An Evaluation of Measures of Textual Similarity

Zachary Elkins
zelkins@austin.utexas.edu

Robert Shaffer
rbshaffer@utexas.edu

September 30, 2017

Abstract

Understanding the similarity of two texts can be enlightening and useful. Unfortunately, human-generated similarity comparisons are expensive and labor-intensive. Supervised and semi-supervised automated approaches are more scalable, but also more difficult to interpret. We leverage a unique dataset of human-coded national constitutions in order to evaluate supervised and unsupervised similarity approaches. First, we assess the scores from a series of plausible unsupervised feature extraction and calculation approaches against a parallel set of human-coded similarity judgments. Next, we use the best-performing feature extraction approaches as inputs into a supervised scheme. We then assess the various automated measures in a set of applied criterion-validity tests. The applied tests illustrate both the utility and the practical limits of the human-machine correspondence. The best automated measures offer high-quality predictive and inferential properties, but limited unit-level descriptive performance.

Word count: 7559

1 Introduction

Measuring the similarity of cases is basic to scientific inquiry (Santini and Jain 1999; Tversky and Gati 1982). Sometimes similarity analysis represents an exploratory step; researchers frequently cluster, match, or categorize observations in order to probe their initial intuitions about a phenomenon of interest. Sometimes similarity is an end in itself. For example, a scholar of electoral campaigns might compare statements made by various Presidential candidates in order to understand the proximity of their agendas. Or, a comparative politics researcher might compare the similarity of national laws in order to study the diffusion of ideas across time and space. Many other inspiring applications abound (see e.g., Strehl et al. (2000); Grimmer (2010); Grimmer and King (2011); Ahlquist and Breunig (2012); Roberts et al. (2015); Purpura and Hillard (2006); Hillard et al. (2008) for discussion and applied examples).

All of the examples given above involve comparison of *textual* information, a form of data that is our specific focus. Much of the raw data on institutional and political phenomena is lodged in texts, such as laws, party platforms, speeches, and advertising materials. As computing resources and data availability have expanded, modeling approaches designed to extract information from these data sources have proliferated.[1] However, it is not always clear how information extracted by such techniques relates to human-generated constructs, particularly in the unsupervised setting. For one, textual similarity is multi-dimensional, and may vary according to content, semantics, style, sentiment, or some combination thereof. Moreover, even with the correct set of features, the functional form relating those features to human-interpretable constructs is usually difficult to define. Supervised machine learning techniques can help ease this challenge, but defining the relevant features and gathering human-generated training data is laborious and offers an unclear payoff. How much training is enough, and at

[1]We will alternately call such methods of content analysis "computational," "automated," or "machine," to distinguish them from "human" approaches, in which the analyzed data result from human interpretations of the text.

what cost?

To explore this set of measurement challenges, we turn to a set of original hand-coded data on the content of national constitutions. Because of their coverage across cases and topics and the authors' close attention to the written text, these data make for a unique reference point against which to evaluate automated measures. Using these data, we develop a baseline similarity approach, *inventory similarity*, that parallels the kind of automated similarity measures that one might compute with large, high-dimensional texts. Following this approach, we generate a set of *inventory similarity* scores with the human-interpreted constitutions data and a corresponding (and varying) set of computational measures derived from the texts themselves. The computational measures represent both unsupervised and supervised approaches, the latter at varying levels of supervision.

We evaluate the validity of the similarity values we produce using three tests. First, we examine the aggregate correspondence between our human-generated target and various machine-generated referent values. We find that some computational approaches predict the criterion measure appreciably better than do others and the best approaches seem to predict the human scores remarkably well ($r \approx 0.7$). Second, we incorporate the human- and machine-generated measures in a set of regression models. We find that analyses produced with both sets of similarity values produce similar causal inferences. Third, we test whether the various measures return the same unit-level descriptive inferences regarding rank-order similarity comparisons. This last test reveals substantial differences across the measures, which reminds us of the substantive limits of their correspondence.

2 Similarity Amongst Constitutions

2.1 Why Measure Similarity?

Both quantitative and qualitative projects often rely on similarity comparisons between cases at some point in the research process. For a motivating example, consider electoral campaigns. As a campaign progresses, we might expect candidates either to cluster or differentiate their messaging, with clustering patterns shifting as the candidates' respective electoral prospects and issue priorities change. We might further expect this pattern to be particularly volatile during a campaign such as the 2016 Republican Primary, which featured a relatively large and open field with no strong initial frontrunner and no clear party-backed candidate. A researcher studying political communication might therefore be interested in searching for points of convergence or divergence in attention paid to campaign issues by the various candidates (Sides 2006; Savoy 2010; Sulkin 2005; Klebanov et al. 2008), or for similarity in rhetoric used by the various campaigns at different points in time (Hart 2009).[2] We view such analyses as potentially enlightening and even pathbreaking, but our optimism is tempered by our uncertainty about measurement error and interpretability.

Unfortunately, opportunities to validate these kinds of automated procedures are rare. In order to study clustering patterns within a set of cases, such as campaign communications or legal contracts, a researcher would need a vector of data on each case, but she would also need a procedure by which to compare and quantify the similarity between those cases. In the computer science and statistical settings, measurement tasks of this sort are generally conducted on short excerpts using an abstract notion of similarity. For example, a researcher might ask subjects to rank pairs of

[2]Researchers have conducted similar analyses in a variety of more or less esoteric settings, including rhetorical comparisons used by Al-Qaeda leaders (Pennebaker et al. 2008) and members of the Beatles (Petrie et al. 2008). Usage comparisons have also been used in more general problem settings, such as unknown authorship problems (Mosteller and Wallace 1963; Argamon et al. 2009; Stamatatos 2009).

words or short sentences based on some undefined notion of similarity, and then attempt to learn a function that replicates their judgments in new cases. By contrast, in the political science context researchers are often more interested in similarity comparisons based on a more specific conceptualization and applied to longer texts. Gathering training data for similarity comparisons between long documents requires human coders to read large quantities of text and judge their similarity (or code their attributes) based on a detailed conceptualization scheme. To our knowledge, no existing study has obtained the necessary training data to conduct this kind of validation exercise.

2.2 The Domain of Inquiry: National "Constitutions"

To explore and compare measures of similarity, we draw on the study of national constitutions, surely one of the more central set of texts in political science. Scholars interested in diffusion of ideas have long studied the intellectual history of these texts, and their patterns of change across space and time. One hears such claims even about esoteric texts, such as the following correction to the record regarding the constitutions of two confederations in ancient Greece:

> It has been suggested that the Arcadian confederate constitution drew on the Boeotian equivalent, but there is in fact little reason to think that the Arcadian constitution was heavily influenced by Boetia (Brock and Hodkinson 2002).

Discussions like these depend primarily on pairwise similarity evaluations of various constitutional texts. Scholars of comparative political institutions deploy these comparisons to evaluate hypotheses related to the origin, spread, and novelty of ideas across jurisdictions. However, as in the campaign examples cited above, generating a valid and interpretable measure of similarity is not straightforward. With respect to ingredients, at least, time-series cross-national data on the content of constitutions

(now recently available) is one promising source. Unfortunately, such data are expensive to produce and represent, ultimately, an idiosyncratic interpretation of political and legal documents. By contrast, recent developments in machine-coded content analysis suggest the possibility of more efficient, and perhaps equally reliable, measures of content similarity. Apart from their competitive advantages, machine-coded measures would also seem to deliver complementary contributions: specifically, such methods may deliver different – perhaps even unexpected – interpretations of constitutional text.

We leverage original data from the Comparative Constitutions Project (CCP), two attributes of which are particularly convenient for the analyses herein. First, the CCP's authors measure aspects of the constitutional *text* itself, not a broader understanding of the "constitutional order," which may include other elements of higher law such as important ordinary laws, judicial interpretations, or norms. This focus on *written* constitutions, and not broader constitutional understandings, makes for a more comparable reference point for automated, text-based measures of similarity. Second, the CCP's scope is extensive. The CCP collects and content-tags constitutional texts for some 600 topics for all founding documents written since 1789. This broad coverage allows analysts to capture and trace the flow of constitutional ideas across countries and time and assess the consequences of different constitutional choices.

2.3 The Target: Inventory Similarity

To generate similarity values from the CCP data, we employ an approach we term *inventory similarity*. As with most human-generated constructs, a measure of similarity is useful only to the extent that it illuminates some underlying concept, and there are many such concepts on which we could choose to focus. For example, we could measure the extent to which two constitutions spend the same proportion of their

verbiage on rights or structural provisions, or the degree to which their substantive treatments of these topics are similar. But we can also ask, more simply, whether two constitutions even *address* the same issues. Here, we focus on the latter idea. If two constitutions address a similar "inventory" of issues, that suggests that the authors of each document were sufficiently interested in those problems to address them in their founding documents. Though each document might address these ideas differently, a shared conceptual inventory suggests a shared set of underlying concerns and challenges.

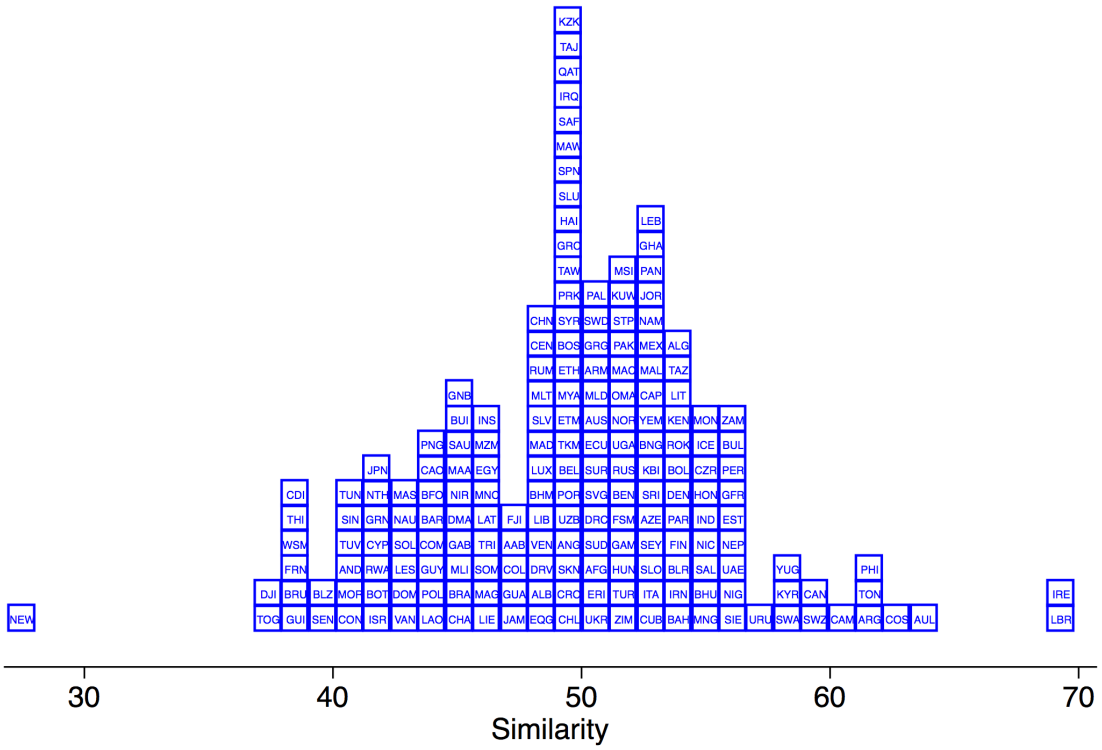
In the constitutional context, the CCP data make inventory similarity scores relatively straightforward to construct. CCP contains extensive information on the inventory of topics included (or not included) in any two constitutions (e.g., does the constitution specify the method of selection for the head of government, mention a central bank, address the accession of new territory, etc.). We select 70 such topics from the CCP along which we can calculate such a measure.[3] We exclude from this list of items many *sub*-topic questions that should be understood as making rather refined distinctions between constitutions (e.g., whether the constitution specifies the selection and removal process for the head of the central bank (excluded) as opposed to whether the constitution specifies a central bank (included)). We also exclude topics that are either highly rare or highly consensual, under the assumption that such low-variance items will be of less informational value.

Thus, our final dataset consists of a set of binary variables indicating whether a given topic is present or not in a given constitution. Using these data, we calculate Jaccard's (1912) similarity coefficient[4] for each unique dyad ($n = 18,528$) across the

[3]Elkins et al. (2009) use the same set of topics as a measure of *scope*, a related concept.

[4]Jaccard similarity is also closely related to mutual information. Defining $I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log(\frac{p(x,y)}{p(x)p(y)}$ as the mutual information of two discrete random variables X, Y and $H(X,Y)$ as their joint entropy, $\delta_J(X,Y) = \frac{I(X,Y)}{H(X,Y)}$

Figure 1: Similarity to the U.S. Constitution (hand-coded, across topics)



full dataset of 193 constitutions, defined as:

$$\delta_J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

The scores across these dyads have a mean of 0.60 (s.d. = 0.10) and range from 0.19 to 0.96.

For those mired in the genre of written constitutions, the similarities exhibit some face validity. Among the most similar pairs are the constitutions of Oman and Qatar (0.90), Armenia and Slovakia (0.90), Serbia and Montenegro (0.91) – pairs that would seem like likely kindred spirits since they were produced in the same parts of the world. Some of the least similar include Brunei and Austria (0.21) and New Zealand and Indonesia (0.24), which upon inspection do look markedly different in their content.

The U.S. Constitution may be more widely known (at least compared to Brunei). In Figure 1, we plot the distribution of similarities scores to the U.S., and identify particular cases – all to add insight on the validity of the measure. We might expect the U.S. to be most similar to those constitutions of its generation, particularly those in Latin America, which are thought to have drawn inspiration from the Madisonian creation. Argentina and Costa Rica are two constitutions in the top five with respect to similarity to the United States. The most similar constitutions to that of the United States are Ireland's and Liberia's. Of course, Liberia was famously founded by ex-slaves from the United States, and is commonly thought to have a similar constitutional structure. In short, the human-generated measure of similarity seems to have face validity, though there are interesting variations in similarity well-worth investigating (which is, indeed, the point of constructing the measure).

2.4 The Referent: Textual Similarity

Text-based feature extraction and dimensionality reduction tools have received substantial attention in the statistics and computer science literature over the last several decades, and have been applied extensively in political science (Grimmer and King 2011; Lucas et al. 2015). Most approaches in this domain work from some variant of a "bag-of-words" approach, in which features are extracted from a vector space presentation of a corpus of interest (usually, a term-document matrix).

Earlier versions of these models were estimated via linear algebra techniques, such as singular value decomposition or principal components analysis (e.g. Latent Semantic Indexing (LSI), as described in Dumais et al. (1988); Deerwester et al. (1990)). More recent approaches, by contrast, have usually generative.[5] Latent Dirichlet Al-

[5]Here, the term "generative" refers to models which specify a joint probability distribution over observed and hidden variables, such that new data can be straightforwardly simulated from the model. For example, LDA is "generative" in the sense that, given a vector of hypothetical topic proportions, we can combine those topic proportions with a fit model to easily simulate a new document. Models like principal components analysis, which use linear algebra techniques to reduce

location (LDA) (Blei et al. 2003; Blei 2012), for example, consists of a Dirichlet-multinomial mixture model, in which documents are represented as a distribution over latent "topics" and "topics" consist of a distribution over words. Subsequent work has expanded on this basic approach in a variety of ways, incorporating time (Blei and Lafferty 2006), authorship (Grimmer 2010), covariates (Roberts et al. 2014), variable dimensionality (Blei and Jordan 2004; Teh et al. 2012) and many other possibilities besides. Some authors (e.g. Mikolov et al. (2013)) have also proposed deep-learning approaches that generate a vector representation of *words* rather than *documents*, which allow for algebraic operations on individual tokens rather than larger texts.

For the purposes of this study, we focus on features generated from four of the most prominent of these models: specifically, LSI, LDA (as implemented in McCallum (2002)), Roberts et al.'s (2014) Structural Topic Model (STM), and Mikolav et al.'s (2013) word2vec model (see Table 1 for details). For each of these approaches, we estimate models based on a variety of topic values, and report results for each model. We also include similarity scores generated from a term frequency-inverse document frequency (TF-IDF)-weighted word count vectors as a baseline point of comparison, which is standard in many natural language processing studies.

Details of parameter settings and pre-processing steps are given in Table 1. For each model we fit, we divide each constitutional text into a number of constituent documents (sentences for word2vec, and articles for all other models). Next, we preprocess the documents in the subdivided datasets. As Denny and Spirling (2016) demonstrate, pre-processing choices in unsupervised text analysis settings can have a substantial effect on downstream model performance. The pre-processing choices we present in this paper are intended to ease computational complexity while discarding as little information as possible. For example, in the broader text analysis literature, dropping non-alphabetical characters, stemming, and dropping words contained in

data dimensionality, do not have this property.

Table 1: Estimation and pre-processing details for feature set under consideration.

Model	Unit	Pre-processing	Hyperparameters
TF-IDF	articles[a]	(1) lower-case; (2) punctuation, stopwords, tokens ≤ 3 characters, tokens in ≤ 10 documents removed; (3) documents ≤ 5 tokens removed	n/a
LSI	articles[a]	Same as TF-IDF	$\{50, 100, 150, 200\}$ topics
LDA	articles[a]	Same as TF-IDF	$\{50, 100, 150, 200\}$ topics; asymmetric alpha prior
STM	articles[a]	Same as TF-IDF	$\{50, 100, 150, 200\}$ topics; constitution and date as covariates
word2vec	sentences[b]	(1) lower-case; (2) punctuation, tokens ≤ 1 character removed	{200, 400, 600, 800}-length feature vector

Where not specified, all parameters left at default settings. LSI and TF-IDF estimated via Gensim (Řehůřek and Sojka 2010). LDA estimated via MALLET (McCallum 2002). STM estimated via the STM R package (Roberts et al. 2014). Asymmetric alpha prior, as described by Wallach et al. (2009).

[a] $n \approx 138000$
[b] $n \approx 201000$

fewer than 0.5-1% of all documents in the dataset are typical pre-processing steps (see, e.g., Grimmer and King (2011); Denny and Spirling (2016)). Here, in our topic modeling specifications we retain all characters, we do not stem terms, and we retain all words contained in at least 10 documents ($\approx 0.01\%$ of the dataset).

As shown in Table 1, the preprocessing standards we use for our word2vec models differ in two respects from the other approaches we present. In particular, for our word2vec specifications we subdivide constitutions into sentences instead of articles, and we retain all words two characters or longer. We adopt this differing specifica-

tion for two reasons. First, word2vec is substantially less demanding to estimate than most of the other approaches we present (particularly STM), and requires fewer preprocessing steps in order to become computationally feasible. Second, these differing standards offer our results some robustness against preprocessing choices. As shown in the following section, the word2vec-based similarity values we generate perform as well as those produced using our topic modeling specifications. We view this finding as suggestive (though not conclusive) evidence that our results are robust to the range of preprocessing standards we test in this paper.

To generate final similarity values, we re-combine all articles/sentence feature vectors extracted by each model into a set of constitution-level feature vectors. Specifically, each constitution-level feature vector P is defined as

$$P_{jk} = \frac{1}{\sum_{i=1}^{N_j} n_{ij}} \sum_{i=1}^{N_j} n_{ij} p_{ijk}$$

Where i indexes the N_j article/sentence-level feature vectors associated with the j^{th} constitution, and k indexes features. Within each constitution, n_{ij} represents the word count of the i^{th} article/sentence associated with the j^{th} constitution, and p_{ijk} gives the feature value of the k^{th} element of the i^{th} article/sentence-level feature vector within the j^{th} constitution. P_{jk} thus represents an average feature vector for each constitution, weighted by the word count of each sub-document composing that constitution.

For LDA and STM, since the relevant feature vectors are constrained to lie on the $(k-1)$-simplex, we calculate a discretized Hellinger distance between each aggregated

feature vector, defined as

$$\delta_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{P_i} - \sqrt{Q_i})^2}$$
$$= \frac{1}{\sqrt{2}}||\sqrt{P} - \sqrt{Q}||_2$$

Feature vectors generated using LSI, word2vec, and TF-IDF are not constrained in this fashion. As a result, we use cosine similarity in these cases instead, defined as

$$\delta_C(P,Q) = 1 - \frac{P \cdot Q}{||P||_2||Q||_2}.$$

We emphasize that these preprocessing steps, parameter settings, and models are not the only plausible feature extraction/similarity approaches. However, for practical reasons, we clearly cannot test all available specifications. Instead, we argue that the choices we present here represent reasonable selection of plausible approaches for applied researchers, which can be adjusted based on the particular problem domain in question.

3 Validating Similarity

Our validation proceeds in three phases, which (we argue) illuminate different aspects of *criterion validity*. Criterion-valid measures are those that correspond closely to a criterion measure, which may be an outcome directly related to the concept or a "gold-standard" measure that is viewed to be an especially reliable and valid measure of a concept. We clarify our process, since the classification and labeling of validation tests are not perfectly standardized (see Adcock and Collier 2001). By way of demonstration, we also suggest our approach as one plausible, and generalizable, method of assessing criterion validity.

In our first test, we conduct a simple (and aggregate) analysis of the covariance between the target and referent measures. Our second test continues the criterion validity logic, although in a more applied and more exacting manner. That is, we test whether models applied to the target and referent measure produce the same inferences on a series of causal questions. Finally, in our third test, we ask whether the referent measure returns the same rank ordering of cases (and even the same ratio-level positions of cases), as does the target measure. In contrast to the previous two tests, this third study involves individual-level instead of group-level comparisons, and is (in our view) the most exacting of all. Taken together, one can think of these last two tests as examples of validation in the "pragmatic" tradition, to follow Collier and Seawright's (2014) typology. Pragmatic approaches, in this sense, are those that validate measures based on their performance in applied settings.

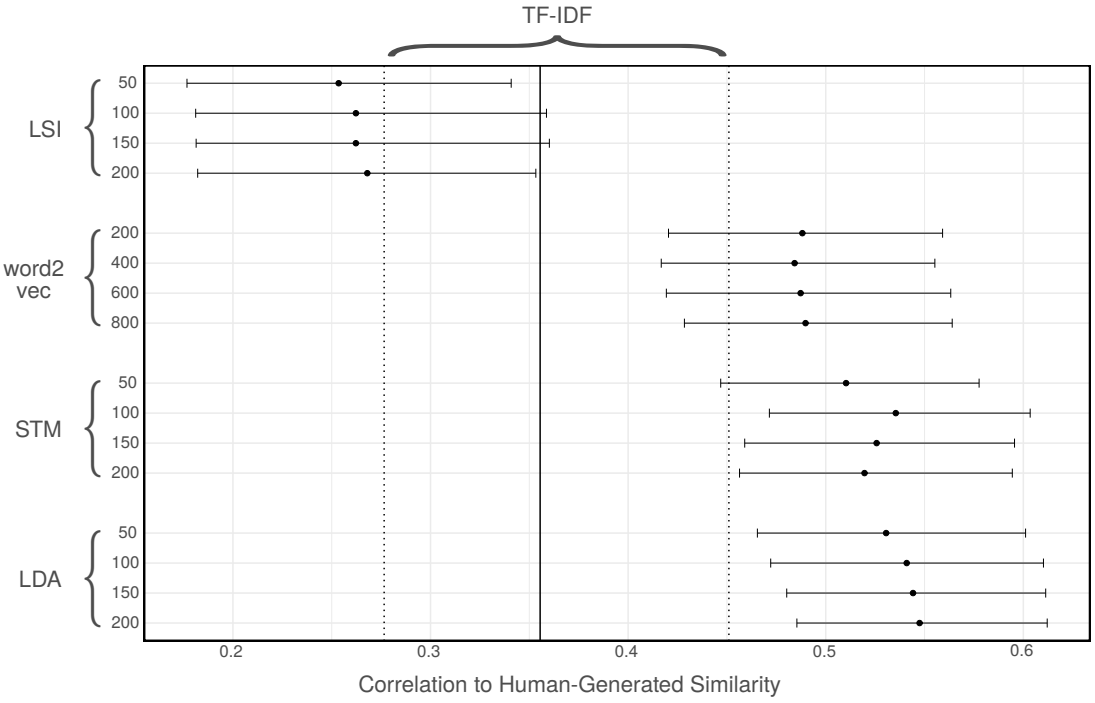
3.1 Test 1: Association between Target and Criterion Measures

To what degree do the machine measures of similarity covary with the human measure? We begin with an evaluation of the unsupervised approaches (see Figure 2). Of the five models under consideration, the three generative approaches (word2vec, STM, and LDA) all perform at a roughly similar level, and significantly better than both the baseline TF-IDF similarity scores and the non-generative alternative (LSI). Similarities created using the three generative models all correlate with hand-coded data in the $r = (.45, .6)$ range, with scores based on higher-dimensional LDA models being the best performers. From our perspective, these numbers suggest a strong association between these particular automated content measures and their human-coded counterpart.

Interestingly, the addition of covariates via STM does not seem to improve performance. For the purposes of this study, we were interested in conducting a fairly

generalizable test of the models under consideration, and thus included few covariates in our estimation approach (specifically, the date of the document's enactment and the constitution from which a given training paragraph was drawn). More expansive covariate sets might improve performance on other corpora; however, in our application, straightforward LDA appears to perform well.

Figure 2: Correlations between machine- and human-generated similarity values



Solid and dashed vertical lines indicate mean correlation and 95% confidence interval between CCP targets and similarities generated using baseline TF-IDF features. Dots and solid horizontal lines indicate mean correlation and 95% confidence interval between similarities produced using other feature extraction approaches. All confidence intervals estimated using critical values from a quadratic assignment procedure (QAP) null hypothesis as described in Dekker et al. (2003).

Importantly, across all of these models, the choice of the number of dimensions appears to have little impact on results. This invariance is heartening. In many modeling settings, dimensionality parameters (such as the number of topics in a topic model) represent a troubling aspect of the research process, with few generally-applicable guidelines or standards. Thankfully, for similarity comparison purposes, this choice does not appear to be particularly consequential.

Though encouraging, the correspondence between machine and human in these initial analyses leaves appreciable room for improvement. We therefore extend our initial approach to a supervised learning setting. In particular, we use feature sets using the LDA and STM models described above as input data for a random forest (Breiman 2001), which we then trained on varying proportions of the CCP data.[6] As before, we emphasize that this is not the only approach one might consider in this context. However, since it is impossible to examine all conceivable specifications, we argue that the approach we take offers a plausible reference point for applied work.

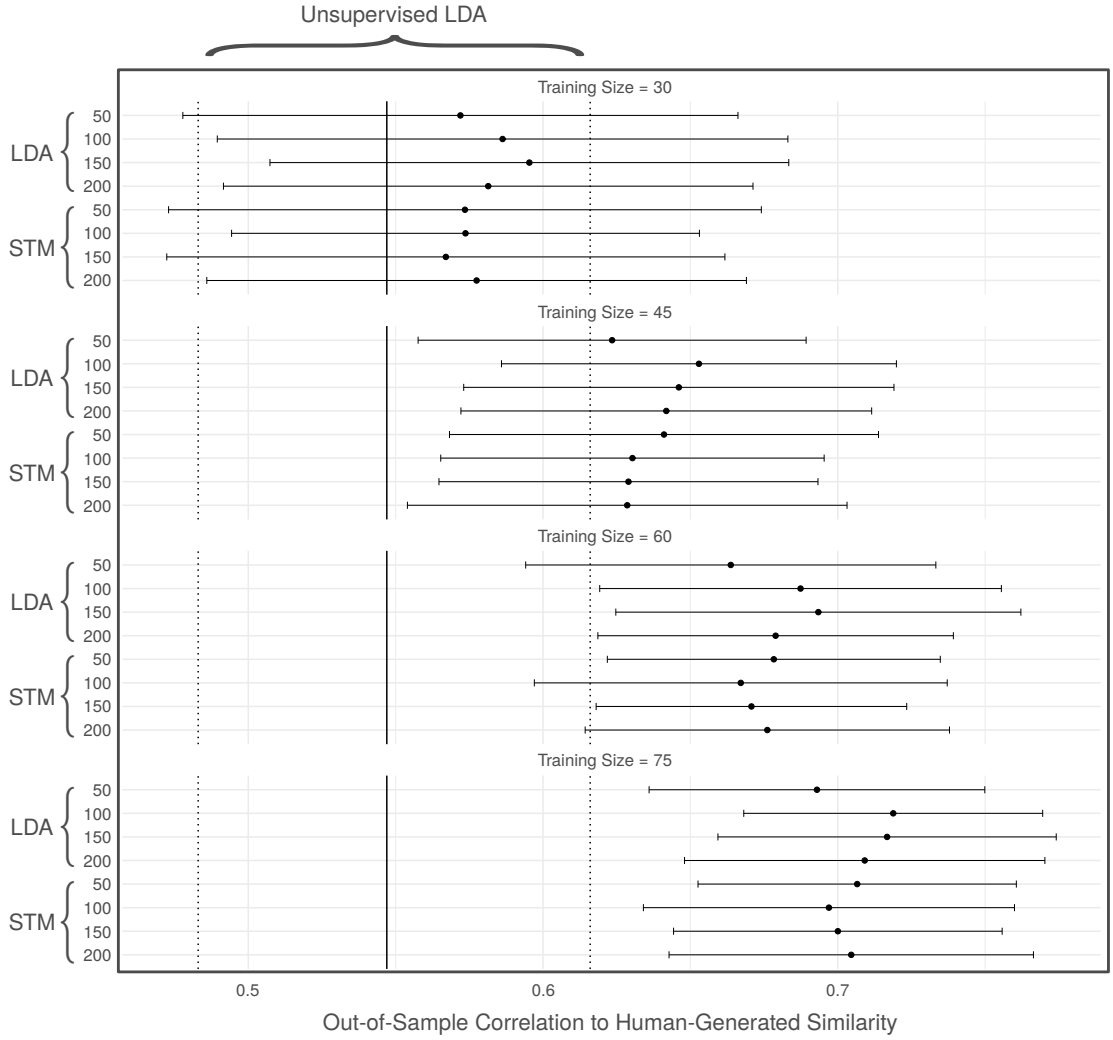
As shown in Figure 3, moving to a supervised alternative offers substantial improvements to predictive accuracy compared with the unsupervised alternative (using LDA_200 features as the point of comparison). With a training set as small as 45 documents ($\approx 25\%$ of the dataset), the supervised predictions consistently correlate more highly with the human-generated similarity measures than the unsupervised comparison shown above. By 75 documents ($\approx 40\%$ of the dataset), the improvements are striking; at that training set size, the supervised predictions correlate at $r \approx 0.7$ with out-of-sample human-coded data, versus $r \approx 0.55$ in the unsupervised comparison. As in the previous section, the choice of model and dimensionality parameter does not appear to make a substantial difference for model performance. Though some small differences remain between the various models, variation based on model specification is overwhelmed by variation generated through selection of the training set.

3.2 Test 2: Causal Inference across Target and Criterion Measures

For enthusiasts of automated content analysis, it will be encouraging to note what seem to be high levels of correspondence between the human measure and at least

[6]Number of candidate variables at each split selected by optimizing out-of-bag error separately for each training set, with 500 trees grown in each model. Training sets were selected by randomly sampling a set of countries, and using all dyads within that set as training examples.

Figure 3: Out-of-sample correlation between predicted similarities generated through a supervised procedure



Solid lines represent ±2 sample standard deviations, estimated from 100 train/test splits. Solid and dashed lines give mean and 95% confidence intervals from the unsupervised LDA_200 model shown in Figure 2 as a baseline comparison.

some of the machine measures. However, a more meaningful test might be one that that evaluates the measures in the context of a set of applied research questions. In that context, the question is whether the machine measure will yield the same causal inferences as would the human measures.

Consider, in this spirit, some basic expectations regarding *isomorphism* in constitutional design. A robust finding among those who have analyzed the CCP constitutions data (e.g. Elkins et al. (2013)) is that the drafter's context – in particular geography and era – matters enormously (call these the "contextual" hypotheses for the sake of reference). Some of these analyses suggest that we can explain as much as half of the variation in a constitution if we know *where* and *when* it was written (Cheibub et al. 2011). We test the contextual hypotheses with a set of regression models that predict similarity across the sample of 18,528 constitutional pairs. The relevant question is whether the relationships between these predictors and pairwise constitutional similarity are consistent across two operationalizations of the dependent variable: (1) a human and (2) a machine measure of similarity.

To test these hypotheses, we include two sets of predictor variables:

1. *Difference in years of enactment*, calculated as the square root of the absolute difference in the years in which the two constitutions in a given dyad were first enacted. Since constitutions written during a similar time period are likely to reflect similar underlying concerns and trends in constitution-writing, we expect the coefficient associated with this term to be negative.

2. *Same region.* A set of dummy variables that equal 1 if a dyad includes constitutions from the same region, for each of the eight geographic regions: East Asia, Eastern Europe, Latin America, Middle East/North Africa, Oceania, South Asia, Sub-Saharan Africa, and Western Europe/North America. We expect each of these coefficients to be positive, since constitutions drawn from the same region are likely to be more similar than those drawn from different regions (the

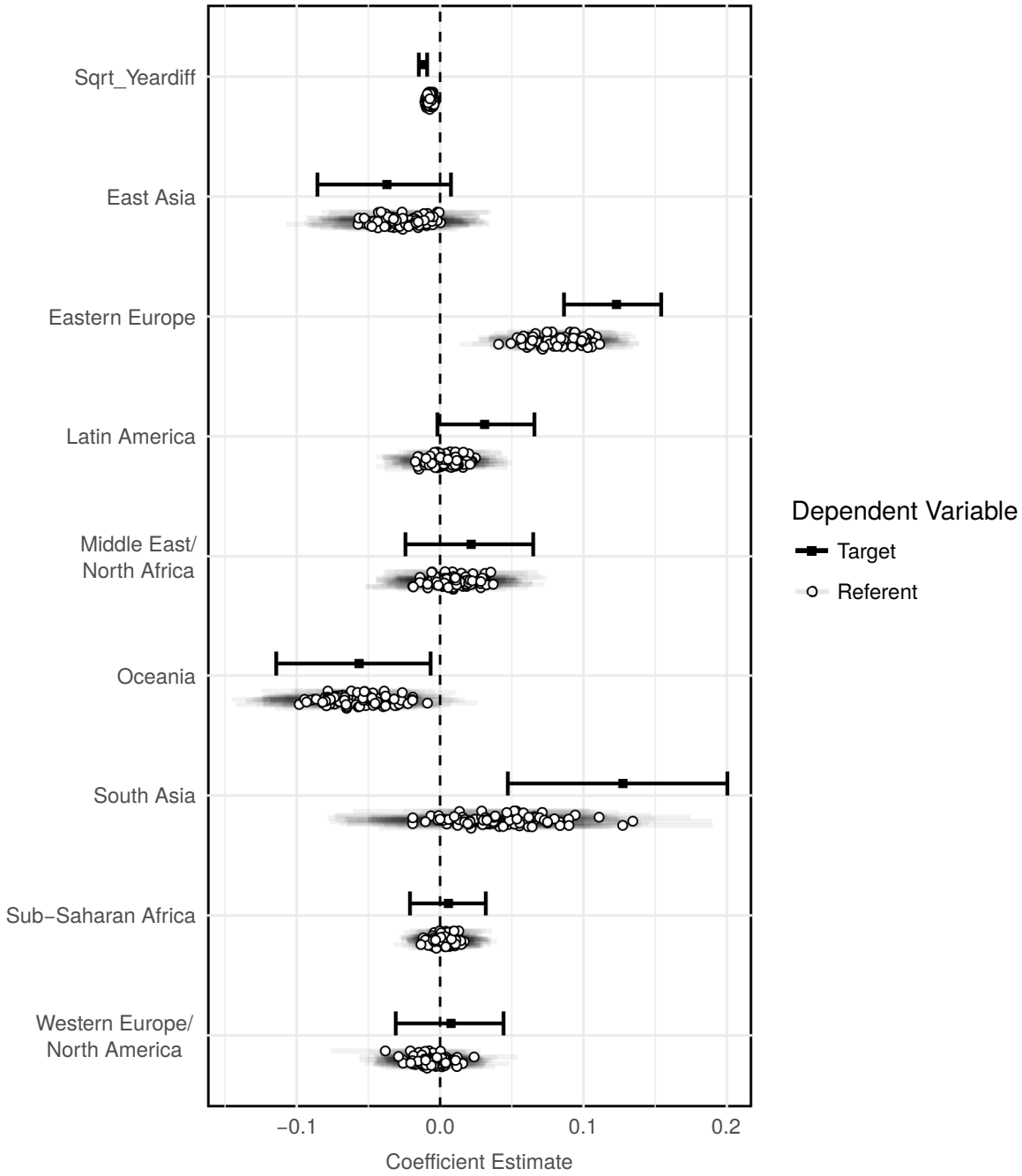
left-out category). However, there is likely to be substantial variation within these categories. For example, we expect constitutions from Oceania to cluster least, since countries in this region share relatively few cultural similarities. By contrast, we expect constitutions from Latin America and Eastern Europe to be more similar than the baseline, since these regions are more culturally homogenous.

Since similarity data are dyadic in nature, standard distributional assumptions for OLS coefficient estimates are inappropriate. Two dyads that share a country are likely to possess positively autocorrelated disturbances, producing artificially narrow confidence intervals for coefficient estimates. To address this issue, we use the permutation-based quadratic assignment procedure (QAP) correction described by Krackardt (1987) and extended by Dekker et al. (2003). Other such corrections are possible; for example, a researcher interested in studying the endogenous characteristics of the similarity network - rather than treating them as nuisance parameters, as in the QAP strategy - might employ a generalized exponential random graph model (see Cranmer et al. (2017) for further discussion). However, since the hypotheses presented in this section focus on coefficient values rather than unmodeled network characteristics, a regression-based approach is reasonable.

Figure 4 reports the results of a linear model estimated by OLS (with QAP-corrected confidence intervals) in which we predict similarity using the variables described above (See Appendix A for numerical coefficient estimates). The dependent variables in these models are the human-generated target values (as obtained from CCP) and the machine-generated referent produced using a random forest model estimated on LDA_200 features with 75 training documents. For the machine-generated referent values, we fit a separate linear model for each of 100 train/test splits, and plotted the estimated coefficients for each model.[7]

[7]For dyads contained in the training set, we substituted human-generated similarity values for

Figure 4: Linear model coefficient estimates for both human and machine-generated similarity values



Critical values, confidence intervals, and p-values for each coefficient produced using a QAP null hypothesis as described in Dekker et al. (2003). Coefficient estimates and confidence intervals drawn from models estimated using machine-generated referent values are overlaid and jittered, and confidence intervals are faded. Intercept omitted for readability.

Beginning with the model estimated on human-coded target values, our results provide some compelling evidence for the contextual hypotheses. For one, a generational effect is readily apparent: a pair of constitutions written in the same year are predicted to be 8 points more similar than a pair written 50 years apart. Geographic location also matters, but the effect varies substantially across regions. Eastern European constitutions exhibit a high degree of clustering, and are estimated to be approximately 12 points more similar than a pair of constitutions drawn from different regions. South Asian constitutions also cluster more than the baseline ($b = 0.127$). The Latin American coefficient is estimated to be positive as well, though it misses conventional cutoffs for statistical significance ($p = 0.069$). Contrary to expectations, however, constitutions drawn from the same region are not always more similar than those that are drawn from different regions. In particular, constitutions from the Oceania region, actually cluster *less* than the baseline (b=-0.056), suggesting that Oceania constitutions are more dissimilar to one another than a random pair of constitutions drawn from different regions. The East Asia coefficient is also estimated to be negative, though the coefficient estimate does not reach conventional significance thresholds ($p = 0.13$).

Again, our focal question in this section is whether we would reach the same conclusions using the machine-generated referent as we would with the human-generated target values. Returning to Figure 4 suggests that the answer to this question is a qualified yes. As in the previous section, the results of the various machine-generated similarity values suggest that the results vary substantially depending on the particular training set selected. Nevertheless, most model replicates returned the same conclusions regarding coefficient significance and sign as the model trained on the target

in-sample random forest predictions. Our rationale for this choice was drawn from the applied hypothetical that our validity tests are intended to approximate. If gold-standard values are available for some subset of observations during the supervised learning step, this same training should also be available during inferential modeling. During this phase of the analysis, substituting predicted values for these gold-standard measurements introduces noise into some observations on the dependent variable, unnecessarily discarding information.

values. Using $p = 0.05$ as a significance threshold, 88% of the non-intercept coefficient estimates across the 100 machine-generated model replicates returned the same conclusions regarding statistical significance and sign as did the human-generated coefficient estimates.[8] Performance was not uniform across all coefficient estimates; for example, all replicates returned a negative and significant estimate for the year-difference coefficient, and a positive and significant estimate for the Eastern Europe coefficient. 86% of the replicates also return a negative and significant estimate for the Oceania coefficient. By contrast, the South Asia coefficient was by far the worst performer, returning a positive and significant estimate in only 26% of replicates.

In general, the coefficient estimates generated using the referent data appear to be somewhat attenuated relative to the human-generated values. On average, estimates for most coefficients were 30-60% smaller in absolute value than the estimates produced using the human-generated referent data, and confidence intervals were approximately 20-30% narrower. Given the prediction strategy used in this paper, these results are unsurprising. Predictive modeling tools like random forests naturally "shrink" data points towards the mean in order to avoid overfitting, which makes the detection of unmodeled clustering structures more difficult. In cases where predictive accuracy is not the objective of interest, other approaches (e.g. Gaussian mixture models with appropriate clustering priors) may reveal these structures more easily. However, the inherent conservativism in predictive approaches can also be advantageous, since researchers can be assured that any unearthed clustering patterns are likely present in the underlying data structure.

[8] We counted a pair of machine/human-generated coefficients estimates as producing the same conclusion if both were negative and significant, both were positive and significant, or neither were significant.

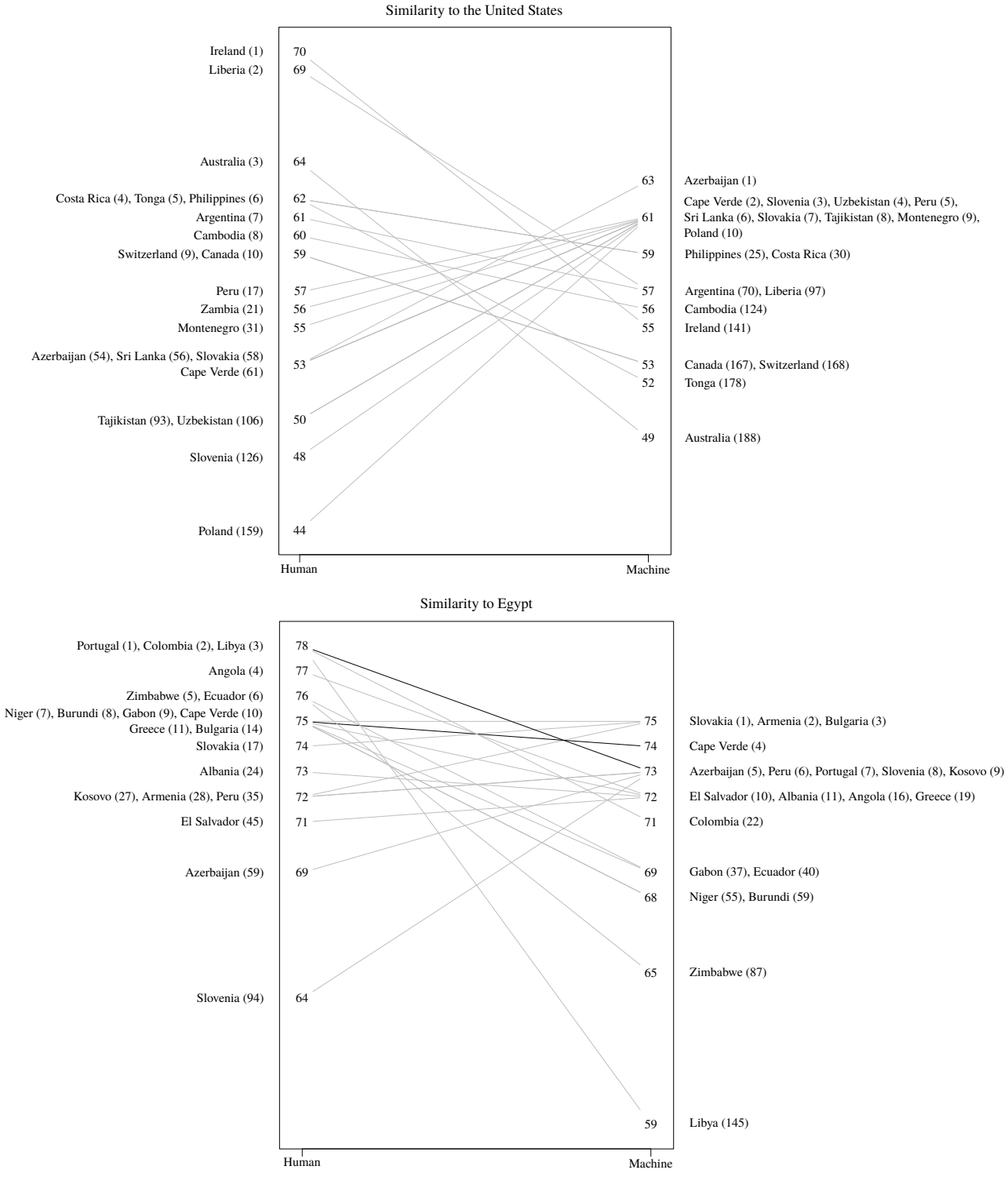
3.3 Test 3: Descriptive Inference across Target and Criterion Measures

In our third – and most exacting – test, we consider a descriptive, unit-level application. As with the causal inference application, the objective is to subject the measures to a more discrete and more meaningful assessment of the degree of correspondence between the measures. In general, we want to know whether a measure yields the same basic descriptive inferences as does the criterion measure. A common application of the measures would be to describe the similarity among particular countries of interest, or more often, the most-similar or least-similar countries to a country of interest. We summarize this sort of assessment as a "top-ten" test. In the stylized version of the test, the question is whether one measure's ten most similar cases matches those in another measure.

This kind of judgment is quite common across research domains, especially among more informal analyses. Consider a customer searching for a top-rated good or service. A standard, almost reflexive, exercise is to compare two lists of recommendations. If one list does not match another, especially if the one is particularly trustworthy, we might wonder about the other's validity. We propose, then, a simple test of descriptive inferential power in which a researcher compares rank-order similarity values for a given constitution with respect to the others in the dataset. We note that a move to a comparison of ordinal (from ratio) data loses some information by design – granular information that a standard measure of association would capture. However, the loss comes with gains in interpretation through the translation of measures to intuitive, everyday benchmarks.

In this spirit, we return to the comparison of human- and machine-generated similarity values used in the previous section. As in the causal inference test, we focus on the supervised similarity scores generated using the LDA_200 features. For each model replicate, we randomly select 75 documents ($\approx 40\%$ of the dataset), and

Figure 5: Similarity to the United States and Egypt, by two measures



Top ten most similar countries to Egypt and the US as identified by both machine- and human-generated similarity values. Machine-generated values represent results from a single randomly-chosen train-test split. Parenthetical values give similarity ranks for each country, by each measure. Dark lines connect observations that are shared in both top-ten lists.

train a random forest on all dyads contained within the 75-document training set. We then estimate similarity values for the held-out dyads, and compare estimated similarity values in the held-out test set with their human-generated target values. We then repeat this process 100 times, and compare results

By way of illustration, Figure 5 depicts a "top-ten" analysis based on a randomly-chosen train/test split for the United States (a Constitution familiar to many readers) and Egypt (a somewhat typical case, in terms of cross-measure correspondence). In both cases, the rank-ordering of countries appears appreciably different between the measures. In the case of the United States, there is no overlap between the lists. In the Egyptian case, the two top-ten lists share only two cases (Portugal and Cape Verde). Another way to think about the two lists is in terms of rank-order differences between the lists. On average, the per-observation difference in rank for the United States similarity set is 50.4, which suggests that a country ranked 100 on one list can be expected to be ranked 50 or 150 on the other list. Egypt, which again is more typical, has an average difference in rank of 30.4, which still seems substantial.

Broadened to the full set of 193 countries and 100 train/test replicates, these results remain largely consistent. Across all countries and train/test splits – 19,300 total comparison sets – the minimum average rank difference was 12.8 (held by Burkina Faso). The average absolute difference in rank across all countries and train/test splits is 31.3, similar to the value for Egypt given above. With respect to the top-ten comparison, the largest correspondence between a country's machine- and human-generated top-ten lists was held by the Bahamas, with an average of 5.7 shared members across all 100 train/test splits. Again, the measures did not correspond particularly well in the case of the United States, for which only 1.2 cases, on average, appeared on both top-ten lists (18th-worst, out of 193 countries). Of course, these differences in correspondence between the two measures, conditional on country, are worth exploring in future research. One might think of these conditional patterns

as violations of "measurement equivalence" – violations that often come with intriguing substantive explanations. That is, why would the machine measures not be as reliable in some sets of texts and as they would in others? Answering such questions would be a productive follow-on analysis here, and in other domains.

How should we think about the differences in descriptive inference between the two sets of measures? What would one think if these were two sets of restaurant or movie reviews? Our sense is that one would see them as written by critics with highly distinct sets of tastes. One might even wonder about the validity of one or the other metric as a measure of quality, depending upon which one were more trusted. More generally, it would seem that unit-level comparisons of this sort push the limit of what one could reliably expect from machine-interpreted similarity. That is, while the machine-interpreted similarity values perform well in terms of aggregate prediction and reasonably well with respect to causal inference, their individual-level accuracy with respect to rank-order comparisons will be less dependable. Or, at least, such measures are the product of enough systematic and non-systematic error of some sort that they do not very reliably replace human-interpreted measures in rank-order descriptive inferences.

4 Conclusion

Automated measures of textual similarity shed light on a host of important and challenging research questions. However, existing work provides little guidance about the performance of these techniques. In this paper, we evaluate the validity and utility of such similarity measures in the context of national constitutions. This application is substantively relevant for many applied researchers, but also presents a unique data opportunity. Using texts and content tags collected by the Comparative Constitutions Project, we construct a similarity metric we term *inventory similarity*,

which we calculate for each pair of in-force constitutions. We assume this human-interpreted measure of similarity to be valid and reliable and, therefore, useful as a point of reference in criterion validity evaluations. We conduct three such evaluations, and offer three conclusions for applied researchers:

1. For researchers interested in aggregate-level predictive accuracy, machine-generated similarity values are highly reliable. In our tests, values generated using a supervised machine learning approach performed best, with an out-of-sample correlation of $r \approx 0.7$ with $n = 75$ documents ($\approx 40\%$ of the dataset) used for training. Encouragingly, these results were largely invariant to the choice of model or parameter settings, at least across the specifications we examined.

2. For researchers interested in causal inference, machine-generated similarity values largely perform well. In our test of the so-called "contextual hypotheses," 88% of coefficients estimated based on the machine-generated referent returned the same conclusions as those estimated using the human-generated target (using $p = 0.05$ as the significance cutoff). Generally, coefficient values estimated using the machine-generated data were more conservative (smaller in absolute value and less likely to reject the null hypothesis) than those produced using the human-generated target. We view this conservatism as preferable to the alternative, giving researchers confidence that the effects they do observe are present in the human-generated target data.

3. For researchers interested in unit-level descriptive tasks (e.g. identifying the top-ten most similar constitutions to a given document), the similarity values we generate in this paper offer a limited correspondence with their human-generated counterparts. For projects requiring a high degree of rank-order correspondence between target and referent quantities, a larger training set or a different estimation procedure than those that we explore might be necessary.

In our view, these results are encouraging. Given a moderate-sized training set, analyses conducted using text-based similarity scores can recreate most results produced using hand-coded similarity comparisons. Like many hand-coding tasks, generating similarity comparisons is difficult and time-consuming for human evaluators, particularly when dealing with large quantities of long documents. In these situations, machine-generated approximations offer a useful way for researchers to test hypotheses regarding diffusion of ideas, content, and rhetoric in text form. The validation exercises we conduct in this paper provide some assurance for researchers uncertain about the promise of such tools, offering a path forward for applied work.

References

Adcock, R. and Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(03):529–546.

Ahlquist, J. S. and Breunig, C. (2012). Model-based clustering and typologies in the social sciences. *Political Analysis*, 20(1):92–112.

Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Jordan, M. I. (2004). Variational methods for the dirichlet process. In *Proceedings of the twenty-first international conference on Machine learning*, page 12. ACM.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brock, R. and Hodkinson, S. (2002). *Alternatives to Athens: varieties of political organization and community in ancient Greece*. Oxford University Press.

Cheibub, J. A., Elkins, Z., and Ginsburg, T. (2011). Latin american presidentialism in comparative and historical perspective. *Texas Law Review*, 89(7).

Cranmer, S. J., Leifeld, P., McClurg, S. D., and Rolfe, M. (2017). Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science*, 61(1):237–251.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407.

Dekker, D., Krackhardt, D., and Snijders, T. (2003). Multicollinearity robust qap for multiple regression. In *1st annual conference of the North American Association for Computational Social and Organizational Science*, pages 22–25. NAACSOS.

Denny, M. J. and Spirling, A. (2016). Assessing the consequences of text preprocessing decisions. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285. ACM.

Elkins, Z., Ginsburg, T., and Melton, J. (2009). *The endurance of national constitutions*. Cambridge University Press.

Elkins, Z., Ginsburg, T., and Melton, J. (2013). The content of authoritarian constitutions. In Ginsburg, T. and Simpser, A., editors, *Constitutions in authoritarian regimes*. Cambridge University Press.

Grimmer, J. (2010). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Grimmer, J. and King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650.

Hart, R. P. (2009). *Campaign talk: Why elections are good for us*. Princeton University Press.

Hillard, D., Purpura, S., and Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50.

Klebanov, B. B., Diermeier, D., and Beigman, E. (2008). Lexical cohesion analysis of political speech. *Political Analysis*, 16(4):447–463.

Krackardt, D. (1987). Qap partialling as a test of spuriousness. *Social networks*, 9(2):171–186.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, page mpu019.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Pennebaker, J. W., Chung, C. K., et al. (2008). Computerized text analysis of al-qaeda transcripts. In Krippendorff, K. and Bock, M. A., editors, *The Content Analysis Reader*, pages 453–465. Sage Press.

Petrie, K. J., Pennebaker, J. W., and Sivertsen, B. (2008). Things we said today: A linguistic analysis of the beatles. *Psychology of Aesthetics, Creativity, and the Arts*, 2(4):197.

Purpura, S. and Hillard, D. (2006). Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225. Digital Government Society of North America.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Roberts, M. E., Stewart, B. M., and Nielsen, R. (2015). Matching methods for high-dimensional data with applications to text.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Santini, S. and Jain, R. (1999). Similarity measures. *Pattern analysis and machine intelligence, IEEE transactions on*, 21(9):871–883.

Savoy, J. (2010). Lexical analysis of us political speeches. *Journal of Quantitative Linguistics*, 17(2):123–141.

Seawright, J. and Collier, D. (2014). Rival strategies of validation tools for evaluating measures of democracy. *Comparative Political Studies*, 47(1):111–138.

Sides, J. (2006). The origins of campaign agendas. *British Journal of Political Science*, 36(03):407–436.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Strehl, A., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64.

Sulkin, T. (2005). *Issue politics in Congress*. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the american statistical association*.

Tversky, A. and Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

A Coefficient Tables for In-Text Models

Table 2: Linear model coefficient estimates produced using human-generated similarity values

	Estimate	P-value
Sqrt_Yeardiff	-0.012	≤ 0.001
East Asia	-0.037	0.130
Eastern Europe	0.122	≤ 0.001
Western Europe/North America	0.007	0.679
Latin America	0.031	0.069
Middle East/North Africa	0.021	0.372
Oceania	-0.056	0.040
South Asia	0.127	0.001
Sub-Saharan Africa	0.005	0.679
Constant	0.656	≤ 0.001
Observations	18528	
Adjusted R^2	0.171	

Note: P-values generated using a QAP null distribution.