An Evaluation of Measures of Textual Similarity

Robert Shaffer[1] **and Zachary Elkins**[1]

[1]University of Texas, Austin

This manuscript was compiled on March 26, 2018

Understanding the similarity of two texts can be enlightening and useful. Unfortunately, human-generated similarity comparisons are labor-intensive to produce. Approaches based on machine learning techniques are more scalable, but their performance is poorly understood, especially as applied to the kinds of long, thematically diverse documents common in many applications. We leverage a unique and substantively important dataset of human-coded national constitutions in order to fill this gap. We begin by comparing the scores produced using a series of plausible unsupervised feature-extraction approaches against a parallel set of human-coded similarity comparisons between constitutional texts. We then use the best-performing feature extraction approaches as inputs into a supervised scheme, which we demonstrate offers a substantial performance boost even at small training set sizes. Finally, we use our estimated similarity scores to test a series of hypotheses regarding spatial and temporal diffusion of ideas in national constitutions, and find that models estimated using our automated similarity scores replicate the substantive conclusions drawn using their human-generated equivalents.

Machine Learning|Text Analysis|Political Science

Measuring the similarity of cases is basic to scientific inquiry (1, 2). 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 media researcher might compare statements made by political candidates in order to understand the proximity of their policy positions.

Both of these examples involve comparisons 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. Modeling approaches designed to extract information from these data sources have become popular in the social sciences in digital humanities in recent decades.* However, it is often unclear how information extracted by such techniques relates to human intuition, and opportunities to validate automated similarity comparisons against a human-generated gold standard are rare.

To explore this measurement challenge, we leverage the Comparative Constitutions Project's original hand-coded data on the content of national constitutions (3). Because of this dataset's scope and substantive importance, it offers a unique opportunity to validate and compare various automated methods of similarity comparison against a human-generated baseline. The best-performing method we test performs remarkably well even with small ($\approx 40\%$) training sets, with an out-of-sample correlation to human-generated similarity scores of $r \approx 0.7$. The approaches we test perform similarly well in an applied examination of study patterns among national constitutions, offering further reassurance for applied researchers.

*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.

1. Textual Similarity: Promise and Challenges

A. 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 (4–7), or for similarity in rhetoric used by campaigns at different points in time (8).

Unfortunately, studies of this kind face substantial practical constraints. Since similarity comparisons involve pairwise evaluations, human-generated similarity scores are resource-intensive to produce. Automated approaches to this problem are more appealing, but opportunities to validate machine-generated similarity scores are rare. In the computer science and statistical settings, measurement tasks of this sort are generally conducted on short excerpts using an abstract notion of similarity. By contrast, social scientists are often more interested in similarity comparisons generated using a more specific conceptualization applied to a dataset consisting of 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 possible without automation. As a result, to our knowledge no existing study has obtained the necessary training data to conduct this kind of validation exercise.

Significance Statement

Similarity comparison is a basic scientific task. Both expert researchers and members of the public employ similarity comparisons to explore datasets, cluster cases, and test theories. Comparisons of this kind are particularly relevant when examining textual data, a paradigmatic high-dimensional data type that resists simple statistical summarization or analysis. Unfortunately, human-generated similarity comparisons between texts are labor-intensive to produce, and the performance of their machine-generated equivalents is poorly understood. In this paper, we leverage a unique hand-coded dataset of national constitutions to assess performance of various automated textual similarity estimation methods. The best-performing approaches we examine offer strong performance both in aggregate and as used to test important research questions, offering guidance to researchers across a variety of substantive domains.

B. Formalizing the Problem. To build intuition and undergird our experimental results, 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 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. Importantly, note that the constraints on $\mathbf{X}$ and $\mathbf{Z}$ need not correspond; 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 $\mathbf{X}$ and $\mathbf{Z}$ should 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, context-agnostic prediction models such as random forests perform best as our $g(.)$ function, but we discuss other approaches in supplementary materials.

2. Similarity Among Constitutions

A. 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. 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:

> 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 (9).

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 and content-tags constitutions for some 600 topics for all founding documents in all countries, offering a rich and substantively significant training set from which to work.

B. The Criterion: Inventory Similarity. Our set of documents $\mathcal{D}$ is the set of all in-force constitutions as of 2014, as identified by the CCP. As mentioned previously, the 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.[†] Using this dataset, we compute a Jaccard similarity value for each document pair, defined as:

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

with $min(.)$ and $max(.)$ the element-wise minimum and maximum functions. Jaccard similarity is particularly attractive in this setting because it avoids inflating the similarity of short texts. 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 speak very differently. They simply choose not to speak much. Naively-selected metrics (say, a procedure which simply counted the number of elements with the same value) would likely produce a very high similarity between these two documents, which is not desirable in this setting.

For those mired in the genre of written constitutions, the similarities produced using this approach 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. Many of the highest-similarity pairs are drawn from countries in more culturally homogenous regions, such as Oman and Qatar (0.90) or Serbia and Montenegro (0.91). By contrast, some of the least similar include Brunei and Austria (0.21) and New Zealand and Indonesia (0.24). Focusing on the United States - a better-known example for many readers - the most similar constitutions to that of the U.S. 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 (10, 782-784).

[†] We exclude from this list of items many *sub*-topic questions that should be understood as making refined distinctions between constitutions, as well as a collection of items that are either very rare or very common. (3) use the same set of topics as a measure of *scope*, a related concept.

3. Candidate Measures of Textual Similarity

A. Feature Extraction. Like most studies focused on textual data, we begin our similarity estimation procedure by converting our textual data into a series of lower-dimensional features. Though many feature extraction options are available, for the purposes of this study, we focus on four approaches: specifically, Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA) (as implemented in (11)), the Structural Topic Model (STM) (12), and word2vec (13). We also include similarity scores generated from term frequency-inverse-document frequency (TF-IDF)-weighted word count vectors as a baseline point of comparison. 5 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 in the social sciences. 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. Co-occurence based models like those we employ in this study work best when trained on thematically-coherent units in order to produce coherent and useful topics. Fortunately, constitution-writers generally organize their texts along thematic lines into Articles, Titles, or some other formalized hierarchy (see supplemental materials for details and examples). For LDA, STM, and LSI, we leverage these internal organization schemes and simply segment constitutions into paragraphs ($n \approx 138,000$). Since word2vec relies on word ordering and localized contextual information, we instead subdivide documents into sentences when training this model ($n \approx 201,000$).[‡]

Second, we pre-process the documents in the subdivided datasets. As recent work has demonstrated, pre-processing choices in unsupervised text analysis settings can have a substantial effect on downstream model performance (14). 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 (14, 15). In our topic modeling specifications we drop only stopwords[§] 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-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 other words, $\mathbf{z}_{ik}$ 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 (18) or sense2vec (19) – also represent reasonable choices. However, for practical reasons, we cannot test all potential specifications. Instead, we suggest that the options we test here are both plausible and commonly-used, and therefore offer a useful baseline for applied work.

B. 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 the latter 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

$$g_H(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{\sqrt{2}} ||\sqrt{\mathbf{z}_i} - \sqrt{\mathbf{z}_i}||_2$$

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

[‡] As identified by the pretrained Punkt sentence tokenizer contained in NLTK.

[§] As defined by NLTK's stopword list.

Table 1. Estimation and pre-processing details

Model	Unit	Pre-processing	Hyperparameters
STM	paragraphs	lower-case; punctuation, stopwords, tokens ≤ 3 characters, tokens in ≤ 10 documents removed; documents ≤ 5 tokens removed	$\{20, 50, 100, 150, 200\}$ topics; spectral initialization; constitution dummies and years since 1789 (spline) used as covariates
TF-IDF	paragraphs	Same as STM	n/a
LSI	paragraphs	Same as STM	$\{20, 50, 100, 150, 200\}$ topics
LDA	paragraphs	Same as STM	$\{20, 50, 100, 150, 200\}$ topics; MALLET hyperparameter optimization
word2vec	sentences	lower-case; 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 (16). LDA estimated via MALLET (11), optimized at 20-document intervals (17). STM estimated via the STM R package (12).

similarity in these cases instead, defined as

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

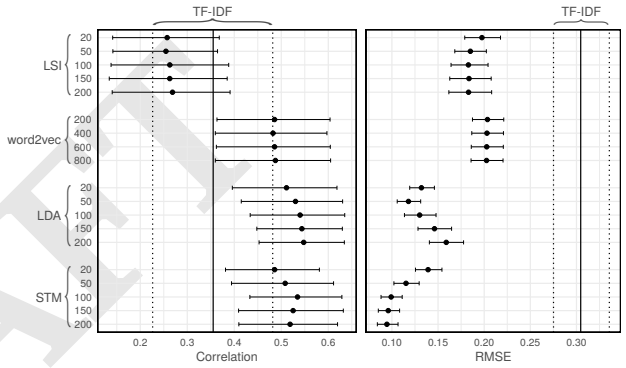
For the supervised similarity values, we begin by concatenating the textual features for each constitution into a $2K$-length vector, which we use as an input into a random forest model (20).[¶] In order to respect the dyadic dependence present in our data, we selected our training sets using a two-step procedure. First, for each training set size we randomly selected a set of training documents. Next, we collected all unique dyads contained within this training set, and used those dyads as inputs into our random forest. Finally, we assigned the remaining dyads (i.e. all dyads containing one or two test-set documents) 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.[‖]

As before, we emphasize that this is not the only approach applied researchers 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.

4. Validating Similarity

A. 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 §3.2, 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.

Fig. 1. Out-of-sample performance for unsupervised similarity values



Dots and solid horizontal lines indicate mean correlation/RMSE and 95% confidence intervals for various feature extraction approaches.

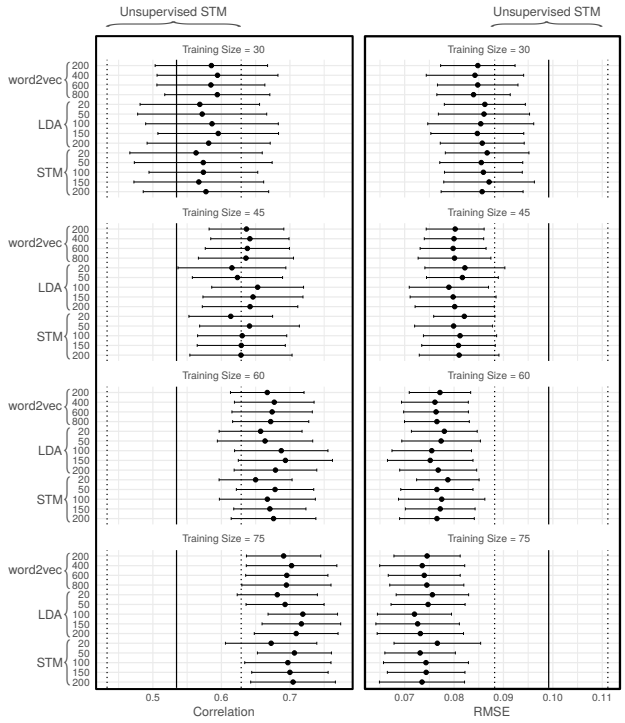
Figure 1 shows the results of this procedure. As measured by out-of-sample correlation, similarity scores based on word2vec, STM, LDA features perform similarly, with correlations to the hand-coded data in the $r = (0.45, 0.6)$ range. By out-of-sample RMSE, differences between the approaches are more noticeable, with STM and LDA-based features performing best. Interestingly, the addition of covariates via STM offers a relatively limited performance boost. 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 set of 193 indicator variables indicating the constitution from which a given paragraph was drawn. In supplemental materials, we compare our results to those generated using an STM model with a richer covariate set; however, these additional covariates do not appear to improve performance.

Though encouraging, the correspondence between machine and human similarities in these initial analyses leaves appreciable room for improvement. Figure 2 describes the performance of the supervised similarity scores generated in §3.2 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 un-

[¶] 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 supplemental materials for details).

[‖] Using the $tuneRF()$ function as implemented in the randomForest package in R.

Fig. 2. Out-of-sample performance for supervised similarity values.



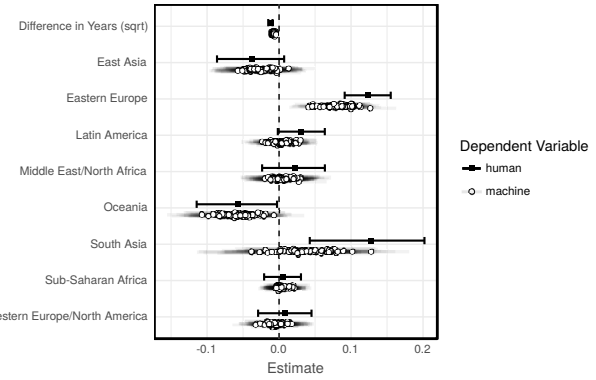
Dots indicate mean out-of-sample correlation/RMSE, estimated from 100 train/test splits. Lines indicate ±2 sample standard deviations.

supervised baseline by both correlation and RMSE. By 75 documents (≈ 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 no 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 represents a troubling aspect of the research process, with few generally-applicable guidelines. 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 for applied work.

B. 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 analyses 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. 21) 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 (22). We revisit these contextual hy-

Fig. 3. Linear model coefficient estimates generated using human and machine-produced similarity values



Critical values and 95% confidence intervals for in-text models, with estimates based on machine-generated data overlaid and jittered.

potheses 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 root absolute difference in the years in which the two constitutions in a given dyad were first enacted.** Since constitutions written in the same period likely reflect similar concerns, we expect the coefficient associated with this term to be negative.

2. *Same region.* A set of dummy variables that indicate whether a dyad's members are drawn from the same geographic region. We expect these coefficients to be positive, since same-region constitutions are likely more similar than those drawn from different regions (the implicit baseline).

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 (23, 24). 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 broad conditions, this null distribution allows us to properly account for unmodeled error dependence introduced by the dyadic data structure (25).

The dependent variables in these models are a blend of the human-generated criterion values (as obtained from the

**We test two other specifications of this variable in supplemental materials. Both approaches yield nearly identical conclusions to the one described in-text.

CCP) and the machine-generated candidates produced using a random forest model estimated on LDA_{100} features with 75 training documents.[††] In particular, for each of 100 train/test splits, we fit a separate model on the training set dyads, and estimated similarity values for all dyads that contain at least one test set member. We then combined these estimated values with their human-generated equivalents for training-set dyads, and estimated a model on this combined vector.

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 in supplemental materials.

Figure 3 reports the results of a linear model estimated by OLS (with QAP-corrected confidence intervals) in which we predict constitutional similarity using the dependent and independent variables described above (see supplemental materials for numerical coefficient estimates). Beginning with the baseline model estimated using human-generated data, 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 and South Asian 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. 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, suggesting that Oceania constitutions are 6 points less similar than the average of different-region constitutions.

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 3 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. Nevertheless, 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.[‡‡] All coefficients besides the South Asia coefficient returned the same substantive conclusions in at least 80% of replicates. By contrast, the South Asia coefficient – the smallest regional group in our dataset – was the worst performer, returning a positive and significant estimate in only 17% of replicates.

[††] 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.

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

5. 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 utility of automated similarity measures in the genre of national constitutions. We find that scores generated using a supervised machine learning approach performed best, with an out-of-sample correlation of $r \approx 0.7$ with the human-generated gold standard at a 40/60 training/test split. In our test of the "contextual hypotheses", 87% of coefficients generated based on human- and machine-generated data returned the same substantive conclusions.

These results should offer substantial reassurance to applied researchers. Given a moderate-sized training set, analyses conducted using text-based similarity scores can reproduce most results produced using labor-intensive hand-coded similarity comparisons. Approximations of this kind therefore offer a useful way for researchers to describe large textual datasets and to test hypotheses regarding the diffusion of ideas in textual form.

1. Santini S, Jain R (1999) Similarity measures. *Pattern analysis and machine intelligence, IEEE transactions on* 21(9):871–883.
2. Tversky A, Gati I (1982) Similarity, separability, and the triangle inequality. *Psychological review* 89(2):123.
3. Elkins Z, Ginsburg T, Melton J (2009) *The endurance of national constitutions*. (Cambridge University Press).
4. Sides J (2006) The origins of campaign agendas. *British Journal of Political Science* 36(03):407–436.
5. Savoy J (2010) Lexical analysis of us political speeches. *Journal of Quantitative Linguistics* 17(2):123–141.
6. Sulkin T (2005) *Issue politics in Congress*. (Cambridge University Press).
7. Klebanov BB, Diermeier D, Beigman E (2008) Lexical cohesion analysis of political speech. *Political Analysis* 16(4):447–463.
8. Hart RP (2009) *Campaign talk: Why elections are good for us*. (Princeton University Press).
9. Brock R, Hodkinson S (2002) *Alternatives to Athens: varieties of political organization and community in ancient Greece*. (Oxford University Press).
10. Law DS, Versteeg M (2012) The declining influence of the united states constitution. *NYUL Rev.* 87:762.
11. McCallum AK (2002) Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.
12. Roberts ME, et al. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4):1064–1082.
13. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality in *Advances in neural information processing systems*. pp. 3111–3119.
14. Denny MJ, Spirling A (2016) Assessing the consequences of text preprocessing decisions. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145.
15. Grimmer J, Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21(3):267–297.
16. Řehůřek R, Sojka P (2010) Software Framework for Topic Modelling with Large Corpora in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. (ELRA, Valletta, Malta), pp. 45–50. http://is.muni.cz/publication/884893/en.
17. Wallach HM, Mimno DM, McCallum A (2009) Rethinking lda: Why priors matter in *Advances in neural information processing systems*. pp. 1973–1981.
18. Le Q, Mikolov T (2014) Distributed representations of sentences and documents in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pp. 1188–1196.
19. Trask A, Michalak P, Liu J (2015) sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. *arXiv preprint arXiv:1511.06388*.
20. Breiman L (2001) Random forests. *Machine learning* 45(1):5–32.
21. Elkins Z, Ginsburg T, Melton J (2013) The content of authoritarian constitutions. in *Constitutions in authoritarian regimes*, eds. Ginsburg T, Simpser A. (Cambridge University Press).
22. Cheibub JA, Elkins Z, Ginsburg T (2014) Beyond presidentialism and parliamentarism. *British Journal of Political Science* 44(3):515–544.
23. Krackardt D (1987) Qap partialling as a test of spuriousness. *Social networks* 9(2):171–186.
24. Dekker D, Krackhardt D, Snijders T (2003) Multicollinearity robust qap for multiple regression in *1st annual conference of the North American Association for Computational Social and Organizational Science*. (NAACSOS), pp. 22–25.
25. Cranmer SJ, Leifeld P, McClurg SD, Rolfe M (2017) Navigating the range of statistical tools for inferential network analysis. *American Journal of Political Science* 61(1):237–251.