On Measuring Textual Similarity

Robert Shaffer
Zachary Elkins

May 30, 2019

Abstract

Understanding the similarity among texts can be enlightening and useful. Unfortunately, similarity comparisons are labor-intensive for human interpreters. Computer-assisted methods are more scalable, but their validity is poorly understood, especially for long, complex documents common in political science research. In this paper, we address this gap in three parts. First, we provide an overview of the *uses* and *types* of textual similarity. Second, we propose a workflow designed to measure *thematic similarity*, which we suggest is the most useful similarity type for applied political science research. Third, we apply our workflow to a unique dataset of expert-interpreted national constitutions. We find that our scores perform well in both out-of-sample prediction and substantive modeling tasks. We also find that a small human-generated training set ($< 25\%$ of the corpus) improves both overall performance and robustness to modeling choices, which emphasizes the importance of human intervention for complex measurement tasks.

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. In fact, 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, unsupervised modeling approaches designed to extract information from these data sources have proliferated. However, it is not always clear how information extracted by such techniques relates to human-generated constructs, or the similarity of those constructs across different texts. This problem is particularly acute with concepts pitched at higher levels of abstraction. Though modern computational tools can easily compare the extent to which one text "borrows" words or phrases from another, less attention has been paid the problem of comparing broad themes contained in a pair of documents, which is our problem of interest.

In this paper, we begin by reviewing existing substantive and methodological work on textual similarity. We then introduce a typology of the various kinds of textual similarity, and situate the concept of thematic similarity—our construct of interest—within this broader context. Next, we describe a machine-assisted workflow that allows applied researchers to develop and measure pairwise semantic similarity scores between documents with minimal human-coding labor. Though some human intervention is preferable in any study of a latent quantity, we argue that our approach offers a principled way for resource-constrained researchers to measure similarity patterns in a large corpus while manually coding a small ($<25\%$) proportion of the content of that corpus.

As an applied example of this workflow, we examine patterns of textual similarity within the Comparative Constitutions Project's (CCP) hand-coded data on the content of national constitutions (Elkins et al., 2009). Using these data, we develop a set of criterion cores of "thematic" similarity, which we use to test a variety of potential feature-extraction and learning approaches within our proposed workflow. In out-of-sample performance comparisons, we find that unsupervised similarity comparison strategies used by applied researchers perform moderately well, but are improved substantially with the introduction of a small quantity of manually-coded training data. Finally, in a series of validity tests, we find that results produced with both the criterion measure and our supervised approximations yield similar substantive conclusions. These findings, we argue, emphasize the need for individual researchers to develop similarity comparison methods tailored to their particular conceptualizations of textual similarity, rather than relying on unsupervised approaches.

1 Uses of Similarity Measures

Why measure similarity? Similarity measures are useful when one is interested in identifying associations between *cases*, as opposed to between *variables*. In methodological work, such scenarios are ubiquitous. Many common measurement and modeling tasks implicitly depend on measures of similarity between cases, including matching, clustering, and nearest-neighbor prediction and recommendation algorithms. However, measures of similarity between cases (textual or otherwise) are also central to a variety of substantive research strategies. We outline three such paradigms here.

1.1 Group Comparisons

A natural similarity use case is to test hypotheses regarding within- and between-group cohesion in some pre-identified grouping structure. Here, similarity measures often serve as a *dependent variable*, which a researcher might attempt to explain using group membership characteristics of the cases in her dataset. For example, a researcher studying political communication might 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). Or, a social media researcher might want to identify differences in social media usage by political party (Jones et al., 2018), gender (Bamman et al., 2014), national identity (Metzger et al., 2016), or age (Jang et al., 2015). One approach to such problems would be to score texts with respect to some well-defined dimension, and to compare mean differences between groups along that dimension. However, if the researcher is interested in many dimensions of comparison, conducting comparisons directly on all features of interest would be unwieldy. A more practical approach would be to define a higher-level similarity metric, which aggregates various features of discourse into a single similarity score that can be easily modeled or described.

Each of the examples outlined above involves the comparison of rhetorical output between members of *known* groups. However, a researcher operating in an unfamiliar setting might also be interested in *clustering* cases to inductively identify novel groupings. Though clustering methods abound (see, e.g. Grimmer and King, 2011), all depend on some defined notion of similarity between cases, which is used as an input in the underlying clustering model(s).

1.2 Diffusion

We conceive of *diffusion* as a flagship term for a host of varied phenomena related to the interdependence of actors or jurisdictions in the creation of policy, rules, and practices (Dolowitz and Marsh, 2000; Shipan and Volden, 2008). One unit's behavior alters the probability of such behavior in related unit, and one prediction might be that the two units' behavior converges (though divergence, through repellant mechanisms, is theoretically possible as well). Thus, similarity between cases serves as a *dependent variable* that scholars might describe or explain through shared social, economic, political, or cultural attributes or direct network ties. Prominent studies across political science subfields examine diffusion patterns in American states (Berry and Berry, 1992; Linder et al., 2018), Congressional legislation (Wilkerson et al., 2015), national governments

(Dolowitz, 1997; Weyland, 2009), and international institutions (Elkins and Simmons, 2005).

Since many policy and legal decisions are represented in textual form, textual similarity presents a tractable resource with which to trace diffusion patterns across political and institutional contexts. As we note below, not all similarity measurement approaches are appropriate for all diffusion questions, but diffusion clearly represents an important similarity comparison use case.

1.3 Innovation

Finally, when applied to time-series data, similarity comparison approaches can help identify shifts and disruptions in discourse. Here, similarity can serve as either an *independent* or *dependent* variable. Tetlock (2011)'s study of investor reactions to "stale" news offers an instructive example of the former case. He argues that since news stories are repeated across publications, sometimes at a much delayed rate, some investors may be in danger of reacting to news that already known (and, hence, already baked into the stock price). Tetlock's key independent variable is therefore the originality of news, as measured by textual similarity across articles over time. Pagliari and Wilf (2019), by contrast, employ the latter research design. These authors test the assumption that financial crises trigger major changes in banking and securities regulation, and their key outcome variable is the (dis)-similarity of regulatory text over time. Shaffer (2017) uses a related approach to identify periods of expansion and contraction of the regulatory policy agenda following crisis events, which he applies to study Congressional discourse following the 2008-2009 Financial Crisis.

2 Towards Thematic Similarity

Analysts focused on the similarity between texts face a number of practical challenges. The first is definitional: which kind of similarity matters? We suggest that a notion we term *thematic similarity* is the most relevant concept for a large majority of political science studies. Thematic similarity, in our usage, refers to the extent to which two documents contain the same "inventory" of themes or ideas. This definition, we suggest, is particularly well-suited for summarizing patterns of similarity within the kinds of long, complex documents frequently studied by political scientists. Unfortunately, despite its importance, thematic similarity is also underspecified in the methodological literature. For practical reasons, computer scientists and computational linguists tend to focus on sentences and short excerpts rather than larger texts. As a result, to our knowledge no existing studies contrast methods of thematic similarity comparison, which motivates our efforts in this paper.

In the remainder of this section, we develop a typology of similarity metrics, and contrast thematic similarity with other plausible similarity comparison approaches such as equivalence or synonymy. We conclude by developing a workflow designed to extract thematic similarity scores in a machine-assisted fashion. This workflow, we suggest, is flexible and generally applicable to an array of applied political science problems that involve substantive research problems we describe above.

2.1 A Typology of Similarity

In studies of similarity patterns, an important—though often unquestioned—decision is the choice of a *type* of similarity measure. A recognizable reference point in text similarity is the set of text re-use measures designed to detect plagiarism (Wilkerson et al., 2015; Linder et al., 2018). But text re-use is clearly only one variant of text similarity. We suggest a hierarchical taxonomy of measures based on three fundamental dimensions: *word usage*, *meaning*, and *theme* (see Figure 1).[1] Higher-level types of similarity generally imply their lower-level counterparts; for example, two documents that share a near-identical set of words will necessarily display near-identical meaning and overall theme. By contrast, two documents that share a meaning or theme may or may not use similar words to express the ideas they contain. We can think of these combinations as constituting various classes of similarity, which we think of as something of a hierarchy of similarity.

As a running example of these similarity types, consider the text of the 4th Amendment to the U.S. Constitution, which describes criminal procedure rights related to the collection of evidence and, according to some interpretations, a general right to privacy.

[1] For parsimony, we exclude stylistic elements such as sentence length or grammatical structure, and second-order quantities such as sentiment or readability. Though quantities like these can be important for some research applications, they fall outside the notions of similarity with which substantive researchers are usually concerned.

Table 1: A Hierarchy of Kinds of Textual Similarity

		Text contains the same...		
Rank	Similarity Type	Words	Meaning	Theme
1	Equivalent	x	x	x
2	Synonymous		x	x
3	Thematic			x
4	Unrelated			

> **United States (1789), 4th Amendment**
> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.[2]

2.1.1 Equivalence

The simplest and most direct notion of similarity is one of *equivalence*. Here, we define an equivalence-based notion of similarity as one in which two texts are more similar to the extent that they use the *same words* in the *same order* to reflect the same *meaning* and *topic*.[3] An example of this scenario is Section 3.1 of the Marshall Islands Constitution, which repeats the U.S. 4th Amendment verbatim (the only instance of direct re-use of the 4th Amendment text in a national constitution that we are aware).[4] But of course, one finds variation in the overlap of text. For example, Article 35 of the Japanese Constitution – famously written during post-World War II American occupation – is very nearly "equivalent" to the 4th Amendment passage:

> **Japan (1946), Article 35**
> The right of all persons to be secure in their homes, papers and effects against entries, searches and seizures shall not be impaired except upon warrant issued for adequate

[2] https://www.constituteproject.org/constitution/United_States_of_America_1992#145

[3] Parenthetically, note that two texts can use the same words arranged in a different order to express the same meaning and topic (e.g. in passive/active voice sentence constructions). However, we view this possibility as an edge case, which is largely restricted to short documents such as phrases or sentences and thus outside our focus in this paper.

[4] https://www.constituteproject.org/constitution/Marshall_Islands_1995#60

> cause and particularly describing the place to be searched and things to be seized, or except as provided by Article 33.
>
> Each search or seizure shall be made upon separate warrant issued by a competent judicial officer.[5]

Equivalence-based notions of similarity offer a number of advantages. Unlike other possible definitions, equivalence similarity is unambiguously observable. If two documents share a sufficiently long sequence of common words, we can be essentially certain that one document's authors drew upon the other for inspiration, or that both drew upon a single common source. Leveraging patterns of text re-use has been a particularly successful strategy for researchers working with legal texts written in a common context, such as Congressional bills or state-level enacted legislation (e.g. Wilkerson et al., 2015; Wilkerson and Casas, 2017). Legal language is usually written by expert drafters according to a shared set of stylistic norms that are specific to a particular legal context, which incentivizes authors to settle on particular phrases and terms of art to describe key ideas. Better still, because legal texts operate in the public domain, their authors can lift language from related documents freely, allowing researchers to easily trace the flow of language and ideas from one text to another.

However, an equivalence-based notion of similarity is not optimal for all research problems. In applied settings, researchers often care more about the similarity of ideas and less about how (with which words) those ideas are expressed. Outside of specialized, technical settings like law, groups of written documents rarely contain many shared terms of art, even if the underlying ideas discussed in those documents overlap. Rather, in most domains writers will seek to express themselves in a manner which respects their training, interests, and idiosyncratic stylistic preferences. And, even within a highly technical area like legal writing, authors are unlikely to re-use text from outside their immediate professional or personal context, which further limits the applicability of an equivalence-based notion of similarity.

[5] https://www.constituteproject.org/constitution/Japan_1946#106

2.1.2 Synonymy

An alternative definition of similarity – which relaxes these constraints somewhat – is *synonymy*. Here, we treat texts as similar to the extent that they share the same meaning or meanings, regardless of the words they contain. Clearly, texts that share a substantial quantity of shared language are also likely to share a similar meaning, but this relationship is not a necessary one. For an extreme example, consider translating a text from one language to another; though the original and translated version of a given text share no words in common, their meaning and themes are identical. But even when restricted to the same language, synonymy is common. An example from the constitutional context might be the first two clauses of Bolivia's (2009) Article 25, which express a version of the the U.S. 4th Amendment, although worded differently:

> **Bolivia (2009), Article 25**
> I. Every person has the right to the inviolability of his home and to the confidentiality of private communications of all forms, except as authorized by a court.
>
> II. Correspondence, private papers and private statements contained in any medium are inviolable and may not be seized except in cases determined by law for criminal investigation, based on a written order issued by a competent judicial authority.[6]

Like an equivalence-based definition of similarity, a synonymy-based definition offers both advantages and disadvantages. On the positive side, synonymy is clearly more flexible than equivalence. In many contexts, direct plagiarism or extensive quotation of another author's work is either legally sanctioned or socially undesirable. Speakers and writers are therefore unlikely to lift large quantities of language from others. However, consciously or not, authors frequently use their own language to re-express ideas originally advanced by their peers. As a result, identifying synonymous or nearly synonymous sections of text offers a useful way to answer research questions related to diffusion, group cohesion, or novelty in settings where an equivalence-based notion of similarity may be inapplicable.

Unfortunately, at least for political science research, synonymy is also limited. From a practical standpoint, compared with equivalence, synonymy is more difficult to detect. As the Bolivia

[6] https://www.constituteproject.org/constitution/Bolivia_2009#174

example shows, synonymous texts may share many or few words, and they may be written in similar or substantially different styles. In the computer science literature, various authors have developed flexible machine learning approaches designed to solve this problem for pairs of individual words, phrases, or sentences, (see, e.g. Agirre et al., 2012; Ganitkevitch et al., 2013; Xu et al., 2015; Agirre et al., 2016). However, we are aware of no existing studies that have attempted to extend these methods to longer and more complex documents. We suspect this gap is due to a conceptual mismatch. Because large, complicated texts usually address an array of themes and ideas and advance a broad set of arguments, these documents cannot express the same underlying meaning. Synonymy, in other words, may be too exacting a standard when applied to larger texts common in political science research.

2.1.3 Thematic

Finally, if we relax the conditions that two passages share the same words *and* that they share the same meaning, we might think of a milder form of similarity, which we term *thematic similarity*. Under this definition, we focus on the extent to which two texts share the same underlying theme or themes. To return to our running example, the 4th Amendment of the US Constitution deals broadly with issues of criminal procedure and collection of evidence, with a focus on rights of individuals. By contrast, the Egyptian Constitution deals with a similar topic, but frames it as a problem of protecting communications:

> **Egypt (2014), Article 57**
> Telegraph, postal, and electronic correspondence, telephone calls, and other forms of communication are inviolable, their confidentiality is guaranteed and they may only be confiscated, examined or monitored by causal judicial order, for a limited period of time, and in cases specified by the law.[7]

In our view, this form of similarity is the most relevant and generally applicable for political science research. For many political scientists, the primary texts of interest are long, complex documents that come from heterogeneous sources such as constitutions, party manifestos, or

[7] https://www.constituteproject.org/constitution/Egypt_2014#208

political speeches. In these contexts, an equivalence- or synonymy-based notion of similarity is likely too demanding. Like all large documents, a given pair of party manifestos or political speeches are not likely to share substantial common language or to carry the same substantive meaning. By contrast, such a pair of documents might well share similar underlying themes or ideas, even if their authors operate under different legal, political, or linguistic norms. For example, a pair of party manifestos from two different countries might be written in highly different styles and focus on policy problems and solutions that are specific to the needs of their countries, while still sharing the same underlying themes—such as environmentalism, multiculturalism, or traditional morality.

Though more general than equivalence- or synonymy-based notions similarity, a thematic notion of similarity carries its own set of challenges. Like synonymy, thematic similarity is difficult to detect, with few obvious linguistic features that indicate its presence. However, compared with equivalence or synonymy, thematic similarity is substantially more labor-intensive for human annotators to identify. In order to investigate patterns of thematic similarity, a researcher would first need to define the set of themes or ideas potentially present in a particular corpus, which usually involves reading and analyzing a substantial portion of that corpus' contents. Then, she would need to code each document according to the presence or absence of the identified set of themes. Finally, she would need to select a similarity function, which she could use to quantify the extent to which two documents' thematic contents are similar. And, as we emphasize throughout this section, these challenges are magnified when comparing long, complex documents, which are the primary use case for political science researchers.

Because of the data-intensive nature of this task, we might reasonably wonder if a machine-driven approach to thematic similarity comparison is feasible. Unfortunately, while a plethora of plausible approaches to generating such similarity scores are available, opportunities to train models of thematic similarity—and to validate their outputs—are rare. As we describe in the previous section, similarity prediction tasks in computer science and natural language processing are generally conducted on short excerpts using an abstract notion of similarity. As a result, to our knowledge no existing study has explored the extent to which automated methods can reproduce human thematic similarity judgments.

2.2 A Workflow for Thematic Similarity

To address this problem, we propose a computer-assisted workflow designed to extract broad "thematic" similarity judgments from a corpus with minimal human input. This workflow proceeds in three steps. First, develop a set of "target" similarity scores for some subset of document pairs within a corpus of interest. Second, conduct a dimensionality reduction step, in which a topic model, word embedding model, or other dimensionality reduction tool is used to convert document texts into vectors of lower-dimensional numeric data. Third, using the numeric document representations as input features, train a flexible, prediction-oriented machine learning model to to recover the human-generated "target" similarity scores for each pair of documents. In the following sections, we provide a broad overview of each of these steps, with a set of specific potential implementation strategies for each step.

2.2.1 Developing the Target

As in any supervised learning project, the first step in building a thematic similarity measurement workflow is to construct a *training set*, or a set of human-annotated reference cases that act as a reference point for a downstream prediction model. For a thematic similarity application, developing this training set will first involve identifying the set of relevant themes in a given corpus, which might involve creating a novel codebook or drawing on an existing typology or resource.[8] Then, given a set of relevant themes, for each document in the training set the researcher will need to code the presence or absence of each theme. Finally, the researcher will need to relate data on thematic presence/absence for each pair of documents using an appropriate similarity function, which yields a similarity score for each unique pair of documents in the dataset.[9]

[8] E.g. the Comparative Agendas Project in public policy research, or the Party Manifestos Project in electoral politics.

[9] For example, in our application, we use a Jaccard similarity metric, which is appropriate for comparing vectors of binary presence/absence data. If the underlying themes are coded in continuous fashion, a Euclidean or cosine similarity metric might be appropriate instead.

Importantly, note that developing a training set is not an absolutely necessary part of this workflow. In principle, researchers could simply generate a lower-dimensional vector representation of each document in an unsupervised fashion, and generate similarity scores by comparing these vectors with an appropriate distance metric (e.g. Euclidean or Hellinger). However, to preview our applied results, we find that adopting a supervised approach with even a small quantity of training data both improves predictive performance and reduces model dependence. Better still, generating training data also gives researchers an opportunity to clarify their conceptualization scheme and validate their similarity comparison approach, which are important steps in any research project involving substantial measurement tasks.

2.2.2 Reducing Dimensionality

The second part of our workflow is a *dimensionality reduction* step, in which each document is converted into a lower-dimensional numeric representation. In virtually all quantitative text-as-data research projects, the researcher must at some point convert her corpus into a vector of numeric information, which can be used as inputs into a prediction model or as quantities of interest to be explained in their own right. The vector of values might contain word or n-gram frequencies, topic proportions from a topic model, or embedding-space features from a document or word embedding model. In our application, we compare performance of a variety of possible approaches to this problem, including simple linear-algebra-based dimensionality reduction techniques (Dumais et al., 1988; Deerwester et al., 1990) as well as more modern topic modeling (Blei et al., 2003; Blei, 2012; Roberts et al., 2014) and word embedding (Mikolov et al., 2013; Le and Mikolov, 2014) approaches.

Whatever a researcher's preferred approach, converting each document into a lower-dimensional numerical representation is a vital part of the workflow we suggest. Importantly, unlike conventional applications for topic models or other dimensionality reduction models, generating a substantive interpretation for each extracted topic or dimension is not necessary in our application. Because our objective is to build a model designed to replicate human-generated similarity comparisons, our performance metric is not the coherence or interpretability of the topics generated by a particular model, but rather the out-of-sample predictive performance of a model built using the topics generated by that model. As we show in our application, most dimensionality reduction approaches and parameter values perform similarly when assessed by this standard, especially when at least some training data is available to the model.

2.2.3 Building a Model

The final step in our workflow is to use the vector representations produced in the previous step to build a predictive model of pairwise document similarity. This model should take as inputs the numeric feature vectors generated in the previous step, and generate a predicted pairwise similarity value for the two documents as an output. As in the dimensionality reduction step, the selection of model is up to researcher preference. However, our experiments suggest that flexible, generally-applicable approaches like random forests are likely to outperform more specialized, theoretically-driven schemes like kernel-based metric learning algorithms.

At first glance, this preference for flexible, ad-hoc methods such as random forests over narrowly tailored metric-learning methods might seem counter-intuitive. After all, metric-learning methods are designed for the purpose of learning pairwise distance metrics, and are intended to exploit the special properties of distance metrics in order to narrow the set of candidate parameter values and functional forms. However, as we show in our appendices (see Appendix B.1.2 for details), metric-learning approaches—which exploit special features of distance metrics to impose additional structure on the learning problem—actually underperform more general machine-learning approaches in out-of-sample testing. The reason for this performance gap is straightforward. Most methods designed to learn distance metrics from data assume that the "true" input features are available to the model at training time. However, in our application, the features we generate for each document in step two are necessarily different from those used to produce the human-generated gold standard values. As a result, a modeling approach that assumes that the features observed by the model are the same as those used to generate the "target" similarity scores performs poorly in this application. A more flexible approach—one that allows the relationship between input features and the target similarity value to exhibit threshold effects or other non-linearities—represents a better alternative.

2.2.4 Summary

In our view, this three-step strategy is simple, flexible, and applicable for most social science research tasks involving similarity comparison. In modern social science research, dimensionality reduction and predictive modeling are common, particularly for researchers focused on textual data. As a result, extracting document-level vector representations from a topic model and fitting flexible, prediction-oriented algorithms are likely to be familiar tasks for applied researchers, or at least more familiar than that of fitting a metric-learning algorithm or some other more complex modeling approach. Best of all, the results of our applied study suggest that this straightforward approach performs well for the kinds of analytic tasks likely to be useful for applied researchers, particularly when at least some training data is available to the model.

3 Application: Similarity Among Constitutions

As an application of our proposed workflow, we draw on the the Comparative Constitutions Project's (CCP) database of national constitutions Elkins et al. (2009). National constitutions are one of the central objects of study in law and political science, and scholars have long been interested in studying the flow of constitutional ideas across space and time (see, e.g. Rutherford et al., 2018). Discussions like these depend on textual similarity comparisons, which scholars deploy to evaluate hypotheses related to the spread of ideas across jurisdictions.

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 with which to work.

3.1 Creating a Target

Our corpus is the set of all national constitutions in-force as of 2014, as identified by the CCP. As mentioned previously, in addition to constitutional texts the CCP data contain extensive information on the inventory of topics included in any two constitutions (e.g., whether the constitution mentions a central bank, addresses the accession of new territory, etc.). From this list, we eliminate *sub*-topic items 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. This process leaves us with a set of 70 topics, the feature set that we analyze.[10]

To generate thematic similarity values from these data, we calculate a simple Jaccard similarity coefficient (1912) between the feature vectors for each unique pair of constitutions. The Jaccard formula takes the following general form[11]:

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

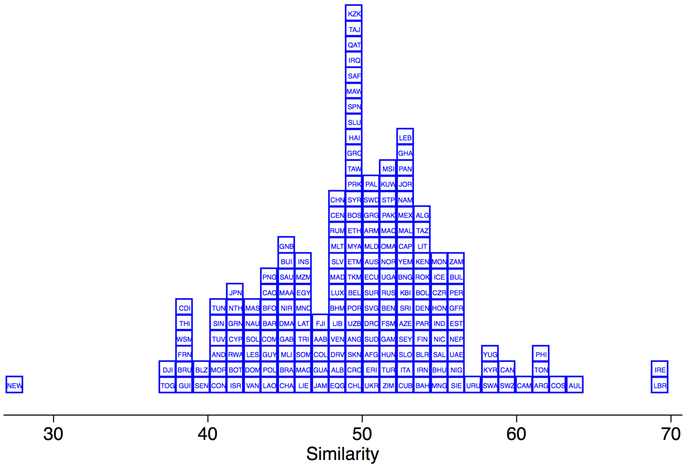
where $min(.)$ and $max(.)$ are the element-wise minimum and maximum functions.

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 (e.g., a procedure that 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.

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

[11] 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.

Figure 1: Similarity to the U.S. Constitution c. 2014 (hand-coded, across topics)



At least 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* appear markedly different in their content.

The U.S. Constitution may be more widely known (at least compared to Brunei). In Figure 1, we strip-plot the distribution of similarity scores to the U.S., and identify cases with three-letter country codes. We might expect the U.S. text 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. Recall the sample is a snapshot of constitutions in 2014. 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. More analysis of this variation in similarity across hand-coded constitutions would be interesting; suffice it to say here that the measure is face-valid.

3.2 Reducing Dimensionality

Having established a target measure of thematic similarity, we then use a series of commonly-applied unsupervised dimensionality reduction schemes schemes to extract candidate feature vectors from each constitutional text. For the purposes of this study, we focus on features generated from four co-ocurrence based latent trait models: specifically, Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (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 2 for details). We also include similarity scores generated from term-frequency-inverse-document-frequency (TF-IDF)-weighted word count vectors as a baseline point of comparison.

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.[12] 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. Co-ocurrence based latent trait models like LDA, STM, and word2vec are designed to extract dimensions that correspond to broad themes or senses of meaning from documents or words. Since our criterion metric is determined by the shared presence/absence of certain high-level ideas 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 pre-processing approach (summarized in Table 2). First, we subdivide each constitutional text into a number of constituent documents (see Appendix A for examples). Like many datasets of interest to political scientists, constitutions are long and thematically diverse. For co-occurrence-based models like LDA, STM, and LSI, these documents

[12] Sparse matrix factorization or sparse LDA (see, e.g., Zhang et al. (2013); Ming et al. (2014)) offer an alternative feature extraction framework. However, since these approaches are not frequently used in applied work, we do not test them here, but these approaches represent a potential direction for future work.

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 segment constitutions based on their internal header structure.[13] Since word2vec relies on word ordering and localized contextual information, we instead subdivide documents into sentences when training this model.[14]

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[15] 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 2, 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 afford *some* robustness assessment 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 do LDA and STM; however, all models perform similarly in our

[13] For headers containing multiple paragraphs, we further subdivide each header into paragraphs.

[14] As identified by the pre-trained Punkt sentence tokenizer contained in NLTK (http://www.nltk.org/).

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

Table 2: Estimation and pre-processing details for feature set under consideration.

Model	Unit	Pre-processing	Hyperparameters
TF-IDF	headers[a]	(1) lower-case; (2) punctuation, stop-words, tokens ≤ 3 characters, tokens in ≤ 10 documents removed; (3) documents ≤ 5 tokens removed	n/a
LSI	headers[a]	Same as TF-IDF	{20,50,100,150,200} topics
LDA	headers[a]	Same as TF-IDF	{20,50,100,150,200} topics; MALLET hyperparameter optimization
STM	headers[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, TF-IDF, and word2vec estimated via Gensim (Řehůřek and Sojka, 2010). For word2vec models, we do not set a maximum window size, and allow the predicted word to be influenced by all words in the sentence under consideration. 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] $n \approx 201000$

supervised experiments. We view this finding as suggestive 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 pre-processing steps, parameter settings, and models are not the only plausible feature-extraction approaches. However, for practical reasons, we cannot test all possible specifications. Instead, we suggest that the options we test here are both plausible and represent a reasonable selection of commonly-used approaches, and therefore offer a useful baseline for applied work.

3.3 Model 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. For the supervised learning approaches, we vary the proportion of the dataset used for training, and assess performance 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 an inverse discretized Hellinger distance between the feature vectors $\mathbf{z}_i$ for each constitution, defined as

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

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 flexible, prediction-oriented modeling 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).[16] 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

[16] 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.2.1 for details).

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.[17]

As before, we emphasize that this is not the only approach one might consider. We investigate a variety of alternative models and specifications in Appendix B, but all perform similarly or worse than the approach we propose. Since our preferred approach is scalable, straightforward to implement, and performs well, we argue that it offers a plausible strategy for applied work.

4 Validating Thematic Similarity

To validate the scores we present in the previous section, we proceed in two phases. First, we conduct a simple (and aggregate) analysis of the covariance between the criterion and candidate measures. Broadly, we find that similarity scores generated using both supervised and unsupervised approaches across all feature extraction strategies approximate the criterion measure. However, we find that a supervised approach—even with a small training set ($< 25\%$ of the corpus)—substantially outperforms the unsupervised baseline. Moreover, scores generated using a supervised approach appear to be robust to all feature selection and parameter setting decisions we examine, which suggests the additional utility of this approach. In our second test, we examine whether models applied to the criterion and candidate measures produce the same substantive conclusions. Here, again, we find that our proposed approach performs well, with models fit to the human- and machine-generated similarity scores producing broadly similar results.

[17] 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.

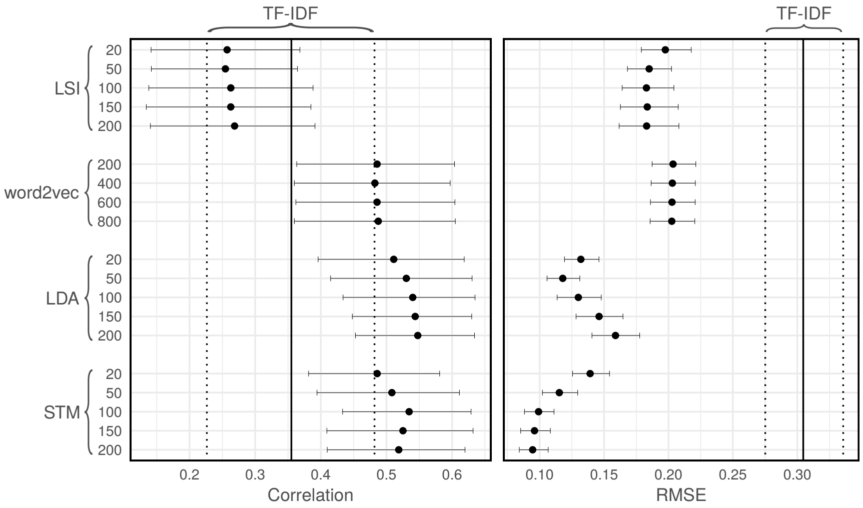
4.1 Test 1: Aggregate Learner Performance

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 above, and assessed the relationship between these scores and the human-generated criterion values. 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.2.2, 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.

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 (≈25% of the dataset), our supervised predictions consistently outperform the unsupervised baseline by both correlation and RMSE. By 75 documents (≈40% of the dataset), these improvements are striking; at $n = 75$ training documents, the

Figure 2: Correlation and RMSE comparisons between machine- and human-generated similarity values

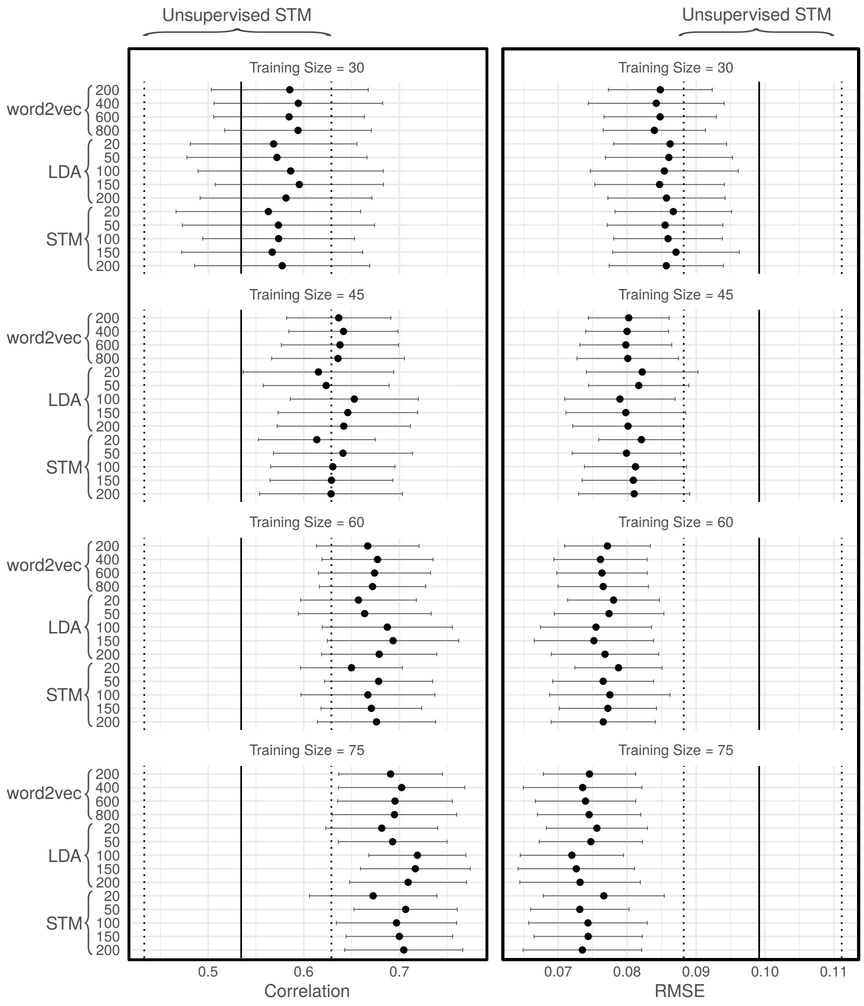


Solid and dashed vertical lines indicate mean out-of-sample correlation/RMSE and 95% confidence interval between CCP targets and similarities generated using baseline TF-IDF features. Dots and solid horizontal lines indicate mean correlation/RMSE and 95% confidence intervals between human- and machine-generated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

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, which provides researchers some reassurance that their results are largely invariant to choice of dimensionality parameter or feature extraction approach.

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



Dots and solid horizontal lines indicate mean correlation/RMSE and ±2 sample standard deviations between human- and machine-generated similarities, estimated from 100 random train/test splits at indicated training set sizes. Solid and dashed lines give mean and 95% confidence intervals from the unsupervised STM_{100} model shown in Figure 2 as a baseline comparison.

4.2 Test 2: Patterns of Constitutional Similarity

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 substantive 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.[18] 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.

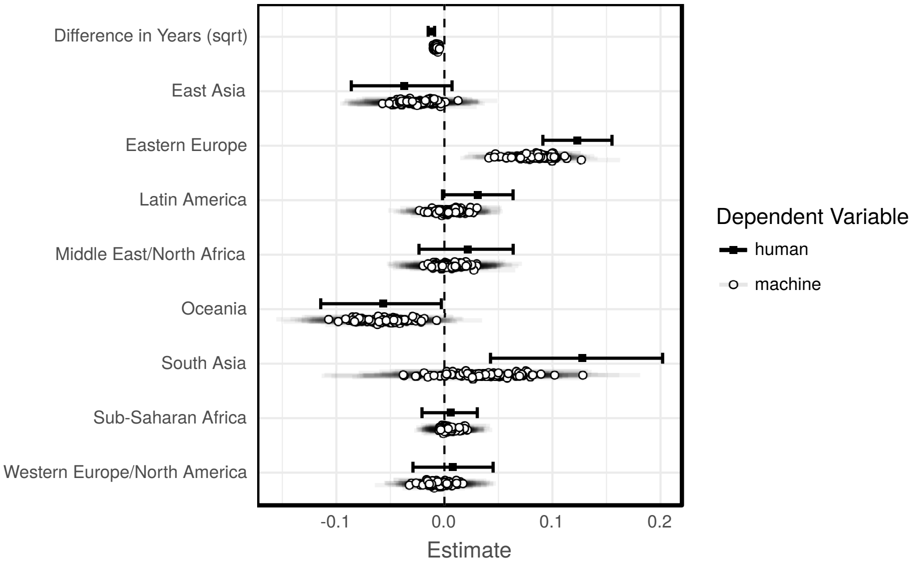
2. *Same region.* A set of dummy variables that equal 1 if a dyad includes constitutions from the same region, for each of eight geographic regions.[19] We expect each of these coefficients to be positive, since constitutions drawn from the same region are likely to be more similar than those drawn from different regions (the implicit left-out category). However, there is likely to be substantial 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 error term estimates for regression models are likely to be inappropriate in this context. 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 depen-

[18] We test two other specifications of this variable in Appendix C.3. Both approaches yield nearly identical conclusions to the one described in-text.

[19] Namely, East Asia, Eastern Europe, Latin America, Middle East/North Africa, Oceania, South Asia, Sub-Saharan Africa, and Western Europe/North America.

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.

dencies 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.[20] 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

[20] As noted in the previous section, 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.

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 in Appendix C.2.

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 is predicted to be 8 points more similar than is 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 to one another than is 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 a pair of Oceania constitutions is more dissimilar to one another than is a pair of constitutions drawn randomly 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. Figure 4 suggests that the answer 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 did the model with the criterion values. Across 100 model replicates, 87% of the non-intercept coefficient estimates returned the same conclusions regarding significance and sign as did the human-generated coefficient estimates.[21] All coefficients besides the South Asia coefficient returned the same substantive conclusions in at least 80% of replicates, with the South Asia coefficient – the smallest regional group in our dataset – returning a positive and significant estimate in only 17% of replicates.

[21] 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

Textual similarity is an important, perhaps even under-analyzed, quantity applicable to a broad set of intriguing research questions. However, existing literature offers little systematic guidance about the ways that texts can align, the relevant kinds of research questions and corpora, and the methods used to measure and estimate similarity scores. This gap is particularly noticeable in the context of long documents, such as national constitutions or political speeches. As we argue, these documents represent a core data source for political science research, and one whose use is only growing.

We address this gap in four stages. First, we offer a simple typology of textual similarity, with a focus on *equivalence*, *synonymy*, and *thematic* similarity. Though all of these conceptions of similarity definitions are potentially useful, we suggest that *thematic* similarity is most useful for applied political science research. Second, we develop a workflow designed to allow applied researchers to extract thematic similarity scores from large corpora with minimal human intervention. Third, we leverage the Comparative Constitutions Project's data on the content of national constitutions to offer an applied example of our workflow. Fourth, we validate and test this workflow, with an emphasis on testing both predictive performance and robustness to modeling choices. We find that a supervised approach with a small ($<$25% of the corpus) training set noticeably outperforms a simple unsupervised baseline and improves robustness to modeling choices.

In our view, these results are encouraging for automated text practitioners. Given a small 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 documents. In these situations, machine-generated approximations offer a useful way for researchers to test hypotheses regarding the diffusion, innovation, and clustering of ideas and discourse. The validation exercises we conduct in this paper provide some assurance for researchers uncertain about the promise of such tools and offer a path forward for applied work.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics, 2012.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, 2016.

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

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

Frances Stokes Berry and William D Berry. Tax innovation in the states: Capitalizing on political opportunity. *American Journal of Political Science*, pages 715–742, 1992.

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

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

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

Jose Antonio Cheibub, Zachary Elkins, and Tom Ginsburg. Beyond presidentialism and parliamentarism. *British Journal of Political Science*, 44(3):515–544, 2014.

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

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 41(6):391–407, 1990.

David Dekker, David Krackhardt, and Tom Snijders. 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, 2003.

Matthew James Denny and Arthur Spirling. Assessing the consequences of text preprocessing decisions. Working paper. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2849145`, 2016.

David P Dolowitz. British employment policy in the 1980s: learning from the american experience. *Governance*, 10(1):23–42, 1997.

David P Dolowitz and David Marsh. Learning from abroad: The role of policy transfer in contemporary policy-making. *Governance*, 13(1):5–23, 2000.

Susan T Dumais, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman. 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, 1988.

Zachary Elkins and Beth Simmons. On waves, clusters, and diffusion: A conceptual framework. *The Annals of the American Academy of Political and Social Science*, 598(1):33–51, 2005.

Zachary Elkins, Tom Ginsburg, and James Melton. *The endurance of national constitutions*. Cambridge University Press, 2009.

Zachary Elkins, Tom Ginsburg, and James Melton. The content of authoritarian constitutions. In

Tom Ginsburg and Alberto Simpser, editors, *Constitutions in authoritarian regimes*. Cambridge University Press, 2013.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

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

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

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

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

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

Jin Yea Jang, Kyungsik Han, Patrick C Shih, and Dongwon Lee. Generation like: Comparative characteristics in instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4039–4042. ACM, 2015.

Kevin L Jones, Sharareh Noorbaloochi, John T Jost, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Liberal and conservative values: What we can learn from congressional tweets. *Political Psychology*, 39(2):423–443, 2018.

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

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

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

Fridolin Linder, Bruce A Desmarais, Matthew Burgess, and Eugenia Giraudy. Text as policy: Measuring policy similarity through bill text reuse. 2018.

Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Megan MacDuffee Metzger, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. Tweeting identity? ukrainian, russian, and# euromaidan. *Journal of Comparative Economics*, 44(1): 16–40, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

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

Stefano Pagliari and Meredith Wilf. Does the g20 affect international financial regulation? a text-based approach.

Stephen Purpura and Dustin Hillard. 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, 2006.

Radim Řehůřek and Petr Sojka. 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, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

Margaret E Roberts, Brandon M Stewart, and Richard Nielsen. Matching methods for high-dimensional data with applications to text. 2015.

Alex Rutherford, Yonatan Lupu, Manuel Cebrian, Iyad Rahwan, Brad L LeVeck, and Manuel Garcia-Herranz. Inferring mechanisms for global constitutional progress. *Nature Human Behaviour*, 2(8):592, 2018.

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

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

Robert Shaffer. Cognitive load and issue engagement in congressional discourse. *Cognitive Systems Research*, 44:89–99, 2017.

Charles R Shipan and Craig Volden. The mechanisms of policy diffusion. *American journal of political science*, 52(4):840–857, 2008.

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

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

Tracy Sulkin. *Issue politics in Congress*. Cambridge University Press, 2005.

Paul C Tetlock. All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5):1481–1512, 2011.

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

Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981, 2009.

Kurt Weyland. *Bounded rationality and policy diffusion: social sector reform in Latin America*. Princeton University Press, 2009.

John Wilkerson and Andreu Casas. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544, 2017.

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

Wei Xu, Chris Callison-Burch, and Bill Dolan. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11, 2015.

Aonan Zhang, Jun Zhu, and Bo Zhang. Sparse online topic models. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1489–1500. ACM, 2013.