An Evaluation of Measures of Textual Similarity

Robert Shaffer
rbshaffer@utexas.edu

Zachary Elkins
zelkins@austin.utexas.edu

January 1, 2018

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 their validity is relatively unknown. We leverage a unique dataset of human-coded national constitutions in order to evaluate a set of automated approaches to similarity. 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 exhibit admirable predictive and inferential properties, but limited unit-level descriptive performance.

Word count: 8468

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. For example, a political behavior researcher might cluster or match responses to open-ended survey questions in order to probe their initial intuitions. Sometimes similarity is an end in itself. For example, a scholar of electoral campaigns might compare statements made by 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 these examples 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. Textual similarity is multi-dimensional, and may vary according to content, semantics, style, sentiment, or some combination thereof. Moreover, even with an appropriate set of features, the functional form relating those features to human-interpretable constructs is usually difficult to define. Supervised machine learning techniques can help address these challenges, but defining the relevant features and gathering human-generated training data is laborious and offers an unclear payoff. How much training is enough,

[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.

and at what cost?

To explore this set of measurement challenges, we leverage the Comparative Constitutions Project's original hand-coded data on the content of national constitutions (Elkins et al. 2009). 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 set of baseline similarity scores using an approach to thematic similarity that we term *inventory similarity*. We think of this measure as something like a criterion measure in a criterion validity approach, with the assumption that a high degree of correspondence between it and a candidate measure would indicate the latter's validity. We attempt to reproduce the *criterion* measure using various automated feature extraction and learning approaches (*candidates*). Since one of our key problems of interest is learner performance as a function of training set size, we explore both supervised and unsupervised learning approaches, with varying training set sizes in the supervised case. We find that some approaches (candidates) predict the criterion (human) measure appreciably better than do others (especially in the unsupervised setting) and the best approaches predict the human scores remarkably well ($r \approx 0.7$).[2]

We further evaluate the validity of the similarity values we produce using two applied tests, which are extensions of a criterion validity approach. First, 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. Second, 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]We refer, interchangeably, to the human measure as the "criterion" and the various machine measures as "candidates."

2 Textual Similarity: Promise and Challenges

2.1 Why Measure Similarity?

Similarity comparisons between cases can be illuminating at different points 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. 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).[3] We view such analyses as potentially enlightening and even pathbreaking, but our optimism is tempered by our uncertainty about measurement error and interpretability.

In the settings described previously, machine-generated similarity comparison approaches offer a natural approach. Unfortunately, opportunities to validate machine-generated similarity scores 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 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 words or sentences based on some undefined notion of similarity, and then learn a function that replicates their judgments. By contrast, in political

[3]Researchers have conducted similar analyses in a variety of more or less esoteric settings, including comparisons of rhetoric 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).

science, researchers are often more interested in similarity comparisons generated using a more specific conceptualization and applied to longer texts. Gathering training data for similarity comparisons across 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, which is rarely practical without some degree of automation. As a result, to our knowledge no existing study has obtained the necessary training data to conduct this kind of validation exercise.

2.2 Defining the Problem

To build intuition and undergird the experimental results that we present later in this paper, we introduce a formalization of the problem described in the previous section. Suppose that a researcher is interested in measuring the pairwise similarity between documents in a some corpus $\mathcal{D}$, $|\mathcal{D}| = n$. Define the "true" similarity between the i^{th} and j^{th} constitution as $\mathbf{Y}_{ij} = f(\mathbf{x}_i, \mathbf{x}_j)$, with $\mathbf{Y}$ the matrix of pairwise similarity values, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n]^t$ an $n \times m$ matrix of m features corresponding to $\mathcal{D}$, and $f(.)$ a similarity function relating a pair of document-level feature vectors (e.g. cosine or Euclidean). In our applied examples, we construct $\mathbf{X}$ using human-coded information and select an $f(.)$ that matches $\mathbf{X}$'s constraints. However, in other situations, a researcher might simply ask human coders to record $\mathbf{Y}$ directly, without formally defining $\mathbf{X}$ or $f(.)$.

In either case, our problem of interest in this paper is to construct a set of machine-generated similarity values that approximate $\mathbf{Y}$ as closely as possible, without allowing the machine to directly observe $\mathbf{X}$ or $f(.)$. As a result, we must instead learn a function $g(\mathbf{z}_i, \mathbf{z}_j) \approx \mathbf{Y}_{ij}$, with $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ...\mathbf{z}_n]^t$ an $n \times k$ matrix of features generated using the raw texts of $\mathcal{D}$ and $g(.)$ a function of a given pair of textual feature vectors.[4] Importantly, note that the constraints on $\mathbf{X}$ and $\mathbf{Z}$ need not correspond;

[4] An alternative approach would be to use $\mathbf{Z}$ to estimate $\mathbf{X}$ directly, which we could then use

while human-coded variables are frequently discrete or bounded, machine-generated textual information is often continuous, creating additional difficulties when selecting an appropriate $g(.)$.

Framing the problem in this fashion raises a number of immediate challenges. Most notably, neither $\mathbf{Z}$ nor $g(.)$ are predefined, and must be constructed based on the problem of interest. Since $\mathbf{X}$ and $\mathbf{Z}$ are both generated using $\mathcal{D}$, as long as the underlying features in $\mathbf{Z}$ are well-chosen we might reasonably assume $\mathbf{X}$ and $\mathbf{Z}$ contain similar information, but the specific relationship between the two feature sets (and any pairwise similarity values constructed using them) is unclear. For example, several human-generated features in $\mathbf{X}$ may be represented using a single feature in $\mathbf{Z}$, or vice versa. Or, if $\mathbf{X}$ is discrete, the best-performing $g(.)$ function may exhibit a threshold effect, in which changes in the textual features $\mathbf{Z}$ affect the predicted similarity value in a logarithmic or other non-linear fashion. For this reason, in our experiments and applied examples we find that flexible, black-box prediction models such as random forests perform best as our $g(.)$ function, but we discuss other approaches in supplementary materials.

3 Similarity Among Constitutions

3.1 The Domain of Inquiry: National "Constitutions"

To explore the problem of similarity comparison in an applied setting, we draw on the study of national constitutions, surely one of the more central sets of texts in political science. Scholars interested in the diffusion of ideas have long studied the intellectual history of these texts, and their patterns of change across space and time. One hears claims regarding constitutional diffusion even about esoteric texts:

to calculate $\mathbf{Y}$. This approach represents a plausible alternative to our existing approach and an interesting direction for future research (though we note that this strategy would involve estimating a substantially larger number of parameters than our existing approach).

> 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 on textual similarity comparisons, which scholars deploy to evaluate hypotheses related to the spread of ideas across jurisdictions. As we discuss in the previous section, similarity scores for large documents are difficult to acquire in many contexts; fortunately, however, the Comparative Constitutions Project (CCP) makes pairwise similarity comparisons in the constitutional setting relatively easy to acquire. As a result, the CCP data offer a natural testing environment for our purposes.

Two attributes of the CCP's data are particularly convenient for the analyses herein. First, the CCP's authors measure aspects of the constitutional *text* itself, offering a useful reference point for automated, text-based measures of similarity. Second, the CCP's scope is extensive. The CCP collects, cleans, and content-tags constitutional texts for some 600 topics for all founding documents in all countries, offering a rich and substantively significant training set from which to work.

3.2 The Criterion: Inventory Similarity

Our set of documents $\mathcal{D}$ is the set of all in-force constitutions as of 2014, as identified by CCP. As mentioned previously, CCP contains extensive information on the inventory of topics included in any two constitutions (e.g., whether the constitution mentions a central bank, or addresses the accession of new territory). We use these data to create our human-generated feature set $\mathbf{X}$, which consists a set of binary variables denoting the presence or absence of 70 such topics in each constitution. We exclude from this list of items many *sub*-topic questions that should be understood as making 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.[5]

To generate similarity values from these features, we employ a thematic approach that we term *inventory similarity*.[6] 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. Or one could focus on presentational elements – those of word choice, organization, or tone. Or, one could even measure the extent to which two constitutions re-use subsequences of text drawn from one another (see, e.g., Wilkerson et al. (2015); Linder et al. (2016); Burgess et al. (2016)). But we can also ask, more fundamentally, whether two constitutions *address* the same topics. Here, we focus on this last idea. If two constitutions address a similar "inventory" of ideas, we might infer that the authors of each founding document were similarly concerned with those topics, or perhaps influenced by one another. Though each document might address these issues differently, a shared conceptual inventory suggests a shared set of underlying concerns and challenges.

We compute similarity scores from this set of 70 binary attributes using Jaccard's (1912) similarity coefficient, which has some helpful properties given that texts in the corpus are of different lengths. The Jaccard formula takes the following general

[5]Elkins et al. (2009) use the same set of topics as a measure of *scope*, a related concept.
[6]Or, synonymously, *thematic* similarity.

form[7]:

$$f(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum min(\mathbf{x}_i, \mathbf{x}_j)}{\sum max(\mathbf{x}_i, \mathbf{x}_j)}$$

with $min(.)$ and $max(.)$ the element-wise minimum and maximum functions. Informally, Jaccard similarity 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.

Consider the second property. 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.

3.3 Validity of the Criterion

The central comparison of measures in a criterion approach to validity rely upon the assumption is that the criterion measure itself is valid. Our understanding of the underlying data provides some confidence in this respect. As we describe above, the

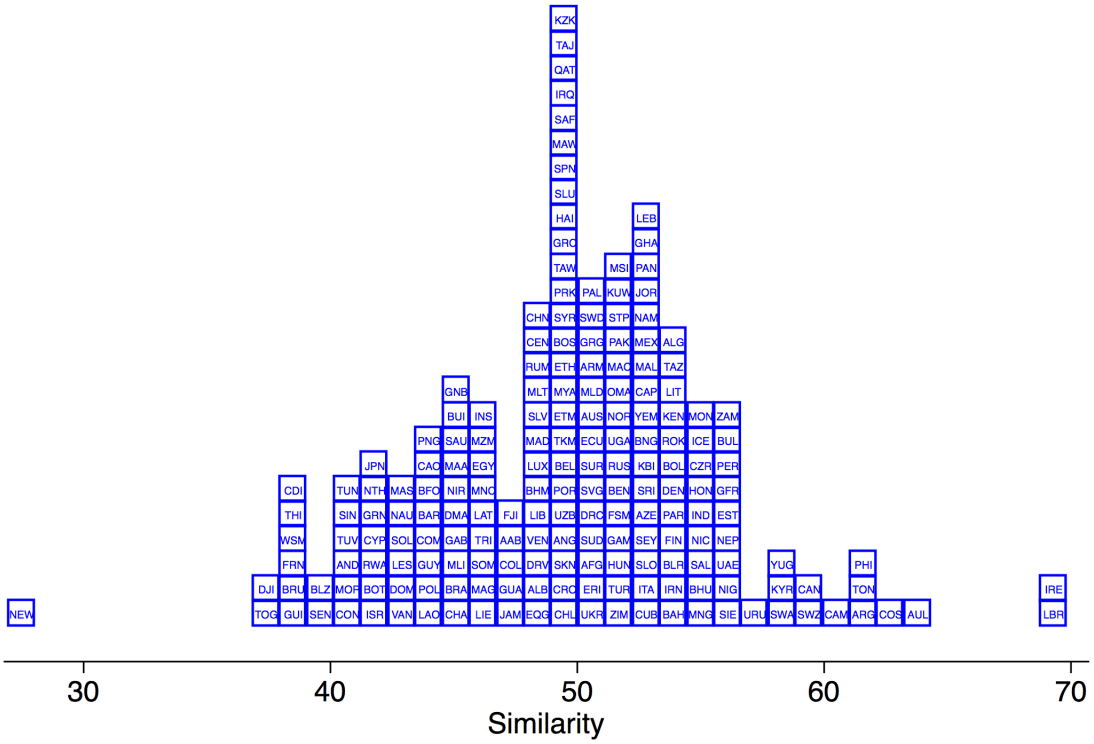
[7]This formulation of the Jaccard coefficient actually describes the generalized version of the measure, which extends to any set of non-negative features. By contrast, Jaccard's original formulation – expressed in set notation as $J(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cap \mathbf{x}_j}{\mathbf{x}_i \cup \mathbf{X}_j}$ – applies only to binary feature vectors. We give the generalized version in-text in order to avoid cumbersome switching between vector and set notation.

CCP data was collected from a close reading by multiple coders of constitutional text and coded for a large set of topics with the intent of measuring thematic similarity. Also, the codings seem to exhibit a high degree of inter-coder reliability (Melton et al. 2013), which is reassuring given the sometimes polemic interpretive battles regarding constitutional clauses.

For those mired in the genre of written constitutions, the similarities produced using our criterion measure exhibit some face validity. Across the full dataset of 193 constitutions ($n = 18,528$ unique dyads), the mean similarity score is 0.60 ($s.d. = 0.10$) with a range of 0.19 to 0.96, suggesting a moderate degree of similarity between randomly-selected constitutions with substantial variation across the dataset. 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* (to us) 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 similarity scores to the U.S., and identify particular cases. 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. The texts of Argentina and Costa Rica – two of the oldest in Latin America – are both 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 exhibit face validity, though there are interesting variations in similarity well-worth investigating (which is, indeed, the point of constructing the measure).

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



4 Candidate Measures of Textual Similarity

Having established a criterion measure of thematic similarity, we follow a set of standard practices in order to produce a comparable set of machine-generated candidate measures. Our process involves two steps. First, we use a series of unsupervised feature extraction schemes (TFIDF, LSI, LDA, STM, and word2vec) to extract candidate feature vectors $\mathbf{Z}$ from each constitutional text. Second, using these feature vectors, we use a series of pairwise similarity functions $g(\mathbf{z}_i, \mathbf{z_j})$, selected in both unsupervised and supervised fashions.

4.1 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 political science (Grimmer and King 2011; Lucas et al. 2015). The earliest versions of these algorithms use simple linear algebra-based dimensionality reduction techniques to extract latent dimensions from a term-document matrix (e.g. singular value decomposition as used in Latent Semantic Indexing (LSI) (Dumais et al. 1988; Deerwester et al. 1990)). More recent approaches, by contrast, usually employ a generative approach.[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 such "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),

[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.

variable dimensionality (Blei and Jordan 2004; Teh et al. 2012), and many other possibilities besides. Some authors (e.g. Mikolov et al. (2013); Le and Mikolov (2014)) have also proposed deep-learning approaches that generate vector representations 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 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 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.

We select this particular set of feature extraction approaches for two reasons. First, the four primary approaches are well-supported in various programming languages and are frequently used both within and beyond Political Science.[9] As a result, all four represent natural choices for applied users. Second, from a performance standpoint, these models are well suited to the task we outline in this paper. Since our criterion "inventory similarity" metric is determined by the shared presence/absence of certain high-level themes in constitutional texts, latent variable approaches based on word co-occurrence are well-suited to extract relevant information from each document.

Before fitting our models, we use a two-step preprocessing approach (summarized in Table 1). First, we subdivide each constitutional text into a number of constituent documents. As described in §3.1, constitutions are long and thematically diverse. For

[9]Sparse matrix factorization or sparse LDA (see, e.g., Zhang et al. (2013); Ming et al. (2014)) offer an alternative feature extraction framework, which may help address some of the problems created by the conceptual disconnect between $\mathbf{X}$ and $\mathbf{Z}$ discussed in §2.2. Since these approaches are not frequently used in applied work, we do not test them here, but these approaches represent a promising direction for future work.

co-occurrence-based models like LDA, STM, and LSI, these documents need to be subdivided into thematically-coherent units in order to produce coherent and useful topics. Fortunately, constitution-writers generally organize their texts along thematic lines using some formalized hierarchy (e.g., Articles and Sections in the US Constitution). For LDA, STM, and LSI, we leverage these internal organization schemes and simply segment constitutions into paragraphs. Since word2vec relies on word ordering and localized contextual information, we instead subdivide documents into sentences when training this model.[10]

Second, we pre-process 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 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)). In our topic modeling specifications we drop only stopwords[11] and punctuation; we do not stem terms, and we retain all terms longer than three characters and contained in at least 10 documents (representing $\approx 0.01\%$ of the dataset).

As shown in Table 1, the pre-processing 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 specification for two reasons. First, word2vec is substantially less demanding to estimate than are most of the other approaches we present (particularly STM), and requires fewer pre-

[10] As identified by the pretrained Punkt sentence tokenizer contained in NLTK (http://www.nltk.org/).

[11] As defined by NLTK (http://www.nltk.org/)'s stopword list.

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

Model	Unit	Pre-processing	Hyperparameters
TF-IDF	paragraphs[a]	(1) lower-case; (2) punctuation, stopwords, tokens ≤ 3 characters, tokens in $\leq$ 10 documents removed; (3) documents $\leq$ 5 tokens removed	n/a
LSI	paragraphs[a]	Same as TF-IDF	$\{20, 50, 100, 150, 200\}$ topics
LDA	paragraphs[a]	Same as TF-IDF	$\{20, 50, 100, 150, 200\}$ topics; MALLET hyperparameter optimization
STM	paragraphs[a]	Same as TF-IDF	$\{20, 50, 100, 150, 200\}$ topics; constitution dummies and years since 1789 (spline) used as covariates
word2vec	sentences[b]	(1) lower-case; (2) punctuation, tokens $\leq$ 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), with default hyperparameter optimization settings, optimized at 20-document intervals Wallach et al. (2009). STM estimated via the STM R package (Roberts et al. 2014), with a spectral initialization used.

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

processing steps in order to become computationally feasible. Second, these differing standards offer our results some robustness against pre-processing choices. As shown in the following section, the word2vec-based similarity values we generate perform somewhat worse in the unsupervised setting than LDA and STM; however, all models perform similarly in our supervised experiments. We view this finding as suggestive (though not conclusive) evidence that our results are robust to the range of pre-processing standards we test in this paper.

To generate a final set of feature vectors, 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 $\mathbf{z}_i$ is defined as:

$$\mathbf{z}_{ik} = \frac{1}{\sum_{j=1}^{N_i} n_{ij}} \sum_{j=1}^{N_i} n_{ij} p_{ijk}$$

Where j indexes the N_i paragraph/sentence-level feature vectors associated with the i^{th} constitution, and k indexes features. Within each constitution, n_{ij} represents the token count (after preprocessing) of the j^{th} article/sentence associated with the i^{th} constitution, and p_{ijk} gives the feature value of the k^{th} element of the j^{th} paragraph/sentence-level feature vector within the i^{th} constitution. In words, $\mathbf{z}_{ik}$ therefore represents a normalized feature vector for each constitution, constructed by summing over the term-level feature values constructed using each feature extraction approach.

We emphasize that these preprocessing steps, parameter settings, and models are not the only plausible feature-extraction approaches. Other featurization setups – such as doc2vec (Le and Mikolov 2014), sense2vec (Trask et al. 2015), or Grimmer and King (2011)'s ensemble approach – also represent reasonable choices. However, for practical reasons, we do not test all available specifications. Instead, we suggest, simply, that the options we test here are both plausible and represent a reasonable se-

lection of commonly-used approaches, and therefore offer a useful baseline for applied work.

4.2 Similarity Estimation

These approaches leave us with a set of feature vectors for each constitution, which we then use to calculate text-based similarity scores for each dyad. As discussed above, since text-based similarity scores often rely upon human-generated training data, one of our primary issues of interest in this project is to compare learner performance across training set sizes. We therefore construct scores based on both unsupervised and supervised approaches for each set of feature vectors we construct, and vary training size in each case.

To generate unsupervised similarity values, we follow a simple procedure. For each feature extraction approach, we take the feature vectors for each pair of constitutions, and calculate a distance measure between the two vectors appropriate to the constraints imposed by the feature extraction approach. 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 the feature vectors for each constitution, defined as

$$\begin{aligned} g_H(\mathbf{z}_i, \mathbf{z}_j) &= \frac{1}{\sqrt{2}} \sqrt{\sum_{k=1}^{K} (\sqrt{\mathbf{z}_{ik}} - \sqrt{\mathbf{z}_{jk}})^2} \\ &= \frac{1}{\sqrt{2}} ||\sqrt{\mathbf{z}_i} - \sqrt{\mathbf{z}_i}||_2 \end{aligned}$$

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

$$g_C(\mathbf{z}_i, \mathbf{z}_j) = 1 - \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{||\mathbf{z}_i||_2 ||\mathbf{z}_j||_2}.$$

For the supervised similarity values, we employ a "black-box" approach. In particular, for each feature type and constitution dyad, we concatenate the textual features for each constitution into a $2K$-length vector, which we use as an input into a random forest model (Breiman 2001).[12] In order to respect the dyadic dependence present in our data, we selected our training sets using a two-step procedure. First, we randomly selected a set of training documents, using 30, 45, 60, and 75 documents as our training set sizes. Next, we collected all dyads within this randomly-selected training set, and used those dyads as inputs into our random forest. Finally, we assigned the remaining dyads (i.e. all dyads with one or both members not included in the document-level training set) to our test set, which we use to assess out-of-sample performance. We repeated this process 100 times for each training set size, allowing us to assess variability in performance induced by training set selection. For each model, we grew 500 trees, with the number of randomly-selected candidate variables at each split determined by optimizing out-of-bag error for each training set.[13]

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.

5 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

[12] We also experimented with an approach in which we trained the random forest using the element-wise absolute difference between each constitution's feature vector, $|\mathbf{z}_i - \mathbf{z}_j|$. However, this approach performed slightly worse than our existing setup (see Appendix B for details).

[13] Using the *tuneRF*() function as implemented in the randomForest (https://cran.r-project.org/web/packages/randomForest/index.html) package in R. All parameters not mentioned in-text left at their default values.

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 criterion and candidate measures. In our second test, we examine whether models applied to the criterion and candidate measures produce the same inferences on a series of causal questions. The similarity values we generate largely perform well in this case, returning the same substantive conclusions for most parameter values of interest. Finally, in our third test, we ask whether the candidate measure returns the same rank ordering of cases (and even the same ratio-level positions of cases), as does the criterion 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.

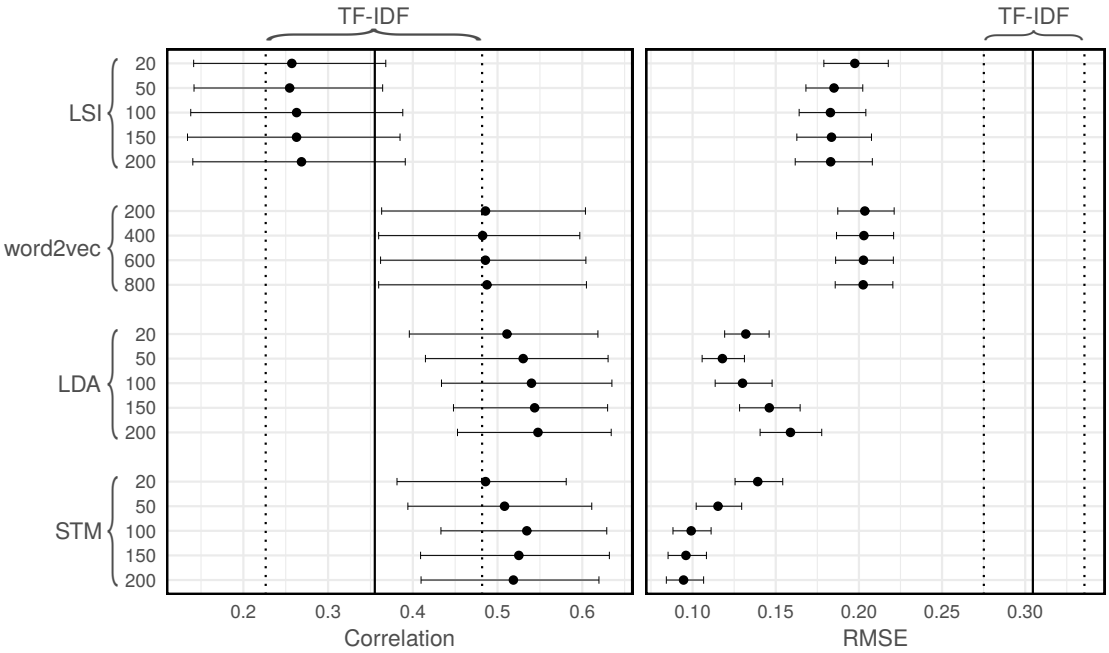
5.1 Test 1: Aggregate Learner Performance

To what degree do the machine-generated measures of similarity replicate the human-generated values? We begin with an evaluation of unsupervised performance. For each feature set, we calculated machine-generated similarity scores as described in §4.2, and assessed the relationship between these scores and the human-generated criterion values in terms of correlation and RMSE. To generate measures of uncertainty, we used a modified block bootstrap procedure. For each bootstrap replicate, we drew a set of countries (with replacement), extracted all dyads within this set, and calculated performance based on these values. Finally, we repeated this process 10,000 times, and reported the 2.5^{th} and 97.5^{th} percentile scores for each feature set.

Figure 2 shows the results of this procedure. By correlation, the three generative approaches (word2vec, STM, LDA) perform similarly, with correlations to the hand-

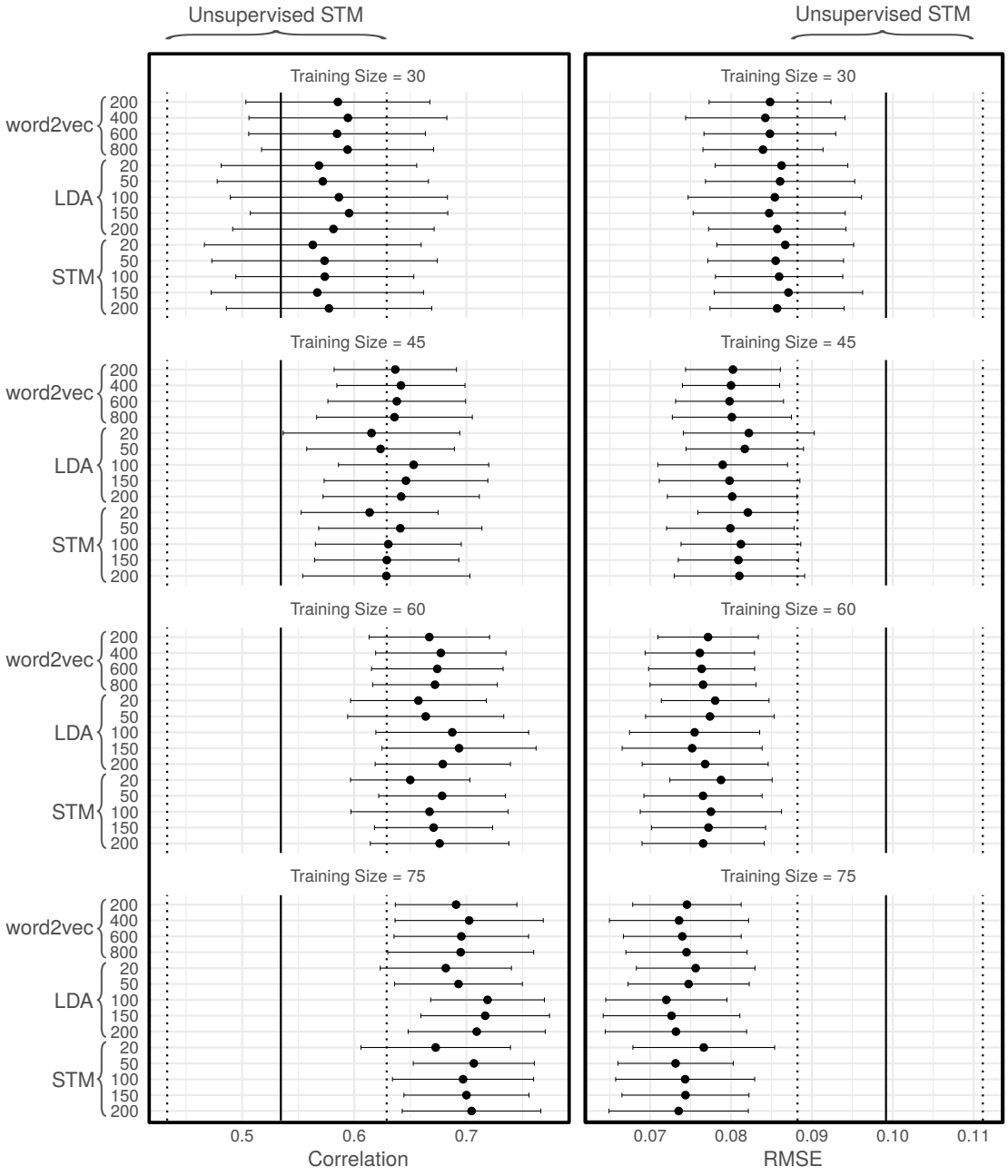
coded data in the $r = (0.45, 0.6)$ range. By RMSE, differences between the models are more noticeable, with similarity scores based on higher-dimensional ($k \geq 100$) STM features performing best. Similarity scores based on simple LDA features are a close second by this metric, with optimal performance at $k = 50$ topics. Interestingly, the addition of covariates via STM only offers a small performance boost relative to the best-performing LDA models. For the purposes of this study, we were interested in simulating a research scenario in which the researcher does not possess a particularly rich feature set; as a result, our only covariates were a spline of the year of the constitution's enactment and a dummy variable indicating the constitution from which a given paragraph was drawn. In Appendix B, we compare our results to those generated using an STM model with a richer covariate set; however, at least for our corpus, these additional covariates do not appear to improve performance.

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



Solid and dashed vertical lines indicate mean correlation/RMSE and 95% confidence interval between CCP targets and similarities generated using baseline TF-IDF features. Dots and solid horizontal lines indicate mean correlation/RMSE and 95% confidence intervals for similarities produced using other feature extraction approaches. Confidence intervals generated using the block bootstrap procedure described in-text.

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 STM_{100} model shown in Figure 2 as a baseline comparison.

Though encouraging, the correspondence between machine and human in these initial analyses leaves appreciable room for improvement. Figure 3 describes the performance of the supervised similarity scores generated in §4.2 compared to the unsupervised STM_{100} features as a baseline. With a training set as small as 45 documents ($\approx 25\%$ of the dataset), our supervised predictions consistently outperform the unsupervised baseline by both correlation and RMSE. By 75 documents ($\approx 40\%$ of the dataset), these improvements are striking; at $n = 75$ training documents, 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.

Importantly - and in contrast with our unsupervised tests - the choice of feature extraction approach and dimensionality parameter appears to have little impact on our supervised results. This invariance is present across all training set sizes, and represents a heartening finding. In many modeling settings, tuning 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. Fortunately, with even a small ($n = 30$) training set, our results are essentially unaffected by the choice of model or dimensionality parameter. Based on these results, moving to a supervised similarity estimation approach with even a small training set offers a substantial payoff, providing researchers some reassurance that their results are largely invariant to choice of dimensionality parameter or feature extraction approach.

5.2 Test 2: Causal Inference across Criterion and Candidate Measures

For enthusiasts of automated content analysis, the high levels of aggregate correspondence between human and machine measures will be encouraging. However, for applied researchers, a more meaningful criterion is the extent to which analy-

ses conducted on human- and machine-generated similarity measures yield the same causal conclusions. Consider, in this spirit, some basic expectations regarding *isomorphism* in constitutional design. A robust finding in comparative constitutional studies (e.g. Elkins et al. (2013)) is that the drafter's context – in particular geography and era – matters enormously. Some of these analyses suggest that we can explain as much as half of the variation in constitutional content if we know *where* and *when* a given document was written (Cheibub et al. 2014). We revisit these 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.[14] Since constitutions written in close temporal proximity likely reflect similar constitution-writing trends and concerns, 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 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 implicit left-out category). However, there is likely to be substantial

[14]We test two other specifications of this variable in Appendix C.2. Both approaches yield nearly identical conclusions to the one described in-text.

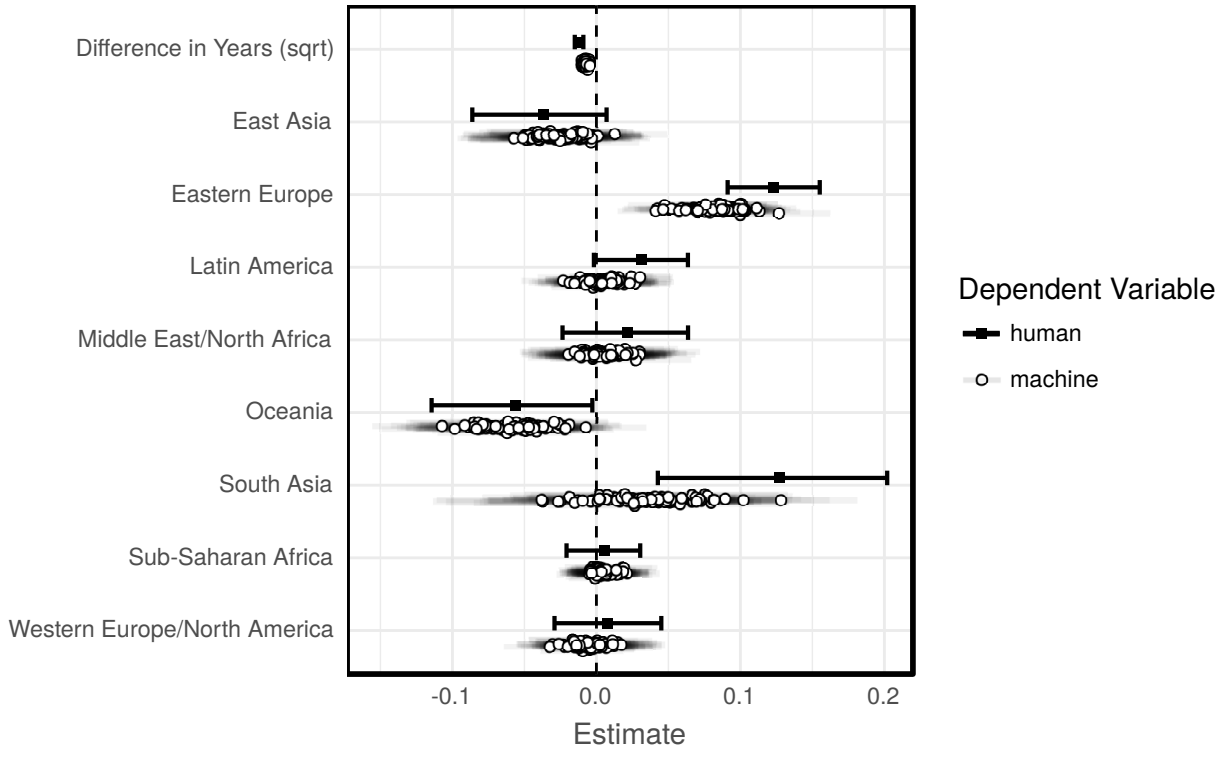
variation within these categories. For example, we expect constitutions from Oceania to cluster least, since countries in this region share relatively few cultural similarities. By contrast, we expect constitutions from Latin America and Eastern Europe to be more similar than the baseline, since these 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). Under this procedure, we simultaneously permute the rows and columns of the matrices of dependent and independent variables included in the regression model. These repeated permutations preserve the dependency structure within the dependent variable matrix while removing dependencies between the dependent variable and the independent variables. Under most conditions, this null hypothesis distribution allows us to properly account for unmodeled dependencies left unaddressed by standard approaches (see Cranmer et al. (2017) for further discussion).

Figure 4 reports the results of a linear model estimated by OLS (with QAP-corrected confidence intervals) in which we predict constitutional similarity using the variables described above (See Appendix C.1 for numerical coefficient estimates). The dependent variables in these models are the human-generated criterion values (as obtained from CCP) and the machine-generated candidates produced using a random forest model estimated on LDA_{100} features with 75 training documents.[15]

[15] As noted in §4.1, in the supervised context similarity values based on LDA, STM, and word2vec perform similarly at all parameter settings we examine. We focus on LDA features in this section because of LDA's simplicity and generalizability relative to both word2vec and STM, and because similarities based on LDA_{100} features perform slightly (though not significantly) better than all other models by both RMSE and correlation.

Figure 4: Linear model coefficient estimates generated using human and machine-produced 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 candidate values are overlaid and jittered, and confidence intervals are faded. Intercept omitted for readability.

For the machine-generated candidate values, we fit a separate linear model for each of 100 train/test splits, and plotted the estimated coefficients for each model. In these models, for dyads contained in the training set we substituted human-generated similarity values for in-sample random forest predictions. Our rationale for this choice is drawn from the applied context our validity tests are intended to approximate. If a researcher has human-generated gold-standard values available for some proportion of her sample during similarity estimation, this same set should also be available during inferential modeling. However, we revisit the implications of this choice below.

Beginning with the model estimated on human-coded criterion 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 exhibit a similar pattern ($b = 0.127$). 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.

Again, our focal question in this section is whether we would reach the same conclusions using the machine-generated candidate as we would with the human-generated criterion values. Returning to Figure 4 suggests that the answer to this question is a qualified yes. Unsurprisingly, since our estimates contain some degree of measurement error, most coefficients estimated using the machine-generated similarity values are attenuated relative to their human-generated counterparts. Nev-

ertheless, most model replicates returned the same conclusions regarding coefficient significance and sign as the model trained on the criterion values. Using $p = 0.05$ as a significance threshold, 87% 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.[16] All coefficients besides the South Asia coefficient returned the same substantive conclusions in at least 80% of replicates, with the year-difference coefficient and the Eastern Europe, Middle East/North Africa, and Sub-Saharan Africa coefficients returning the same conclusions in 100% of replicates. By contrast, the South Asia coefficient was by far the worst performer, returning a positive and significant estimate in only 17% of replicates.

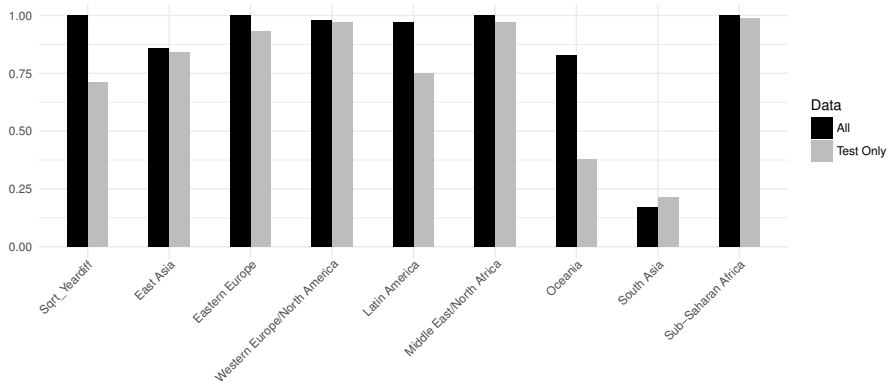
As mentioned previously, in Figure 4 we estimate models using human-generated training data for training-set dyads. We believe that this approach offers the most useful results for applied researchers, but methodologically-oriented readers might reasonably object that this approach produces a baseline that is too optimistic. To address these concerns, we reestimated a set of models using only those dyads with neither country contained in the training set ($n = 6903$ with a training set of 75 countries), using the same 100 train/test splits used in Figure 4.[17] The results of this experiment are shown in Figure 5.[18] Unsurprisingly, most coefficients perform somewhat worse in the test set-only models than those that include all available data, with variables whose "true" coefficients that are closest to the statistical significance

[16]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.

[17]Note that, since this approach excludes all dyads with at least one member drawn from the training set, the underlying dataset (and resulting "criterion" human-generated coefficient estimates) changes for each train/test split. As a result, the "correct" substantive conclusion for each coefficient (i.e. the coefficient drawn when examining a model estimated using human-generated similarity values) may differ for each replicate.

[18]In two train/test splits, the test set contained no countries drawn from the South Asia region, leading this coefficient to be excluded from the model. As a result, the sample size for this coefficient is 98 rather than 100.

Figure 5: Proportion of coefficients that return the same substantive conclusion for human- and machine-generated models, using all data and test-only dyads.



Proportion of model replicates which returned the same substantive conclusions for all covariates across 100 train/test splits. Values for models estimated using all data and models restricted to dyads with both countries are drawn from the test set are compared.

threshold suffering the largest performance hits. However, even in this more stringent test, some 76% of coefficients return the same substantive conclusions when estimated using test set-only dyads, suggesting that our estimates still perform reasonably well in this set.

5.3 Test 3: Unit-Level Descriptive Inference

In our final 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

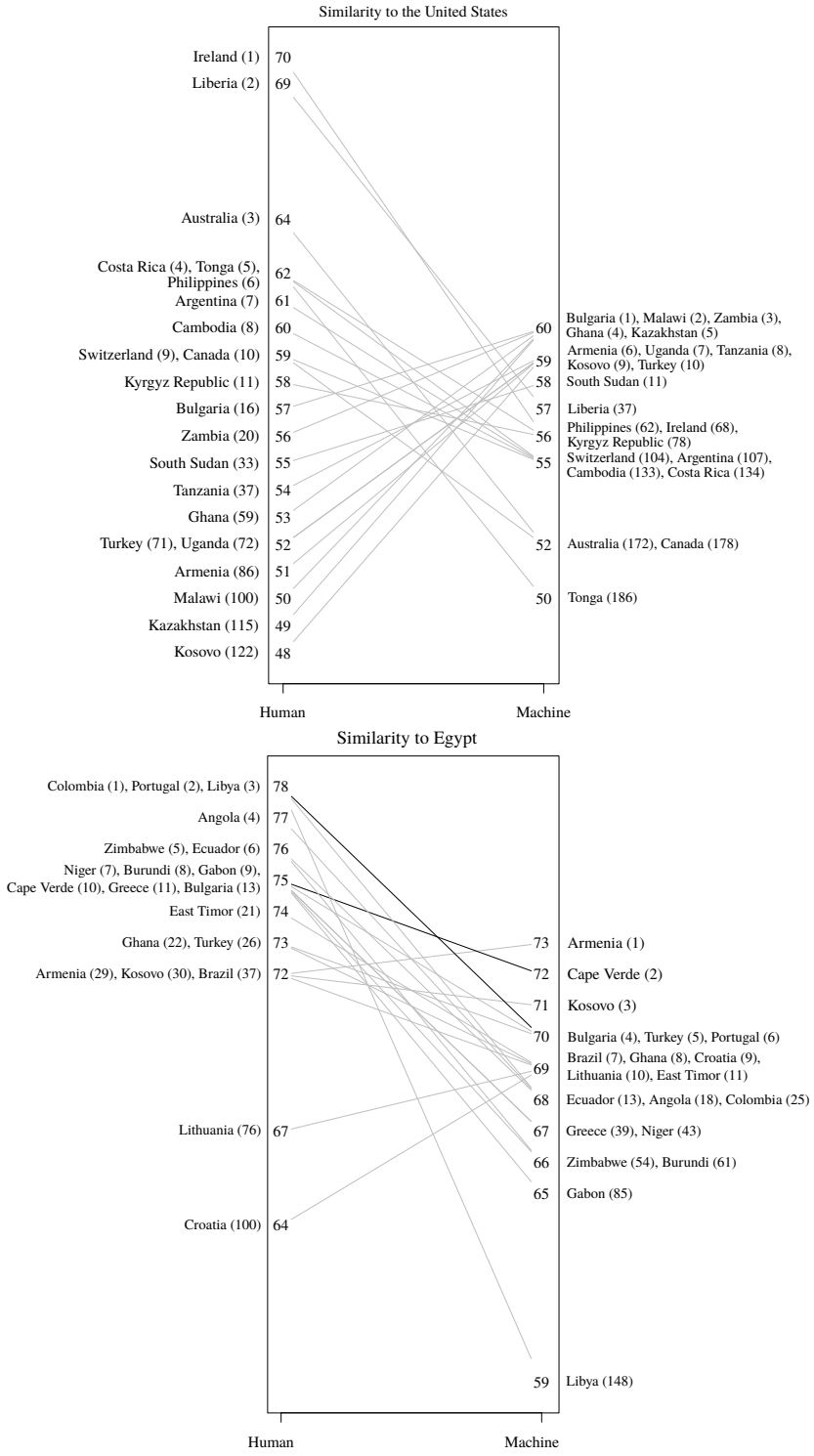
more informal analyses. Consider a customer searching for a top-rated good or service. A standard, almost reflexive, exercise is to compare two lists of recommendations. If one list does not match another, especially if one is particularly trustworthy, we might wonder about the other's validity or reliability. Put differently, we might suspect that the measures are either tapping different concepts, or that one or both contain substantial unsystematic error. 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_{100} features.[19] As in previous sections, for each replicate we randomly select 75 documents ($\approx 40\%$ of the dataset), and 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 values. We then repeat this process 100 times, and compare results

By way of illustration, Figure 6 depicts a "top-ten" analysis based on a randomly-chosen train/test split for the United States (whose Constitution will be familiar to many readers) and for Egypt (a somewhat typical case, in terms of cross-measure correspondence). In both cases, the rank-ordering of similar countries appears appreciably different between the measures. In the case of the United States, there is

[19] As in our previous test, we focus on LDA features in this section because of LDA's simplicity and generalizability relative to both word2vec and STM, rather than any performance differences between the various models.

Figure 6: 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.

no overlap between the lists, and some rather unsettling differences. For example, Australia is in the top ten of the human list, but in the *bottom* ten in the machine list. In the Egyptian case, the two top-ten lists share only two cases (Portugal and Cape Verde). One way to summarize the difference in lists is in terms of rank-order differences between cases on the lists. On average, the per-observation difference in rank for the United States similarity set is 43.5. Egypt, which again is more typical, has an average difference in rank of 29.1, 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 21.3 (the case of Georgia). The average absolute difference in rank across all countries and train/test splits is 30.8, 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 exhibited by the Bahamas, with an average of 6.2 shared members across all 100 train/test splits. 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 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? We might think of the reviews as ones written by critics with highly distinct sets of tastes. 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.

6 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 validity of these measures, especially as applied to long documents. In this paper, we evaluate the validity and utility of automated similarity measures in the genre 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*, a measure of thematic similarity that we calculate for each pair of in-force constitutions. We adopt a criterion-validity testing framework, in which we assume this human-interpreted measure of similarity to be valid and reliable and, therefore, useful as a benchmark. We conduct three specific evaluations, which offer varying perspectives on the degree of criterion validity of the automated measures.

Our first test is one of aggregate correspondence. We find that each of the four approaches to automated similarity under evaluation is highly correlated with our criterion measure. We conclude that the various measures all tap a common dimension of thematic similarity. This correspondence is our central and baseline finding, and one that should (in our view) inspire confidence in such techniques. In our tests, those scores 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, in our supervised experiments these results were invariant to the choice of model or parameter settings at all training set sizes, suggesting that supervised methods can adapt to the differences between

various training sets given very little training data.

In a second set of tests, we incorporate the various measures in a set of regression analyses of constitutional similarity, in order to determine whether the machine-generated measures yield the same causal inferences as their human-generated counterparts. In our test of "contextual hypotheses," inspired by theories of diffusion, 87% of coefficients estimated based on the machine-generated candidate measure returned the same conclusions as those estimated using the human-generated criterion. 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 criterion measure. One can view this conservatism – likely due to differences in random measurement error – benignly. Researchers using automated measures in regression models can have some confidence that the effects they do observe are present in the human-generated target data.

Finally, we subjected the measures to a set of unit-level descriptive tasks – identifying the top-ten most similar constitutions to a given document – in order to illuminate the degree of correspondence among the measures more clearly. These tests suggest the limits of automated measures, at least for some descriptive purposes. A comparison of top-ten lists across measures – a typical exercise in recommendation services – yields very few common cases across the measures. For projects requiring a high degree of rank-order correspondence between the criterion and subject 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 for automated text practitioners. Given a moderate-sized training set, analyses conducted using text-based similarity scores can reproduce 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 with large quantities of long docu-

ments. In these situations, machine-generated approximations offer a useful way for researchers to test hypotheses regarding the 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.

Burgess, M., Giraudy, E., Katz-Samuels, J., Walsh, J., Willis, D., Haynes, L., and Ghani, R. (2016). The legislative influence detector: Finding text reuse in state legislation. In *KDD*, pages 57–66.

Cheibub, J. A., Elkins, Z., and Ginsburg, T. (2014). Beyond presidentialism and parliamentarism. *British Journal of Political Science*, 44(3):515–544.

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.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Linder, F., Desmarais, B. A., Burgess, M., and Giraudy, E. (2016). Text as policy: Measuring policy similarity through bill text reuse.

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.

Melton, J., Elkins, Z., Ginsburg, T., and Leetaru, K. (2013). On the interpretability of law: Lessons from the decoding of national constitutions. *British Journal of Political Science*, 43(2):399–423.

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.

Ming, F., Wong, F., Liu, Z., and Chiang, M. (2014). Stock market prediction from wsj: text mining via sparse matrix factorization. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 430–439. IEEE.

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.

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*.

Trask, A., Michalak, P., and Liu, J. (2015). sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*.

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.

Wilkerson, J., Smith, D., and Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, 59(4):943–956.

Zhang, A., Zhu, J., and Zhang, B. (2013). Sparse online topic models. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1489–1500. ACM.