An Evaluation of Measures of Textual Similarity

Zachary Elkins
zelkins@austin.utexas.edu

Robert Shaffer
rbshaffer@utexas.edu

January 12, 2017

.

v.3.0

Abstract

Understanding the similarity of two texts can be extraordinarily enlightening and useful. Unfortunately, pairwise similarity comparisons are generally expensive and labor-intensive for human evaluators to conduct. Supervised and semi-supervised automated approaches are more scalable, if sometimes mysterious. Opportunities to evaluate these approaches are rare, especially in the political context. We leverage a unique dataset of human-coded national constitutions in order to evaluate automated measures, across both supervised and unsupervised approaches. First, we assess the scores from a series of plausible unsupervised feature extraction and similarity 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, and examine the sensitivity of out-of-sample performance to training set size and parameter values. We find that some of the automated measures correspond highly with the human-interpreted measure. We then assess the various automated measures in a set of applied criterion-validity tests, in which the reference point is, again, a trusted set of human-interpreted data on content similarity. The applied tests illustrate – more exactly and more vividly – the practical limits of the human-machine correspondence. The best automated measures yield the same set of causal inferences, as does the criterion measure, but not necessarily the same descriptive inferences.

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 might 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 the statements 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 diffusion of ideas across time and space. These are but two important examples. Many other inspiring

applications abound (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).

Both of our examples imply an evaluation 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 statements, advertising materials, or any number of multi-media publications. Resourceful scholars have long examined such documents in order to analyze political phenomena across time and space. But opportunities for such analyses have exploded. Modern scholars have at their disposal a seemingly bottomless trove of text, both contemporary and historical. Corpora of interest commonly contain thousands or even millions of documents, each of which may range from short utterances to long, detailed statements. Of course, the modern scholar also has a growing set of computational tools to accompany data at this scale.

As a result of these trends, unsupervised modeling approaches designed for the text-as-data setting (e.g. topic models) have proliferated in political science.[1] Results based on automated methods are captivating and relatively easy to generate, but it is not always clear how features extracted by such models relate to human constructs. For one thing, 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. Estimates of the uncertainty surrounding these similarity scores are similarly challenging to generate. A common approach in these situations is to "train" the models to interpret text according to a set of interpretative criteria. Compared with unsupervised procedures, supervised learning approaches offer a sharper focus on a concept of interest. 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 what cost?).

To explore this set of measurement challenges, we turn to a set of original hand-coded (human-interpreted) data on the content of national constitutions. Because of their cover-

[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.

age 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. Indeed, we think of the data as providing something akin to a "gold standard" in a *criterion-validity test*. 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 what we will call *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 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 ($r = 0.6$). In order to understand the full implications of this seemingly high correspondence, we compare the measures in two applied settings. We think of these two subsequent analyses as further illustrations and more exacting evaluations of criterion-validity. Indeed, the analyses evaluate criterion validity in progressively demanding ways. The first applied test integrates the logic of criterion validity with that of construct (nomological) validity. In that test, we incorporate – sequentially – the human- and machine-generated measures in a set of models predicting the similarity of national constitutions. We find that analyses with the (1) human- and (2) machine-generated measures point to essentially the same causal inferences. In the second applied setting, we test whether the various measures return the same basic *descriptive* inferences about "top-ten" cases. This last test reveals substantial differences across the measures, which reminds us of the substantive limits of their correspondence.

2 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] Those similarity values could then be used as an object of analysis in their own right (e.g. comparing average similarity between two candidates over the course of the campaign), or as a component in a subsequent analysis (e.g. grouping candidates into clusters and comparing within- and between-group variation over time). We view such analyses as potentially enlightening and even pathbreaking, but our optimism is tempered by our uncertainty about measurement error. To what degree can we trust any such similarity scores?

Similarity calculations rely on an implicit definition of both the relevant features and notion of similarity under consideration. 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. Thus, we might think of two central concerns in building such measures: (1) the ingredients of the data vector; and (2) the comparison procedure.

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 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 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).

ten more interested in similarity comparisons based on a more specific conceptualization and applied to longer texts. Unsurprisingly, this use case has not been as closely studied. 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. As a result, opportunities to evaluate automated similarity procedures in these cases are rare.

3 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. In general, there is no shortage of claims (and counterclaims) about the diffusion and influence of ideas in this genre. 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).

Swirling around these kinds of discussions is the idea of constitutional similarity, as important a concept as it is intriguing. Indeed, the concept would seem to be the lifeblood of scholars of comparative political institutions, who depend upon it to evaluate hypotheses related to the origin, spread, and novelty of ideas across jurisdictions. However, as in the examples given in the previous section, 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

Figure 1: Visitors to the NCC Interact with Textual Similarity Scores

Photo courtesy of Jessie Baugher

of constitutional text.

Importantly, challenges regarding the construction of valid and interpretable similarity measures are not restricted to the academic domain. For example, in February of 2014, the National Constitutions Center in Philadelphia commissioned a new interactive exhibit featuring data drawn from the Comparative Constitutions Project (CCP). As part of the exhibit, researchers created two interactive visualizations that featured similarity scores between (a) successive drafts of the U.S. Bill of Rights, and (b) each right in the U.S. Constitution and that in constitutions from other countries.

The exhibit design process, and the reactions of museum visitors and NCC staff, exhibited many of the same promises and concerns raised in the preceding discussion. Observation and analysis of visitor experiences suggested that most visitors and staff seemed drawn to

the similarity scores (to the exclusion of other elements in the interactive), and the scores prompted as much skepticism as they did intrigue (citation suppressed). Most notably, the similarity scores did not always exhibit face validity. One particular comparison – that of the rights regarding the collection of evidence in the U.S. and Japan – became an especially central, and vexing point of discussion. The fourth amendment to the U.S. Constitution was, according to some, an influential model for the drafters of the Japanese Constitution of 1946. Indeed, any reader will notice striking similarities. Nevertheless, many textual similarity measures did not score the pair of texts very highly. In cases such as these, where scores did not match expectations, viewers wondered how the scores were constructed. Were the algorithms matching ideas, matching words, or something else? The audience reaction reflected an impression that we have of textual similarity analyses in the scholarly domain more generally. Even in more technical settings, text similarity methods *seem* to retain something of an alluring – but mysterious and untested – quality to them.

4 A Proposed Series of Criterion-Validation Tests

Our validation proceeds in three phases, which we view as illuminating different degrees and kinds 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).

We consciously mix a set of insights from several types of validation, though our central comparisons are rooted in a criterion-validation logic. Specifically, we envision a series of tests in which we evaluate measures at progressively demanding levels of scrutiny. The first test is a basic step in criterion validation and something akin to convergent, or predictive, validation: i.e., does the measure in question correlate strongly with the criterion measure. We note that the evaluation of this test involves fairly impressionistic benchmarks (notions of high and low correlation), though perhaps benchmarks to which social scientists are well accustomed. The second test is more applied and, even, more exacting in that it establishes

some brighter lines for adjudication. It incorporates the idea of construct, or nomological, validation – the question of whether the target measure predicts outcomes in accordance with widely theorized causal relationships. To this "nomological" logic, we combine a layer of criterion validation by introducing the criterion measure as a reference point. So, does the target measure predict a widely-hypothesized outcome in the same way as does the criterion measure? Finally, in our third stage of validation, we incorporate what we see as an even stricter bright-line test. While the first two tests assess average levels of association among concepts, the third test puts measures to the task of descriptive inference. The question in the third test is whether the target measure returns the same rank ordering of cases (and even the same ratio-level positions of cases), as does the criterion measure. We view this last test as potentially unfair, but a test that is especially common and intuitive. 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 conceptualization, are those that validate measures based on their performance in applied settings.

The key point of leverage in our analysis is our criterion measure, the design and conceptualization of which matter significantly. We develop a criterion measure using a comprehensive dataset on the content of constitutions. The unique features of the data lead us to think of the measure as promising material for a gold-standard measure of constitutional similarity. Clearly, there is an array of possible approaches by which one might extract features from, and assess the similarity of, a pair of high-dimensional texts. For our purposes, we focus on a notion we term *inventory similarity*, which we argue represents a useful conceptualization for investigations of document content. We first outline this idea in more detail, and then use the basic framework to extract similarity scores from data generated by the Comparative Constitutions Project (CCP).

4.1 Features of the Reference (Criterion) Data

Two attributes of the CCP data are convenient for the analysis 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 records comprehensive data on some 600 topics in written constitutions. This broad coverage was intended to allow analysts to capture and trace the flow of constitutional ideas across countries and time and assess the consequences of different constitutional choices. Arguably, this comprehensive review of a text will be more comparable to that of an automated analysis, especially one untrained to focus on certain topics.

The CCP's temporal and geographic coverage are similarly comprehensive. The CCP's sample includes the constitution (and its updates) for nearly every independent state since 1789. In this paper we analyze a slice of this dataset consisting of all constitutions in force in 192 countries in 2014, and generate and replicate similarity scores between each constitutional pair in that year.

4.2 A Document's Inventory

A primary concern in building a measure of similarity based on human annotations has to do with the measure's *ingredients*. While the CCP's dataset is extensive, not all attributes contained in the CCP codebook are appropriate for generating textual similarity. A guiding question here has to do with what sort of similarity, exactly, we seek to measure. For example, we could ask whether two constitutions make the same choices on any given list of provisions. But we can also ask, more simply, whether two constitutions even *address* the same issues. Here, we focus on the latter idea. The measure we develop (which we term *inventory similarity*) tells us something about the basic proto-architecture of a constitution – that is, not so much how a feature is constructed, but what its designers thought appropriate or necessary to design in the first place. That table-of-contents inquiry is sequentially prior to one about downstream design choices, and it is perhaps more consequential. But also, a measure of inventory similarity is particularly well matched to the machine-interpreted measures that we evaluate here, most of which (e.g., topic models) purport to measure inventory. In that sense, such a measure contributes to a more equivalent comparison set.

Thus, our measure of *inventory similarity* compares the array of topics included (or not included) in any two constitutions. The dataset consists of a set of binary variables indicating whether topic x or y is present or not (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 have identified 70 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, though our early exploration (citation suppressed) suggests that the variance in items may not be especially consequential. Nonetheless, the goal is to identify topics at a high level of generality and thus measure broad areas of inclusion or exclusion in constitutions.

4.3 Similarity, not Distance

We digress briefly to clarify the concepts of similarity and distance, which differ in subtle ways that have important analytic implications. Informally, "similarity" is sometimes understood as the inverse of distance. The concept of similarity, however, is a general one without precise (or, at least, intersubjectively shared) definitional attributes. By contrast, in the mathematical domain, a distance $\delta(a, b)$ conventionally refers to a metric distance, or a function which satisfies the metric axioms:

$$minimality : \delta(a, b) \geq \delta(a, a) = 0$$

$$symmetry : \delta(a, b) = \delta(b, a)$$

$$triangle\ ineq. : \delta(a, b) + \delta(b, c) \geq \delta(a, c)$$

These axioms formalize basic intuitions regarding distance in physical space. The distance between two distinct points should be greater than the distance between a point and

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

itself (*minimality*); the distance between two points should be the same regardless of where we start (*symmetry*); and, the straight-line distance between two points should be shorter than any other possible route (the *triangle inequality*). Metric distances are ubiquitous in computer science and statistical applications, forming the basis for clustering methods, matching, and, most directly, distance learning applications.

Though the metric axioms are mathematically and intuitively appealing, we do not restrict the similarity values we generate to obey these constraints. From a formal perspective, the metric axioms represent restrictions on the space of functions that we are willing to consider as we attempt to approximate the human-generated similarity values we consider. Though traits like symmetry and minimality are intuitively appealing, in exploratory testing we found that relaxing these requirements allowed us to boost performance substantially. In one series of tests, for example, we fit a weighted Mahalanobis distance function[4] to human-created similarity values; however, out-of-sample values generated using this approach substantially underperformed non-metric alternatives.[5]

4.4 Calculating Similarity

One can assess the association between two units across a set of binary items using a seemingly endless set of formulae. Here, however, we employ Jaccard's (1912) similarity coefficient[6], defined as

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

This basic approach is only applicable for binary feature vectors; however, the Jaccard coefficient can easily be extended to compare sets of arbitrary features consisting of non-negative continuous values (see, e.g., Strehl et al. (2000)). In their basic forms, both the

[4]Using L-BFGS with an L_2 regularization penalty proportional to the squared sum of the weights.

[5]Philosophically, the move to non-metric similarity approximations is also appealing. As a substantial body of psychological research demonstrates, human similarity comparisons frequently violate all three of the metric axioms (Tversky 1977; Tversky and Gati 1978, 1982; Santini and Jain 1999). For the purposes of this paper, our objective function is a proper metric distance; however, for more general applications, the flexibility to replicate non-metric human-generated similarity functions is useful.

[6]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, $\delta_J(X,Y) = \frac{I(X,Y)}{H(X,Y)}$

binary and extended Jaccard coefficients fulfill the metric axioms, but both functions can also be generalized to non-metric, asymmetric variants depending on the depending on the setting of interest (Tversky 1977).

Informally, this approach has at least two attractive properties compared with other possible approaches:

1. Jaccard similarity is easily interpretable and efficient to compute.
2. Jaccard similarity avoids inflating the similarity of short documents.

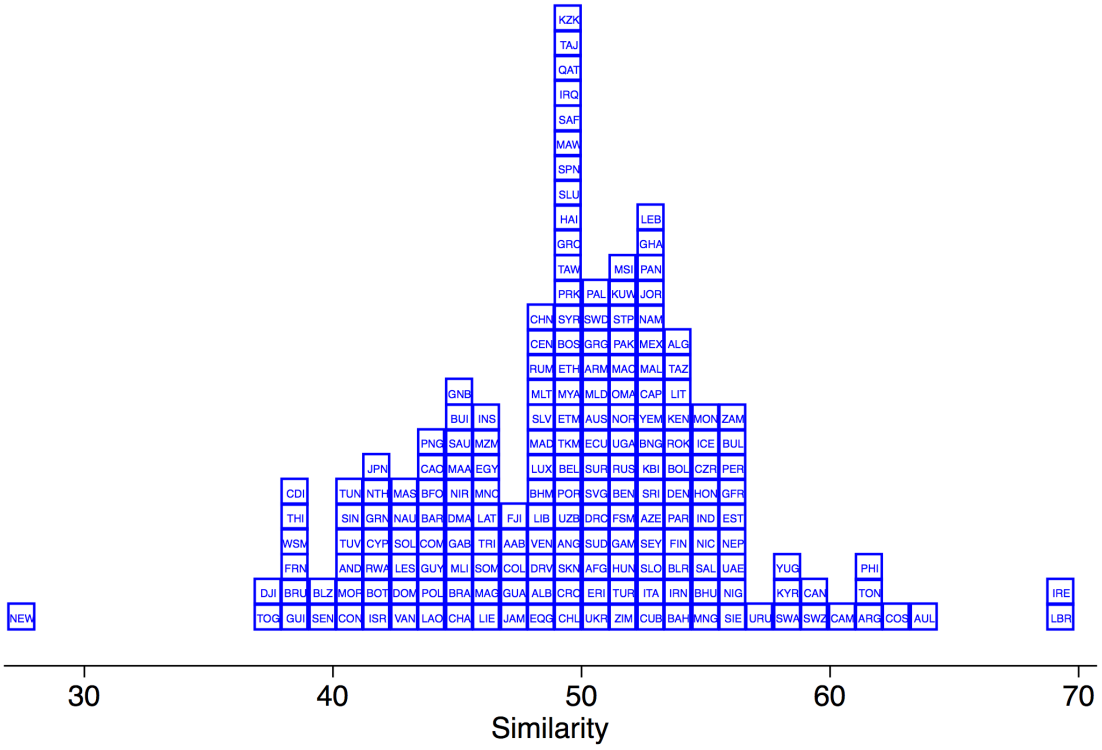
As in many text analysis domains, constitutions tend to contain a sparse array of features, with a large number of absent elements in any given document. In the extreme case, consider two constitutions that discuss one topic each; say, executive power in one, and legislative power in the other. Clearly, when these constitutions *do* speak, they may speak very differently. They simply choose not to speak much. Naively-selected metrics (say, a procedure which simply counted the number of decisions with the same value, on either omissions or inclusions) would likely produce a very high similarity between these two documents. Jaccard similarity avoids this issue by restricting the comparison to the features present in at least one of the two cases under consideration.

4.5 A Measure of Inventory Similarity

We thus calculate a measure of similarity as described for the sample of 194 constitutions in force in 2014 across 70 binary topics. The result is a matrix with some 18,721 unique constitutional dyads (one score for each of the 194 constitutions and its 193 counterparts). 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.

Figure 2: Similarity to the U.S. Constitution (hand-coded, across topics)



The U.S. Constitution may be more widely known (at least compared to Brunei). In Figure 2[7], 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).

5 Machine-interpreted measures of Constitutional Similarity

The next step is to generate computational measures of the similarity of the raw text of the constitutions in the sample. Our general approach is to develop measures similar to those that are widely used by leading practitioners. Our measurement and analysis proceeds in three steps. First, we fit a series of unsupervised data reduction models models on our constitutional corpus, and use the model parameters to extract a vector of features from each constitution. Second, we use these features to calculate distances between documents, and compare these unsupervised scores to the human-derived similarity judgments. Third, we use the those feature vectors as inputs to a supervised similarity learning algorithm, and examine out-of-sample predictive accuracy. We also compare performance at various training/test set sizes, and discuss implications for other studies.

5.1 Feature Extraction

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 the political science literature (Grimmer and King 2011;

[7]Note that Figure 2, like some of the analysis later, re-scales the measure by multiplying it by 100).

Lucas et al. 2015). Generally speaking, 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). In early versions of these models, estimation proceeded 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, are usually generative.[8] Latent Dirichlet Allocation (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 [1].

To generate similarity values, we calculate a weighted average of document-level features extracted by each model into a constitution-level feature vector (weighted by word count in each document), and calculate a distance between between each pair of documents. For LDA

[8]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 data dimensionality, do not have this property.

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

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

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 models and parameter settings are not the only available feature extraction/similarity approaches, and many other plausible approaches exist. However, each of these approaches has been employed extensively in a variety of problem domains, and represent a reasonable starting point for our initial investigations.

6 Criterion Validity Tests

Recall that we propose to test the set of machine measures of similarity (the target set) against a human-derived measure (the criterion, or referent). We do so in three progressive steps. First we analyze the co-variation between the measures in the target set and the referent measure. That analysis informs an initial assessment of the convergent validity of the target measures. However, such analysis remains somewhat abstract, which is why we test the behavior of the various measures in applied settings in two subsequent tests. In a second test, we evaluate the performance of the two sets of measures in a statistical model that predicts the similarity scores with a widely hypothesized set of explanatory variables. The question in that test is whether the two sets of measures yield the same causal inferences. Finally, we evaluate whether the two measures deliver the same descriptive inferences.

Table 1: Estimation and pre-processing details for models 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

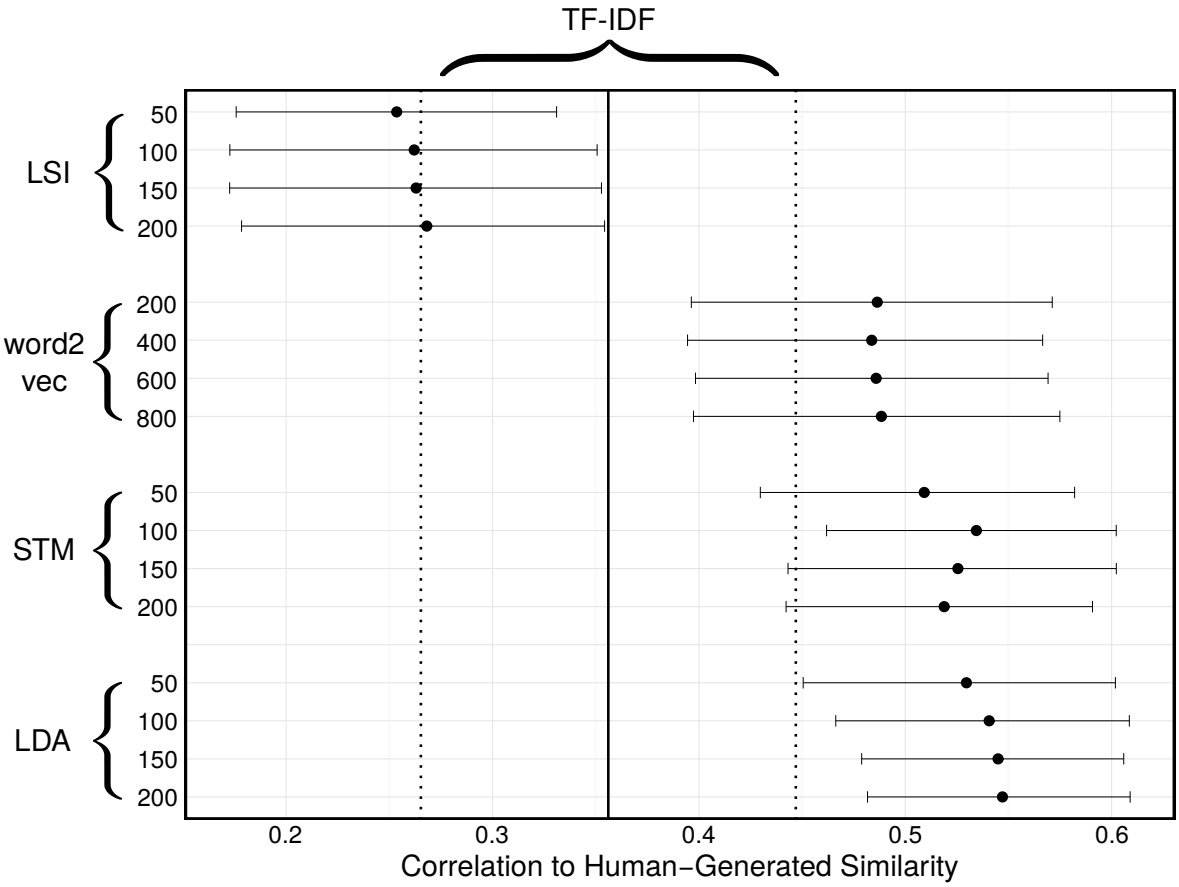
Where not specified, all parameters left at default settings. LSI and TF-IDF estimated via Gensim (Řehůřek and Sojka 2010). LDA estimated via MALLET (McCallum 2002). STM estimated via the STM R package (Roberts et al. 2014). Asymmetric alpha prior, as implemented in Wallach et al. Wallach et al. (2009).

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

6.1 Test 1: Association between Target and Criterion Measures

To what degree do the machine measures of similarity covary with human measure? We begin with an evaluation of the unsupervised approaches, a fairly agnostic place for an analyst to begin (see Figure 3). 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 the non-generative alternatives (LSI, TF-IDF word count vectors). Similarities created using the three generative models all correlate with hand-coded data in the $r = (.45, .55)$ 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 3: Correlations between machine- and human-generated similarity values



Solid and dashed vertical lines indicate the baseline correlation between similarity generated using baseline TF-IDF features and CCP data. Dots and solid horizontal lines indicate correlations between similarities generated using other feature extraction approaches. All confidence intervals represent ± 2 standard deviations, generated using a modified block bootstrap procedure. For each bootstrap replicate, we resampled a dataset of countries with size equal to the original data vector, then retrieved all pairwise similarity values between members of the bootstrapped dataset.

Importantly, across all of these models, the choice of the number of dimensions appears to have little impact on results. This invariance is encouraging. In many modeling settings, selecting dimensionality parameters (such as the number of topics in a topic model) represents 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. As before, we emphasize that this is not the only (or necessarily the optimal) approach one might consider in this context; however, it offers a useful starting point for future analysis.

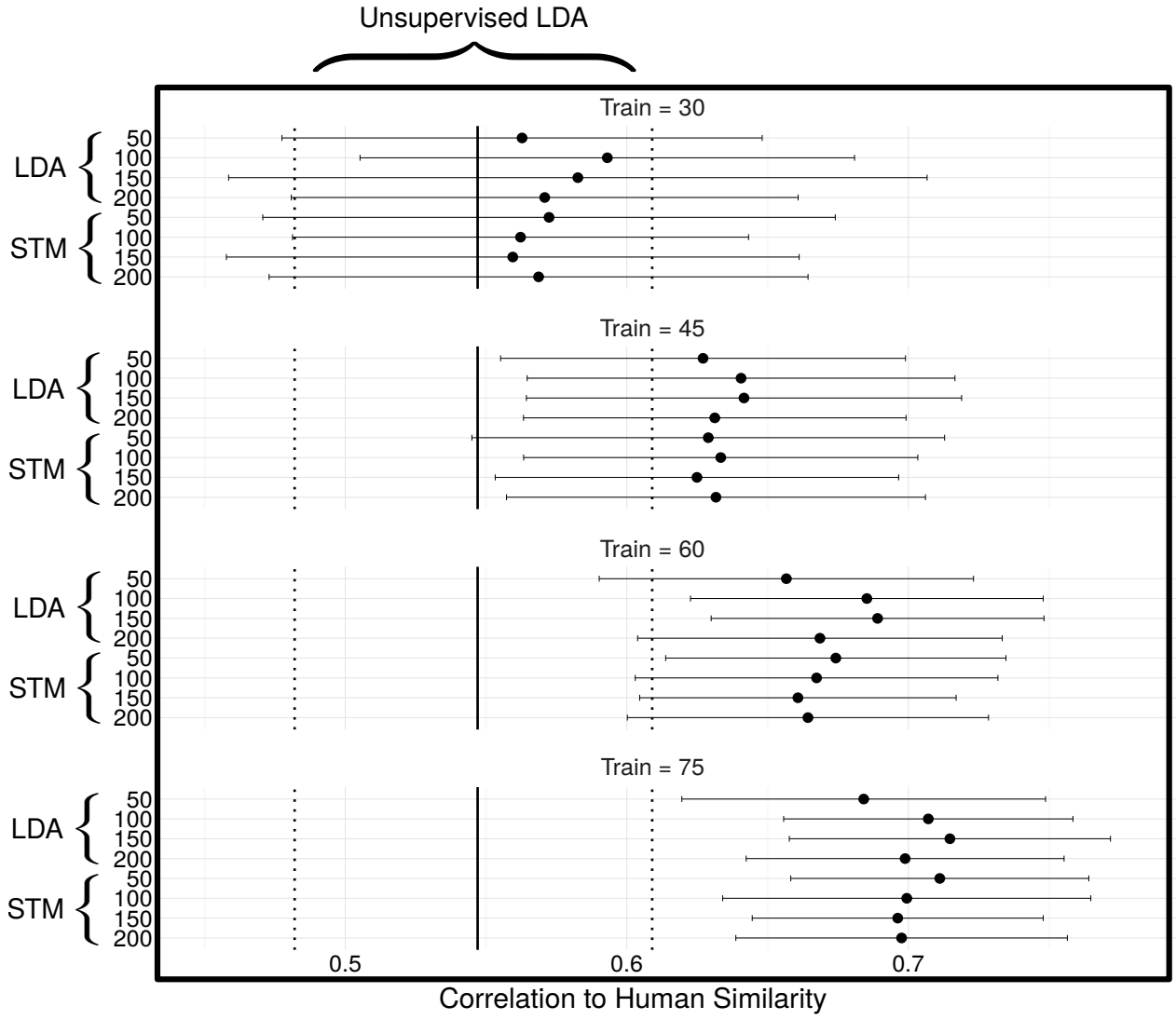
As shown in Figure 4, moving to a supervised alternative offers substantial improvements to predictive accuracy compared with the unsupervised alternative (using LDA_200 features as the point of comparison). With as few as 45 documents ($\approx 25\%$ of the dataset), the supervised predictions consistently correlate more highly with the human-generated similarity measures than the unsupervised comparison shown above. By 75 documents ($\approx 40\%$), 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 .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 and model specification is overwhelmed by variation generated through selection of the training set.

6.2 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 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 drafters 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). A second finding has to do with path dependence and the stickiness of constitutional ideas within a countrys historical series

Figure 4: Out-of-sample correlation between predicted similarities generated through a supervised procedure



Facets indicate the number of documents used to form the training set, out of 192 total. Solid lines represent ± 2 sample standard deviations, estimated from 100 independent model runs. Training sets were selected by randomly sampling a set of countries, and using all dyads within that set as training examples. Solid and dashed lines give mean and 95% confidence intervals from the unsupervised LDA_200 model as a baseline comparison.

of constitutional texts. Constitutions produced in the same country understandably bear a striking resemblance to one another; a research question, then, becomes the identification of those constitutions in the series that break new ground. We explore the first these questions in the analyses below, and identify the second as a promising avenue for other research.

We test the contextual hypotheses with a set of regression models that predict similarity across the sample of 18,721 constitutional pairs. The relevant question is whether this basic set of predictors performs equivalently across two dependent variables: a (1) human and a (2) machine measure of similarity.

The models include a very basic set of predictors that test the contextual hypotheses and control for aspects of the documents (length and scope) that we reason might contaminate the similarity measures. Thus, the explanatory variables are:

1. *Difference in year of enactment (100's)*, calculated as the absolute value of the difference in the years in which the two constitutions were first enacted. This variable captures any generational effect.

2. *Same region.* A set of dummy variables that equal 1 if a dyad includes constitutions from the same region, for each of four geographic regions. Any regional classification will entail controversial decisions. We use a 8-part regional classification developed for use in the CCP. The four regions included in the model are Latin America, Sub-Saharan Africa, Eastern Europe, and Middle East/North Africa. The excluded regions in the reference set are East Asia, Oceania, South Asia, and Western Europe/USA/Canada.

3. *Difference in Scope.* The absolute value of the difference in a measure of *scope* (Elkins et al. 2009) between the two constitutions in the dyad. Scope is a count, ranging from 0 to 70, of the number of topics included in a given constitution.

4. *Difference in Length (words, 10,000's).* The absolute value of the difference in the number of words between the two constitutions in the dyad.

Table 2 reports the results of an OLS regression in which we predict similarity (as measured by four different methods) with the set of explanatory variables described above.

Table 2: Models of Inventory Similarity, Human- and Machine-coded

	Human	LDA	RF (25)	RF (50)
Difference in year of enactment (100's)	-5.41	-6.09	-4.12	-2.17
	[0.16]	[0.38]	[0.14]	[0.11]
Both Latin America	3.59	-3.43	-1.96	2.10
	[0.36]	[0.89]	[0.31]	[0.27]
Both Eastern Europe	10.70	10.77	7.88	7.47
	[0.42]	[1.02]	[0.35]	[0.31]
Both Middle East/North Africa	1.68	14.86	0.02	1.17
	[0.63]	[1.54]	[0.53]	[0.47]
Both Sub-Saharan Africa	0.93	9.42	0.68	-0.70
	[0.26]	[0.65]	[0.23]	[0.20]
Difference in scope	-0.76	-0.48	-0.29	-0.24
	[0.01]	[0.02]	[0.01]	[0.01]
Difference in length (words 10,000's)	-0.33	-3.80	-0.52	-0.56
	[0.04]	[0.09]	[0.03]	[0.03]
Constant	68.56	73.60	63.97	64.50
	[0.12]	[0.30]	[0.10]	[0.09]
R^2	0.36	0.17	0.19	0.19
N	18,145	18,145	17,155	14,050

Again, we test three machine-coded methods against the human-coded referent. Standard errors are clustered at the country level, since each country's score is repeated across some 190 dyads, whose errors are likely interdependent. For presentation purposes, we have re-scaled the similarity measures to range from 0 to 100, instead of 0 to 1.

We start with the reference point. The results from the human-coded measure (column 2) provide some compelling evidence for the contextual hypotheses. For one thing, a generational effect is readily apparent: constitutions separated by 100 years are about 5 points less similar on average (b = -5.41). Geographic location also matters, but the effect varies substantially across regions. (In other regressions (not shown here), a general same-region variable suggests that constitutions from the same region are, on average, 2.5 points more similar). Eastern European constitutions exhibit an extraordinary degree of clustering (b = 10.70). On the other hand, the effect for sub-Sahara is much smaller (b = 0.93). The scope and length variables are also strongly associated with the measure of similarity, as we suspected, which suggests that our similarity measures are picking up effects related to size and not just to content.

Our focal question is whether the predictors behave comparably when we predict similarity as measured by machine? To some extent, yes. The LDA analysis returns the same inference regarding the era variable and the Eastern Europe dummy. On the other hand the other regional variables in that equation suggest either greatly enhanced clustering (Middle East and Africa) or, in one case, negative clustering (Latin America), compared to the results in the human-coded measure. The Latin America effect is especially perplexing. Upon investigation, it appears that the LDA measure returned a comparatively low level of similarity between South American countries and the Caribbean island countries (including the non-Latin ones, which are included in the regional dummy). In other words, the LDA measure may have picked up some systematic differences (and, perhaps, substantively interesting differences) between co-regional countries. This difference, then, is a case of the machine recovering meaningful elements that may have been lost in a human-coded measure.

The supervised measures, to varying degrees, tend to recapture the effects that we saw in the human-coded measures. That correspondence is especially evident in the case of the

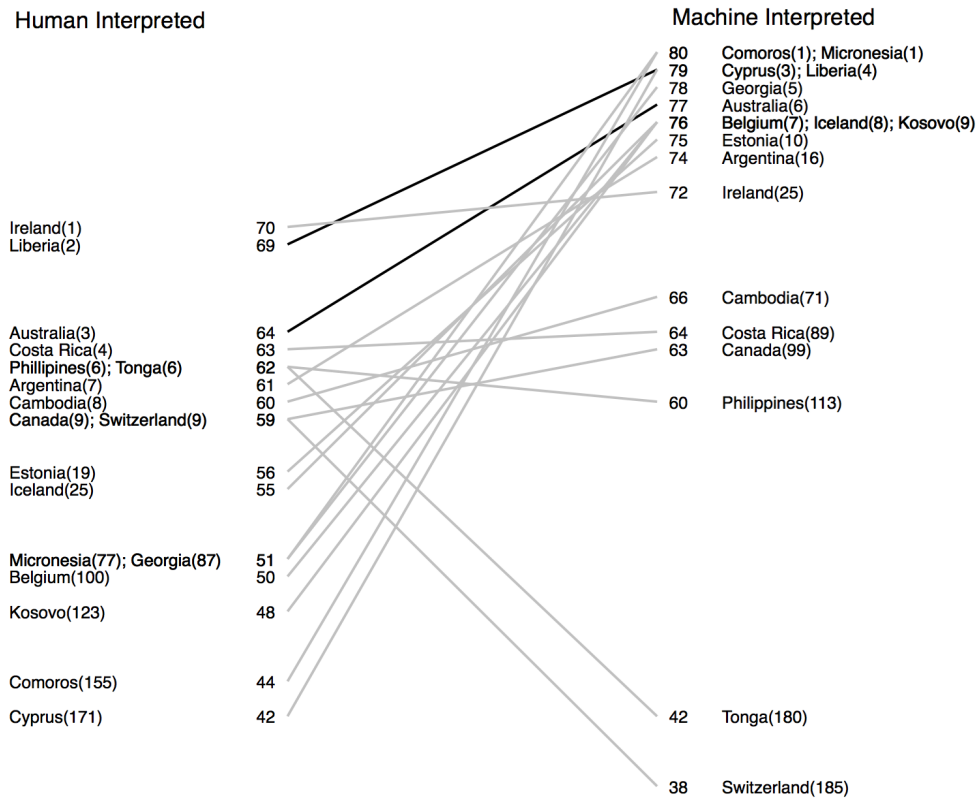
supervised measure trained on 50 percent of the cases, whose estimates are nearly identical to those from the human-coded measure. It seems fair to say that the machine measures passed Test 2. One would reach the same interpretation about patterns of diffusion in the data, regardless of which measure of similarity one employed.

6.3 Test 3: Descriptive Inference across Target and Criterion Measures

We now turn to a test of descriptive inference, or what we'll refer to as the "Japan problem," after the museum visitor described on page 7. Recall that our visitor became fixated on an important data point: in this case, a low similarity score between the fourth amendment in the U.S. Constitution and the corresponding clause in the Japanese Constitution. Of course, it is too much to ask of any measure to render face-valid point estimates for each and every case. In general, however, we *would* want to know whether a measure yields the same basic descriptive inferences as does the criterion measure. For example, one might wish to identify the most and least similar constitutions within a particular constitutional orbit. This kind of judgment is quite common across domains. Think of customers who look for top-rated goods and services. A standard, almost reflexive, exercise is to compare two lists of recommendations. If one list does not match those from another, especially if the latter is particularly trustworthy, we might wonder about the former's validity. We propose, then, that one simple test of descriptive inferential power is a comparison of top-ten lists. We note that such a comparison *loses information by design*, information of which measures of association make good use. At the same time, the test approximates a common (perhaps inevitable) exercise on which individuals have strong intuitions.

In this spirit, we return to the U.S. Constitution, a familiar benchmark. We consider the similarity of constitutions to that of the U.S., according to each measure. The slopegraph in Figure 5 plots human and machine similarity scores for those countries ranked in the top ten in either the human or machine measure (LDA). Note that the figure positions cases vertically according to its similarity score, though much of what we may care about here is its ordinal rank (which is noted parenthetically). Clearly, the two measures are telling very different top-ten stories. Only two of the human's top ten cases crack the machine's top ten, and many of those that do not overlap miss by wide margins (e.g., Cyprus, Comoros,

Figure 5: Ten Most Similar Constitutions to the USA, by two measures;



Notes: (a) countries are arranged by their score on each measure with ordinal rank in parentheses; (b) **darker** lines indicate cases that appear in both top-ten lists

Tonga, and Bangladesh).

What would one think if these were two sets of restaurant or movie reviews? One might wonder about the validity of one or the other metric as a measure of quality, depending upon which one were more trusted. Note that this is a judgment, at this point, about validity, not reliability; one might not (yet) surmise that one has more unsystematic error than the other. Rather, one might suspect that the two sets of reviewers had wildly different tastes, though scored with equal amounts of random error. It is, in fact, in looking at cases "off the diagonal" that score differently on the two measures that one learns something about the meaning of the two scales. Such analysis is a logical next step in an comprehensive evaluation of the measures in question here. We have done so, but leave that analysis for a separate treatment.

7 Conclusion

Automated measures of textual similarity promise to shed light on a host of important and challenging research questions. The promise is rooted in the relative mountain of historical documents with political relevance, and the continuing accumulation of such content, already in digital form. Conceivably, automated content analysis can return measures of interest quite efficiently. One quantity of particular interest to many analysts is the similarity of any two documents, which would also seem to be one of the central strengths of such methods. One could imagine all sorts of applications in political science for such methods. We happen to write on the heels of a heated political election and campaign, which suggests one application. Candidates for office produce volumes of content, perhaps *ad nauseum*, about their beliefs and intentions. How do the ideas of these candidates shift and drift, and which candidates cluster or diverge in their rhetoric and ideas? Automated methods could provide a vivid set of answers to such questions. However, the validity of similarity measures generated automatically is not at all clear to us, despite the wide use of topic models and the like. Hence our exploration here.

We evaluate the validity and utility such similarity measures in the context of national constitutions, a domain of particular interest to us and others, but also one that offers

unique analytic leverage. Specifically, we exploit an original set of hand-coded (human-interpreted) data on the content of constitutions with which we construct a measure of topic similarity between any two constitutions. We assume this human-interpreted measure of similarity to be highly valid and reliable and, therefore, useful as a point of reference in criterion validity tests. We thus develop and calculate a parallel set of automated measures of similarity for constitutional text. Our interest is, in part, in discovering differences in validity across different approaches to these measures, including feature extraction methods and degrees of analytic supervision. The results suggest appreciable variation in the degree to which measures under evaluation predict the criterion measure. However, the more general message is that the best performing of these measures correlate very strongly with the criterion measure. This apparently strong correlation is despite understandable differences in the substantive focus of the machine, as against the human interpreters. As such, the strong association between the best-performing automated measures and the criterion measure suggests that the automated measures are capturing meaningful levels of similarity and may be fulfilling their promise. Indeed, when we explore the behavior of these high-performing automated measures in explanatory models of constitutional similarity, we note that the automated measures mostly return the same substantive results as do human-interpreted measures. We see that finding as a sign of construct (nomological) validity, which gives us even more confidence in the automated approach.

Nevertheless, we sought to subject the machine measures to a more stringent stress test in an applied setting. In this spirit, we demonstrate a simple comparison of "top-ten" lists, as scored by the criterion measure and the top-performing machine measure. We see such a comparison as a common exercise in descriptive inference, and one that illuminates the degree of correspondence between the measures. Indeed, the test provides some perspective on the real limits of the measures' correspondence. Specifically, the tests revealed that – despite the measures' strong correlation and similar predictions – the two measures are not by any means substitutes. They do not reproduce the same cases in ways that one might expect in, say, product reviews. The inspection of these anomolies, of course, lead to a deeper understanding of what, exactly, is measured in each instrument. We leave that analysis as a natural next step in the research program, especially since machine measures

suggest reason to be optimistic about their utility.

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).

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.

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.

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. (1977). Features of similarity. *Psychological Review*, 84(4):327.

Tversky, A. and Gati, I. (1978). Studies of similarity. *Cognition and Categorization*, 1(1978):79–98.

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.