively estimate the performance of these classification models, we split our sample into a training and test set. For the democracy model, we have "ground truth" data from Polity through 2013, so we made 1999-2008 (n = 1,557) the training set and 2009-2013 (n = 707) the test set. For the three authoritarian archetypes, we only have "ground truth" data from GWF and HT through 2010, so we made 1999-2006 (n = 1,138) the training set and 2007-2010 (n = 583) the test set.

5. METHODS

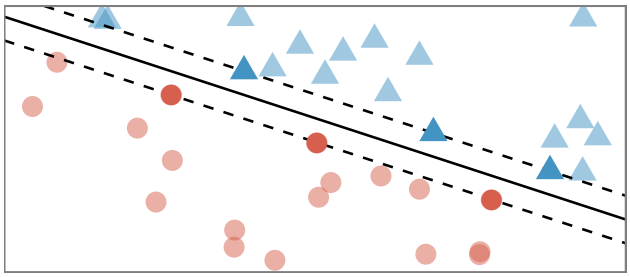
We use support vector machines (SVMs) to classify our text-based data sources. SVMs work by using a hyperplane that separates the classes of labeled input data with the maximum margin of separation (Vapnik 2000, **?**). Documents on either side of the hyperplane correspond to the classes of the binary dependent variable, new documents are then classified according to which side of the hyperplane they are on.

Figure 1 illustrates how a separating hyperplane may be drawn in the case of two features and perfect separability.[1] The two classes of points are designated by the blue triangles and red circles. The goal of an SVM is to find the optimal separating hyperplane between the two classes of points that maximizes the margin between the classes' closest points.

In this example, the optimal separating hyperplane is designated by the solid line and the margins the dashed line. The points which fall on the boundaries are called support vectors, these are designated by a darker shade of red and blue. Having found this hyperplane for a given set of features enables us to then easily classify new observations based on their features.

[1] In real-world applications, however, training data are rarely perfectly separable. To deal with this we utilize soft-margin SVMs, which weight down the influence of data points that fall on the "wrong" side of the separating hyperplane.

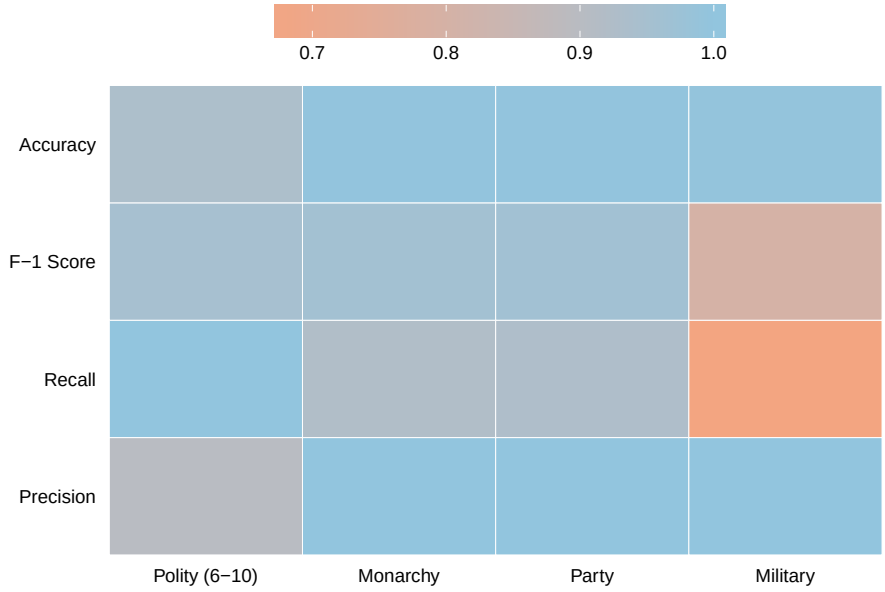
Figure 1. SVM Illustration



6. RESULTS

The heatmap in figure 2 summarizes the accuracy of the SVM classifier across our variables of interest in the out of sample test periods. Accuracy provides a measure for the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications. The accuracy statistic for polity is approximately 90%, while the SVMS on monarchy, party, and military classify over 99% of the cases correctly.

Figure 2. Aggregate Performance

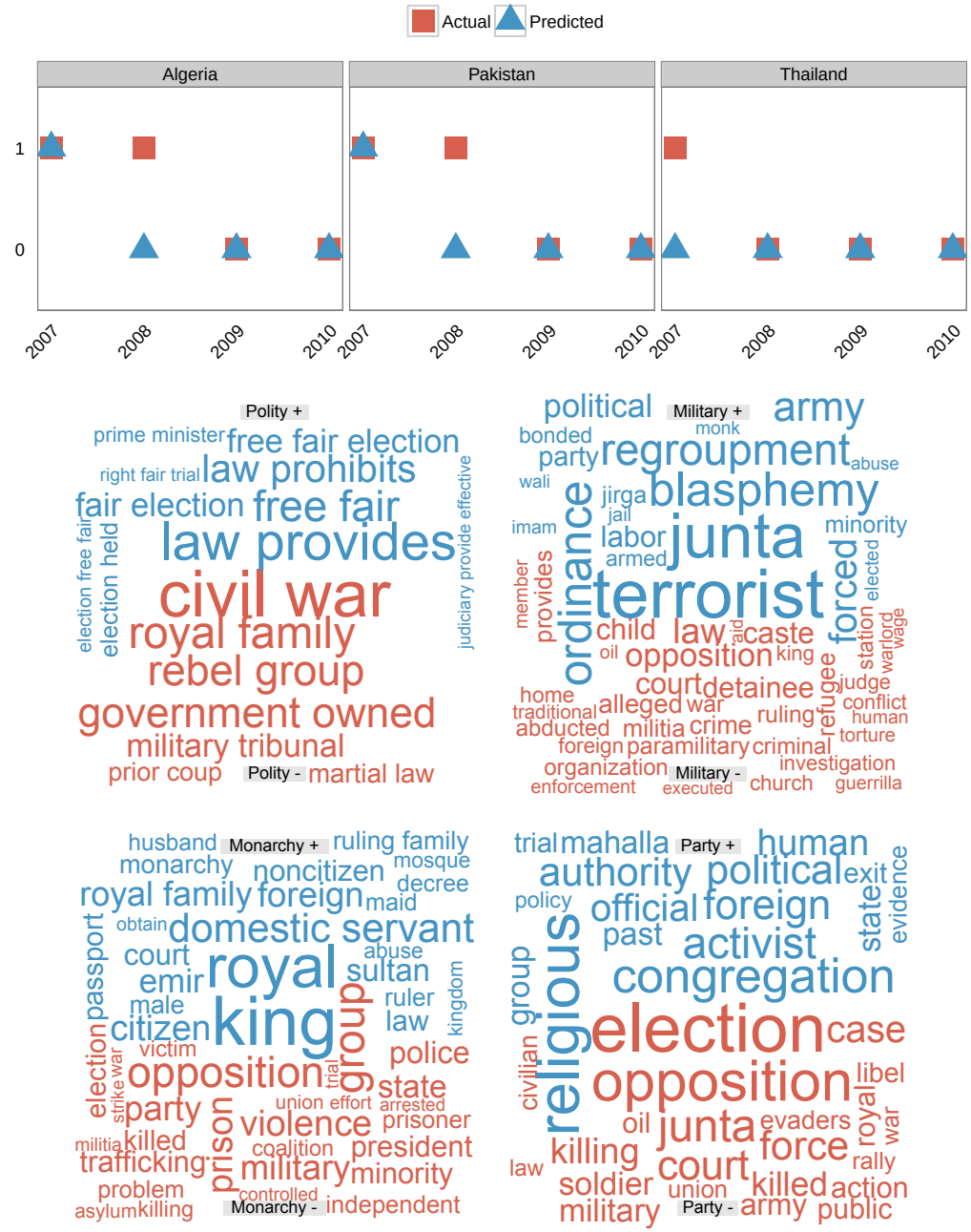


Separation plots are also useful for visualizing the discriminatory power of these models. The series of separation plots shown below in figures **??** and **??** reinforce the point evident in the accuracy statistics: the SVMs do an excellent job discriminating among cases in the test set.

Also track changes over time

** Then go into features that predict classes again using word clouds

Figure 3. Change in military ratings



** Discussion of what we are actually capturing here....

** Provide maps to give some context of how we are classifying countries and keep Jay's comments about it in there

If we compare these predicted probabilities to the observed values in the test set, we see that these classifiers are doing a very good job discriminating between 1s and 0s on the observed values for all four of the regime types. The column chart in Figure 2 summarizes the out-of-sample

accuracy of SVM and logistic regression for each of the four types. SVM consistently outperforms logistic regression and produces extremely good results overall, albeit with some variation across the four types. For the three authoritarian types, precision (the proportion of cases classified as positive that actually belong there) was better than recall (the proportion of positive cases that were correctly classified), but that pattern is reversed for democracy. Logistic regression does noticeably poorer for all four types than SVM, but still not terribly.

The histograms in Figures 1A and 1B, on the following two pages, summarize the distribution of the out-of-sample predicted probabilities from SVM and logistic regression, respectively, for each of the four regime types we have explored so far. The plots for the SVM estimates show sharp differentiation among cases, especially for the three authoritarian dimensions. Nearly all cases score very close to 0 or 1, producing bimodal distributions with steep peaks and few observations in between them. By contrast, the plots for the predicted probabilities from logistic regression are all unimodal, with a peak close to 0, maximum values much lower than 1, and thicker tails than the SVM output.

Figure 4. Histograms of Out-of-Sample Estimated Probabilities of Four Regime Types from Support Vector Machines (SVM)

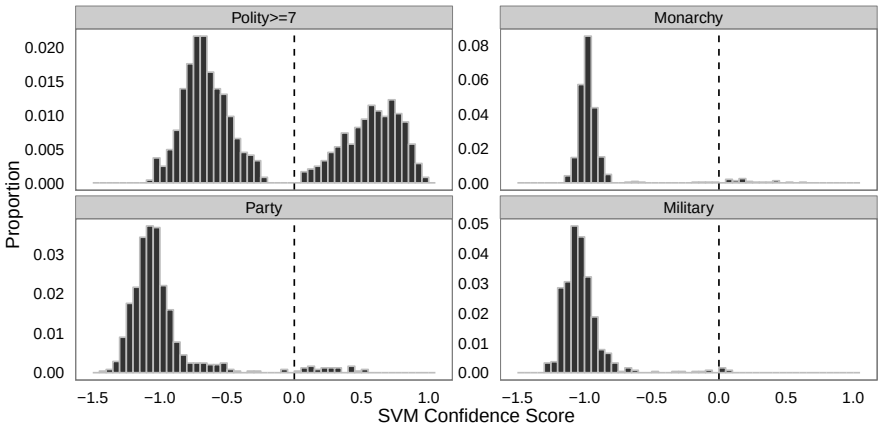
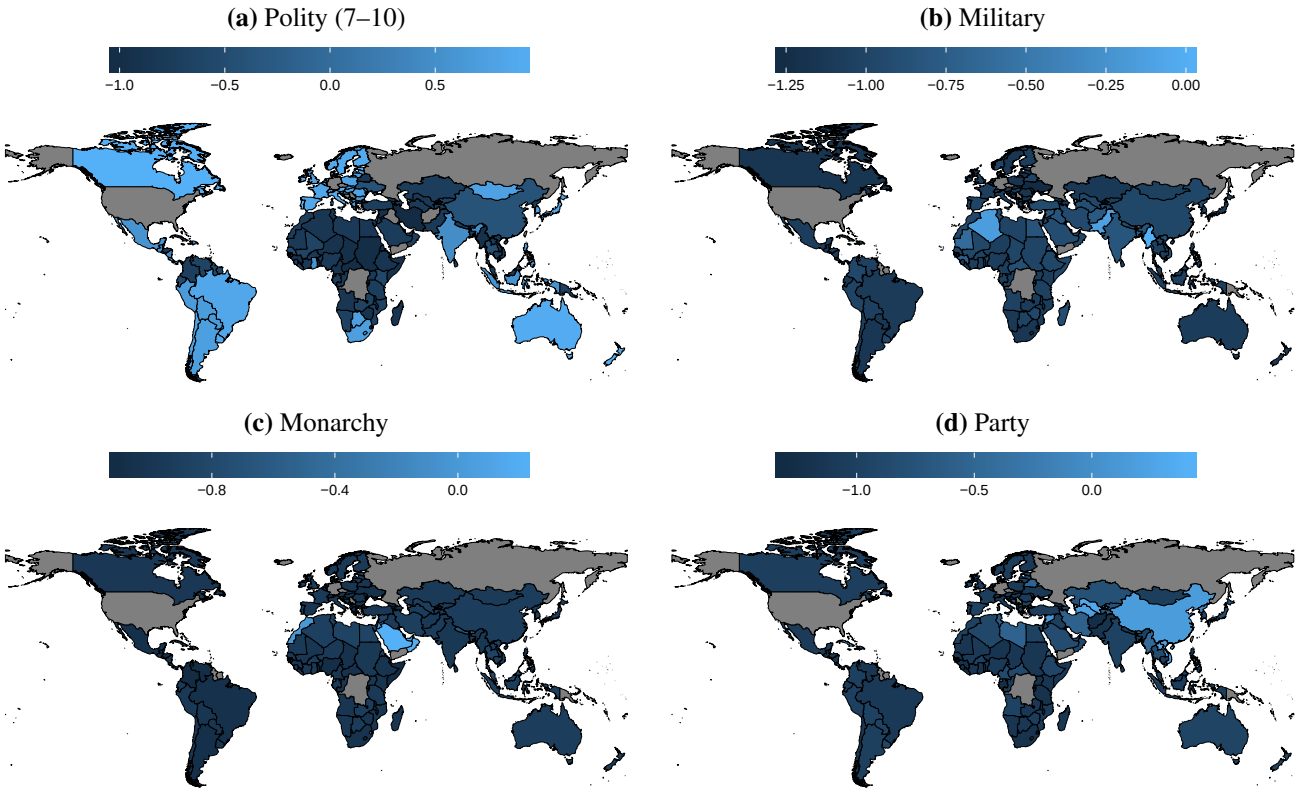


Figure 4, on the following two pages, uses heat maps to show variance in the SVM confidence scores from the most recent year available for all four regime dimensions: 2012 for democracy and 2010 for the three authoritarian forms. These confidence scores indicate the distance and direction of each case from the hyperplane used to classify the set of observations into 0s and 1s. The probabilities used in the earlier summaries are derived from them, but we choose to map the

Figure 5. Heat Maps of Out-of-Sample Confidence Scores from Support Vector Machines (SVM) of Four Regime Types.



raw confidence scores because they exhibit more variance than the probabilities and are therefore easier to visualize in this form.

On the whole, cases fall out as we would expect them to. The results seem valid on their face. For example, the democracy classifier confidently identifies Western Europe, Canada, Australia, and New Zealand as democracies; shows interesting variations in Eastern Europe and Latin America; and confidently identifies nearly all of the rest of the world as non-democracies (defined for this task as a Polity score of 10). Meanwhile, the monarchy model correctly identifies most of the countries of the Arabian Peninsula, Jordan, and Morocco as monarchies and the rest of the world as not. The military rule classifier sees Myanmar, Pakistan, and (more surprisingly) Algeria as likely cases in 2010, and is less certain about the absence of military rule in several West African and Middle Eastern countries than in the rest of the world. The map of results from the one-party rule classifier is probably the least informative because it is missing data for a few critical cases, including China and North Korea. Still, it seems to do well with the rest of the world,

including confident identification of one-party regimes in Syria, Vietnam, Laos, Turkmenistan, and Uzbekistan and midrange values for Iran, Kazakhstan, Libya, Eritrea, and Myanmar.

7. CONCLUSION

The results described here demonstrate that it is possible to generate measures of political regime type from publicly available texts at relatively low cost. An initial application using features derived from a small and corpus produced out-of-sample estimates that closely match observed data on several very different regime types.

The out-of-sample classification accuracy of our models is very good, but that doesn't mean we are satisfied with the output and ready to move to data production. We have two main concerns about the results of this initial implementation. First, the texts included in the relatively small corpus we have assembled so far only cover a narrow set of human-rights practices and political procedures. These practices and procedures are all fundamental to how we conceive and observe political regime type, but they aren't the only things that are.

To address this issue in future iterations, we plan to expand the corpus to include annual or occasional reports that discuss a broader range of features in each country's national politics. Eventually, we hope to add news stories to the mix. If we can develop models that perform well on an amalgamation of occasional reports and news stories, we will be able to implement this process in near-real time, constantly updating probabilistic measures of regime type for all countries of the world at very low cost.

Second, the stringent criteria we used to observe each regime type in constructing the binary indicators on which the classifiers are trained also appear to be shaping the results in undesirable ways. We started this project with a belief that membership in these regime categories is inherently fuzzy, and we are trying to build a process that uses text mining to estimate degrees of membership in those fuzzy sets. If set membership is inherently ambiguous in a fair number of cases, then our approximation of a membership function should be bimodal, but not too neatly so.

Most cases most of the time can be placed confidently at one end of the range of degrees of membership or the other, but there is considerable uncertainty at any moment in time about a non-trivial number of cases, and our estimates should reflect that fact.

If that's right, then our initial estimates are probably too tidy, and we suspect that the stringent operationalization of each regime type in the training data is partly to blame. To address this problem in future iterations, we hope to experiment with two alternatives. First, we may lower the bar for observed membership in each category in the training data. For example, we could lower the Polity threshold for observing democracy from 10 to 8 or 6. For the authoritarian regime types, we could include hybrid cases instead of restricting the training data to instances where both source data sets concur on pure types. For example, instead of only coding regimes as 1 for military rule when both GWF and HT agree that a case belongs in that set and only that set, we could code regimes as 1 on military rule when either source identifies them as having that feature, whether pure or hybrid.

Second and more ambitious, we could use Bayesian measurement error models to derive continuous measures of cases' degree of membership in each regime category from multiple data sets and then use those continuous measures as the targets in our supervised learning stage. This is the technique that Pemstein, Meserve and Melton (2010) use to derive the Unified Democracy Scores from scalar measures of regime type, but it can be generalized to situations in which the observations are categorical or binary as well. In the special case that we have no prior belief about regime type independent of the observed data and we believe the sources are equally reliable, we could simply use the unweighted average of the observed "votes." By making uncertainty about category membership more explicit in the training data, we should be able to develop models that are (properly) less confident in their readings of the relevant texts as well.

Those are the main ways we plan to try to improve the work described here, but they aren't the only ones. In addition to the machine learning and logistic regression models that we have already used, we plan to try supervised topic modeling via LDA as another classification strategy for all of the regime types. Last but not least, we also plan soon to add two other regime types

to the mix–personal rule and collapsed states–and then to replicate the process described here on them.

REFERENCES

Boix, Carles, Michael Miller and Sebastian Rosato. 2012. "A Complete Dataset of Political Regimes, 1800-2007." *Comparative Political Studies* 46(12):1523–1554.

Casper, Gretchen and Claudiu Tufis. 2003. "Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets." *Political Analysis* 11(2):196–203.

Chadefaux, Thomas. 2014. "The Triggers of War: Disentangling the Spark from the Powder Keg.". **URL:** *http://ssrn.com/abstract=2409005.*

Chang, Chih-Chung and Chih-Jen Lin. 2011. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2(3):27.

Cheibub, Jose Antonio, Jennifer Gandhi and James Vreeland. 2010. "Democracy and Dictatorship Revisited." *Public Choice* 143(1/2):67–101.

Coppedge, Michael, John Gerring, Staffan I. Lindberg and Jan Teorell. 2013. "Aggregating Varieties of Democracy Data.". http://ssrn.com/abstract=2303580.

D'Orazio, Vito, Steven T. Landis, Glenn Palmer and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22(2):224–242.

Escribá-Foch, Abel and Joseph Wright. 2010. "Dealing with Tyranny: International Sanctions and the Survival of Authoritarian Rulers." *International Studies Quarterly* 54(2):335–359.

Geddes, Barbara. 1999. "Authoritarian Breakdown: Empirical Test of a Game Theoretic Argument." Paper presented at the annual meeting of the American Political Science Association.

Geddes, Barbara. 2002. "The Effects of Foreign Pressures on the Collapse of Authoritarian Regimes." Paper presented at the Annual Meeting of the American Political Science Association, Boston, MA.

Geddes, Barbara, Joseph Wright and Erica Frantz. 2014. "Autocratic Breakdown and Regime Transitions: A New Data Set." *Perspectives on Politics* 12(2):313–331.

Goldstone, Jack A, Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder and Mark Woodward. 2010. "A global model for forecasting political instability." *American Journal of Political Science* 54(1):190–208.

Grimmer, Justin. 2010. "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19(1):32–47.

Hadenius, Axel and Jan Teorell. 2007. "Pathways from Authoritarianism." *Journal of Democracy* 18(1):143–15.

Hegre, Håvard. 2014. "Democracy and Armed Conflict." *Journal of Peace Research* 51(2):159–172.

King, Gary and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57:617–642.

Marshall, Monty G. and Keith Jaggers. 2002. "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2002, Dataset Users' Manual." Center for International Development and Conflict Management, University of Maryland.

O'Connor, Brendan, Brandon M. Stewart and Noah A Smith. 2013. Learning to Extract International Relations from Political Context. In *Conference Proceedings of the Association of Computational Linguistics*, ed. Hinrich Schuetze Hinrich Schuetze Hinrich Schuetze. Sofia, Bulgaria: p. tbd.

Pemstein, Daniel, Steven A. Meserve and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.

Platt, John C. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods.".

Slapin, Jonathan B. and Sven-Oliver Proksch. 2010. "Look Who's Talking: Parliamentary Debate in the European Union." *European Union Politics* 11(3):333–357.

Ulfelder, Jay. 2005. "Contentious Collective Action and the Breakdown of Authoritarian Regimes." *International Political Science Review* 26(3):311–334.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Wahman, Michael, Jan Teorell and Axel Hadenius. 2013. "Authoritarian Regime Types Revisited: Updated Data in Comparative Perspective." *Contemporary Politics* 19(1):19–34.

Wilson, Matthew. 2014. "A Discreet Critique of Discrete Regime Type Data." *Comparative Political Studies* 47(5):689–7.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology & Politics* 5(1):33–48.