Text Scaling for Open-Ended Survey Responses and Social Media Posts

William R. Hobbs

Assistant Professor, Cornell University

Abstract

Open-ended survey responses and social media posts contain valuable information about public opinions, but can consist of only a handful of words. This succinctness makes them hard to summarize, especially when the vocabulary size across all respondents is large. Here, we propose a method to characterize and score opinions in these data. The approach scores respondents' opinion justifications based on their use of common words and the words that tend to accompany them, so that we can summarize opinions without relying on rarely used vocabulary. This common word regularization identifies keywords for interpreting text dimensions, and is able to bring in information from pre-trained word embeddings to more reliably estimate low-dimensional attitudes in small samples. We apply the method to open-ended survey responses on the Affordable Care Act and partisan animus, as well as Russian intelligence linked Twitter accounts, to evaluate whether the method produces compact text dimensions. Unlike comparison unsupervised techniques, top dimensions identified by this method are the best predictors of issue attitudes, vote choice, and partisanship. Although the method estimates issue-specific "gist" scales that differ in important ways from ideological scales trained on politicians' texts, multi-dimensionality and extremity on these scales are nonetheless associated with opinion change.

1 Introduction

Open-ended survey responses allow for unexpected discoveries and help researchers avoid inserting their own biases into findings. Gleaning systematic information from unstructured open-ended responses can be challenging, however, since many respondents use only a small number of words. For example, in one of the data sets studied here, the mean number of words in the responses is only 7 and 20% of the responses use 3 or fewer words not contained in a widely used stopword list.[1] Although the number of words used by respondents is large, most people use only a few common words.

Bag-of-words approaches, including topic models (Blei, Ng and Jordan, 2003; Blei and Lafferty, 2007; Roberts et al., 2014) and scaling models (Deerwester et al., 1990; Slapin and Proksch, 2008), can produce interpretable or predictive text models when trained on texts in which authors use many specific words. But these standard methods are designed for long and diverse text corpora, rather than short survey responses on a single issue. Because of difficulties inherent to studying general corpora, they by design do not take full advantage of information contained in common words, especially frequently used words that span many topics (Wallach, Mimno and McCallum, 2009). They also provide substantial user control, so that researchers can search for patterns in word use that might vary dramatically from data set to data set.

The task of training text models on eclectic corpora is so difficult that there is often a trade-off between a text model's fit and humans' ability to interpret the model (Chang et al., 2009; Grimmer and Stewart, 2017). In practice, some models are re-run until they "make sense" to the researcher, since there are many valid specifications and many topics to analyze. This is very likely necessary when much of a text corpora is generated by language choices that are not relevant to the researcher's interest.

This paper proposes a method to better estimate meanings for more focused text corpora that

[1] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop (SMART)

are now very commonly used in political science – short[2] and probably keyword-based text such as open-ended survey responses or certain social media post corpora, such as those using a particular hashtag on Twitter. The method is similar to scaling methods commonly used to estimate elite ideology (Slapin and Proksch, 2008; Lowe, 2016), but places greater weight on common words so that output dimensions capture more high-level summaries of text. These high-level summaries avoid over-fitting on rare words and are meant to faithfully capture vague political opinions in the public.

More technically, the method is closely related to approaches for estimating principal components when there is a persistently high ratio of variables to observations (Zou, Hastie and Tibshirani, 2006; Johnstone and Lu, 2009). In these settings, large numbers of randomly correlated variables will reduce the ability to accurately estimate a population's low-dimensional covariance structure (Johnstone and Paul, 2018).

The method, which we call pivoted text scaling,[3] uses few to no researcher defined hyperparameters to help remove the researcher from the measurement process. It scales text based on all words' associations with common words, and these common words are selected using an objective criterion – they are words that are on average written more than their accompanying words. In its simplest form, using a hard cutoff for common versus rare words, the method is simply a principal component analysis of a transformed word co-occurrence matrix, where we have subset the rows to leave only k common words and left the p columns as is, including both common and rare, accompanying words.

Other than the improved performance on short text and its interpretable, nearly fixed specification, the approach has three additional advantages: 1) it produces dimensions that can be described with a very small number of words, 2) in application, it reliably provides multiple, interpretable

[2]Although we do not provide a universal definition of a "short" text response here, we intend this method for cases where very few words are used more than once in a single text response, and where scoring only the most specific words in a corpus would remove many responses from consideration.

[3]This is a nod to (Blitzer, McDonald and Pereira, 2006), which has some similarities with the transfer leaning part of the approach used in this paper.

dimensions of attitudes, and 3) it can incorporate data that has been pretrained from larger text corpora.

We apply the method to open-ended survey responses on attitudes toward the Affordable Care Act, dislike of opposing political parties, and self-descriptions from Russian intelligence linked Twitter accounts to evaluate whether the method produces compact, meaningful text dimensions. We contrast the results with output from topic models and text scaling methods. We find that the unsupervised method is dramatically better at uncovering predictive, low-dimensional dimensions than existing unsupervised methods – models based on the identified dimensions can classify attitudes at levels similar to supervised methods. Multi-dimensionality in the responses predict opinion change. Extremity in the scales measures how closely a response conforms to the keywords on a dimension, rather than how left or right a response is, since we measure attitudes on one issue rather than coherence across multiple issues. This extremity resembles partisan strength in its association with confidence in the direction of attitudes.

2 Motivation for Political Science Research

Our motivation in developing this method is to summarize information contained in open-ended responses, supplementing analyses based on closed-ended responses. Although it is often straightforward to collect closed-ended questions, we are limited in the number of questions we can ask, we do not always know what to ask ahead of time, and it is possible that our questions will create opinions that respondents did not hold before we asked them (Zaller and Feldman, 1992). The summaries would hopefully be able to discover multiple dimensions of attitudes and do this without supervision. For example, since ACA attitudes are correlated with partisanship at 0.65 in our data, supervised methods that project words onto a single dimension will recover that variable.

These motivations help decide what technique we use to analyze the data. Topic models, such as latent Dirichlet allocation (Blei, Ng and Jordan, 2003), correlated topic models (Blei and Laf-

ferty, 2007) and stuctural topic models (Roberts et al., 2014), split documents and sets of vocabulary across distinct categories. Each text response is a mixture of these categories. This probabilistic source separation is often appropriate for long and diverse corpora and it typically requires the researcher to specify the number of categories in the data a priori.

So-called scaling or spatial methods, in contrast, compress variance in text usage onto a small number of continuous and potentially polarized variables. In political science, text scaling methods, including WordFish (Slapin and Proksch, 2008) and WordScores (Laver, Benoit and Garry, 2003; Lowe, 2007), are used as spatial "ideal point" methods, often with scales intended to be similar to those from Poole and Rosenthal's NOMINATE on roll call votes (Poole and Rosenthal, 1985). Scaling methods often do not require the user to specify the number of dimensions of the output, and the dimensions of the output have a natural ordering that is the amount of variance in the source data that an output dimension explains.

In analyzing open-ended survey data, we might prefer a text scaling method over a topic model. Some survey responses are about the same issue (i.e. the same topic), and so are hard to separate into distinct categories. In these cases, a topic model can produce few categories with little separation or find uncommon distinctions across a large number of categories. Next, scaling methods tend to have fewer researcher degrees of freedom than topic models. Well-fitting topic models naturally estimate many topics and researchers typically select only a few of the topics for study. This is especially problematic when there is no multiple testing correction.

Unfortunately, unsupervised scaling methods often fail to produce useful variables in open-ended surveys and social media posts. This paper will provide a method that constructs interpretable and predictive variables from these data.

At the same time, topic models do have certain advantages that scaling methods lack: 1) they provide high-level summaries of text and 2) they more easily provide multiple dimensions of attitudes, while scaling methods used in political science tend to reliably separate only one dimension from noise, especially when using fixed or random effects. The method here adds these advantages

to text scaling.

3 Validation

Text scaling will not generally estimate ideology when trained on focused open-ended survey responses and social media posts. In these texts on a specific issue, scaling will not in general be able to measure similarity in language choices *across* several issues, providing a generalized score to predict attitudes on all policy issues, unless scales are combined across multiple issues.[4]

Instead, pivoted text scaling creates scales for a single issue, and is intended to 1) summarize broad and interpretable variation in attitudes in a particular domain and 2) predict attitudes relevant to the same issue. For example, will someone who scores highly on a dimension describing opposition to the ACA due to high insurance prices be more likely to like the ACA if insurance prices decline?

Interpretation of polarity in the model will be concrete: someone who scores highly on a dimension either uses the keywords of that dimension, or uses words associated with the keywords for that dimension. Broadly, a person scoring highly on a dimension might be more likely to focus on a single aspect of an issue. In a sense, we estimate dimensions of attitudes for an "issue public" (Krosnick, 1990).

Importantly, pivot scaling will also describe the main variation in a data set, avoiding cases where hand-coding or supervision might accidentally focus on attitudes that are rare, specific, and/or not representative of broad and often vague public attitudes – missing the forest for the trees.

Combining the two purposes (summarization and prediction), validation of this method will focus on showing that it is an accurate and general-purpose compression of the text. Accuracy of the compression will be measured by our ability to predict attitudes in a small number of dimensions.

[4]For example, combined in a way similar to Lauderdale and Herzog (2016).

This focus on prediction using on a small number of dimensions reflects the technical focus of this article: that a small number of observations and a large number of noisy variables can impede the compression of text. The approach for addressing the problem can be implemented in other text methods, including topic modeling.

4 Technical Background and Challenges Addressed

Like most spatial methods, our proposed method for scaling open ended survey responses will be based on a matrix decomposition. In this, we would like to estimate a low-dimensional space in which document authors with distant policy attitudes will inhabit distant locations. If variation in our data is generated by distance in the latent space, then, for example, the vector explaining the greatest variation in our data estimates a first, dominant policy dimension.

Most scaling methods use some data transformation so that variation in their data is not generated by latent variables other than the latent variable of interest. Other than sample selection to limit data to political content, the most common transformation conditions on common words or social media activity and popularity. This accounts for the probabilities that any author will use a particular word or follow a particular political leader on Twitter (Barberá, 2015; Bond and Messing, 2015).

Our method will also reduce variation contributed by rare words. This transformation should improve performance when 1) variation in rare word use is orthogonal to policy attitudes, for example representing unrelated language choices, or 2) variation in rare word use is *indistinguishable* from variation due to random noise.[5]

We expect to see improved performance with this adjustment because conditioning on common words could worsen a problem fundamental to covariance estimation in wide matrices. When there are a large number of variables, p, and when p grows with the sample size, then random

[5]Another possibility in text is that two words of the same meaning will be scored less closely because they are substitutes, i.e. where the choice of which word to use is random and exclusive.

correlation in the variables will naturally generate eigenvalues and eigenvectors that are not good representations of the covariance in the population – see Johnstone and Paul (2018) for a review of principal component analysis in high dimensions. Much like existing methods for estimating principal components in samples from high-dimensional data (Johnstone and Lu, 2009), our approach boils down to the need to select a subset of the vocabulary that reliably predicts covariance not due to noise.

Overall, our method differs from methods like sparse PCA and its variants primarily by assigning document scores whether or not a respondent uses a common word that falls above the reliability cutoff, and through the addition of pre-trained word embeddings, which will not be used in the validation here.[6]

4.1 Other Related Work

Our focus on keywords means that we prioritize estimating locations for a small proportion of words, rather than many rare words. Matrix factorization techniques used in computer science, increasingly applied in political science (Rheault and Cochrane, 2019; Rodman, 2019), tend to do the opposite of this. For example, word2vec (Mikolov et al., 2013) and SVD with PPMI standardization (Levy and Goldberg, 2014) discriminate between common and rare words to obtain precise estimates for a full vocabulary. Otherwise, these models are closely related to pivot scaling, in that they involve (sometimes implicit) matrix factorization.

Of course, orienting around common words probably ignores subtleties and idiosyncracies in sophisticated text. However, this relative ignorance allows us, we hope, to produce interpretable vector representations.

[6]Pre-trained "word embedding" matrices that compress word co-occurrences to a low-dimensional space based on their usage in a large text corpus. These matrices are similar to those produced by LSA or PCA (Levy and Goldberg, 2014), but typically estimate local/low-level similarity (e.g. synonyms or very close word associations) rather than high-level word distances ("government" → "shoe").

5 Approach

We begin with a standard scaling method, where we estimate the greatest variation in word co-occurrences after conditioning on a word's tendency to appear with other words in general. Below, M is a document-term matrix, where rows are documents and columns are words. G is a the word co-occurrence matrix of M with a square root transformation, that is $G = (M^\top M)^{\circ\frac{1}{2}}$, where $\circ\frac{1}{2}$ is the element-wise square root. A standard scaling method estimates k orthogonal dimensions of

$$X = \|g_i\|^{-1} G, \tag{1}$$

where $\|g_i\|^{-1}$ is row-wise division by the sum of each row. This division creates a matrix where the Euclidean norm of each row is equal to one, and conditions on how often a word is used with other words.[7]

The eigenvectors representing ideological or policy attitude dimensions are typically estimated using singular value decomposition and are ordered by the amount of variance they explain in word co-occurrences:

$$U_k \Sigma_k V_k = SVD(X), \tag{2}$$

where $U_k\Sigma_k$, or equivalently XV_k, maps X to the k-dimensional latent space, or principal components, and which can be ordered by variance explained.[8] In some models (as is done here due to the element-wise square root transformation) $j = 1$ is a random effect for word frequency and $j = 2$ is the first substantive dimension. For example, U_2 estimates y_1 a first, latent dimension:

[7] The Euclidean norm here allows us to estimate a random effect on the first dimension of our decomposition, and does still place greater weight on words that appear with a *variety* of words (forcing greater separation in these vague words). This square root transformation also allows us to more easily compare our results to a null matrix, where words are used randomly with an expected frequency. In that null model, the first eigenvalue will approximately equal our first eigenvalue, and subsequent eigenvalues will diverge.

[8] X here is technically centered, but covariance can be closely (and much more quickly) approximated without centering in our method.

$$y_j \propto (\text{-})U_{j+1}. \tag{3}$$

This approach is closely related to the factorized matrices in topic models (Roberts, Stewart and Tingley, 2016) and existing text scaling methods, including LSA (Deerwester et al., 1990) and correspondence analysis (Lowe, 2007; Bonica, 2014). Our method adjusts this scaling so that it relies less on words that are likely to be noisy.

In our adjusted scaling, we select a subset of the vocabulary that is more likely to be reliable – words that are used very often. This reliability for common words is visible in a principal component analysis of the original G word co-occurrence matrix, where the eigenvalues and associated eigenvectors are closely related to the inverse-rank frequency distribution of words – the most common word loads primarily on the first dimension, the second on the second, etc..

With this subset, we would like to penalize words' contribution to variance when noise begins to exceed reliable signal. Here, given our interest in an interpretable model with few researcher degrees of freedom, we will focus on a single solution that is interpretable. We select a subset using words, which we will call pivot words, that appear more often than their accompanying words:

$$\|x_j\| > 1, \tag{4}$$

meaning that keywords (pivot words) in X have column Euclidean norm greater than one.[9] These words are usually a small fraction of a dataset's vocabulary – perhaps 10%, for example, and hopefully declining with increasing sample size.

Our subset scales all words around those common words, and we assign the remaining words word/"gist" scores based on their associations with the common words.

[9]This sum will be approximately the same as a version of the X matrix without the diagonal.

5.1 Implementation: canonical correlation analysis

We implement the approach described above using a variant of canonical correlation analysis sometimes called orthonormalized partial least squares (Rosipal and Krämer, 2006). Our method in its simplest form – using a hard cutoff for common versus rare words – is simply a principal component analysis of a transformed word co-occurrence matrix, where we have subset the rows to leave only k common words and left the p columns as is, including both common and rare words. The left singular vectors of $SVD(X_{k,p})$ give us keyword/pivot scores, while the right singular vectors project our original X matrix to create word/"gist" scores that use to project our term-document matrix into a low-dimensional space. [10] The CCA procedure implements a smooth cutoff, permits the inclusion of out-of-sample word embeddings, and provides a framework for assigning both "gist" scores and "pivot" scores.

Since the step-by-step CCA implementation is less important than the high-level description above, we include the full details of the implementation in the appendix. Table 4 in the appendix summarizes the steps. The method is implemented in the R package available here: `https://github.com/wilryh/parrot`.

As a sketch of those steps, a matrix Z is used to adjust the standard text scaling on X. We subset to approximately $2k$ pivot words by truncating to the top k principal components of $Z = D_g^b X$, where b is a large exponent (e.g. 2 or greater) and D_g is the diagonal of the unstandardized word co-occurrence matrix (essentially, the number of times a word was used in the corpus). When not using word embeddings, the exact value of b, at least beyond squaring word counts, has little effect on results, but larger values of b linearize the association between keywords/"pivot" scores and word/"gist" scores.

With inverse-rank frequency word counts, the truncation of Z creates a logistic function that smoothly separates pivot words from the remaining words. The truncation k and the exponent

[10] PCA on $X_{k,p}$ avoids imaginary values in the eigenvectors of $X_{k,p}$, but is otherwise very close to the real parts of that decomposition.

b control the shape of that function. We can add out-of-sample word embeddings to our data by replacing X with a word embedding matrix W and using the original word counts rather than exponentiated word counts. This introduces the out-of-sample data around the smooth cutoff, where our data is less reliable, and leaves them below the most common words in our data to keep our word score estimates close to our own data.

6 Application to Open-Ended Surveys on the ACA

We now apply our adjusted scaling to study its behavior on open-ended survey responses and social media data. We begin with data on the Affordable Care Act.

Over 9,000 open-ended responses on the ACA were collected by the Kaiser Family Foundation between 2010 and 2015. We add to this data approximately 2,000 responses in 2009 from a survey conducted by Pew, along with 1,000 responses in 2016 from a national representative sample conducted by researchers at the University of Pennsylvania. The panel respondents are not included in the word score training, except in the word embedding illustration shown in the appendix.[11] The KFF and Pew data sets are publicly available and have been analyzed using topic models in prior work (Hopkins, 2017).

In the data, the KFF and 2016 panel respondents were asked two questions at the beginning of a longer survey on health care policy attitudes (KFF) and the 2016 election (panel). The first two questions were: 1) "As you may know, a health reform bill was signed into law in 2010. Given what you know about the health reform law, do you have a generally favorable or generally unfavorable opinion of it?" 2) "Could you tell me in your own words what is the main reason you have a favorable/unfavorable opinion of the health reform law?". Around 2,000 thousand respondents were asked two similar questions by Pew before the ACA was signed into law.[12]

[11] Around 4 out of 5 survey takers responded to the open-ended questions in the KFF and Pew samples and 9 out of 10 in the panel sample.

[12] Closed-ended: "As of right now, do you generally favor or generally oppose the health care proposals being discussed in Congress?". Open-ended: "What would you say is the main reason you favor or oppose the health care

We use pivot scaling on this data without word embeddings for the comparisons to other methods, since 1) competing methods do not use pre-trained word embeddings and 2) pivot scaling becomes more subjective if we add choices about which pre-trained word embeddings to incorporate. We provide an example of the process using "meta" word embeddings (Yin and Schütze, 2016) in the appendix, where we use 2018 only data ($\sim$ 700 responses) to reconstruct the history of ACA attitudes.

In the ACA data, $b = 2$ is sufficient to induce pivoting (see Figure A2 in the appendix), and we are able to use b as large as 4 before the matrix is no longer invertible. Results below will use this $b = 4$ parameter. Results are insensitive to its precise setting. To improve interpretability, all dimensions from the model are set so that positive sides of dimensions favor the ACA while negative sides of dimensions oppose it.

We first show the keywords from the top 2 dimensions of our output in Table 1. The keywords here are a word's pivot scores on a dimension multiplied by its total activation (Euclidean norm of pivot scores of all k dimensions). We named the dimensions ourselves. These keywords appear to be highly informative. They pick up both specific components of ACA policy and broad opinions on it. In appendix Table A2, we show example responses that scored highly and uniquely on one of the top two dimensions. Although not perfect, of course, these examples suggest that the method works well on the document level.

6.1 Visualizing Pivot Scores

Because pivots are central to our method, as keywords and regularizers, we next visualize the distributions of pivot scores in our data.

The two panels of Figure 1 show that the pivot words have the same scores in projections to the shared space for the two views (pivot and overall) of the data. The rest of the words appear in only the overall word scores. We display the same information using an alternate visualization in

proposals being discussed in Congress."

Keywords			
Dimension 1		*Dimension 2*	
"role of government"	"patient protection"	"personal cost"	"universal access"
Anti	Pro	Anti	Pro
involved	conditions	medicare	step
socialism	existing	cut	universal
government	parents	premiums	access
telling	pre-existing	went	available
run	condition	age	world
unconstitutional	allows	keep	direction
federal	children	medicaid	america
anything	age	benefits	states
much	kids	doctors	provides
medicine	stay	wont	affordable
socialized	covers	paying	everyone
want	access	away	needed
economy	insured	kids	needs
business	provides	theyre	right
everything	helps	pay	important

Table 1: *ACA Keywords.* This table shows the keywords in the top two dimensions of the open-ended responses on favoring or opposing the Affordable Care Act.

appendix Figure A1.

6.2 Predictive Evaluation and Comparisons to Other Continuous Methods

We evaluate the dimensions of the pivot scaling output by predicting responses to the closed ended question in the surveys on favorability toward the Affordable Care Act. This is a straight-forward mapping to the closed-ended response.[13]

In the evaluation, we predict the closed ended response on favorability toward the Affordable

[13] The prediction based approach allows us to get around complicated survey based evaluations of our scores. In evaluations that test whether people are able to group words in the same way as a model, for example, we would likely perform better if we estimate a high number of dimensions on which clusters of words are tightly packed. These tightly packed clusters would be easy for non-experts to associate. However, the large number of clusters would no longer be in line with our goal of producing latent and low dimensional representations of attitudes. The low dimensional representations should link loosely associated arguments.

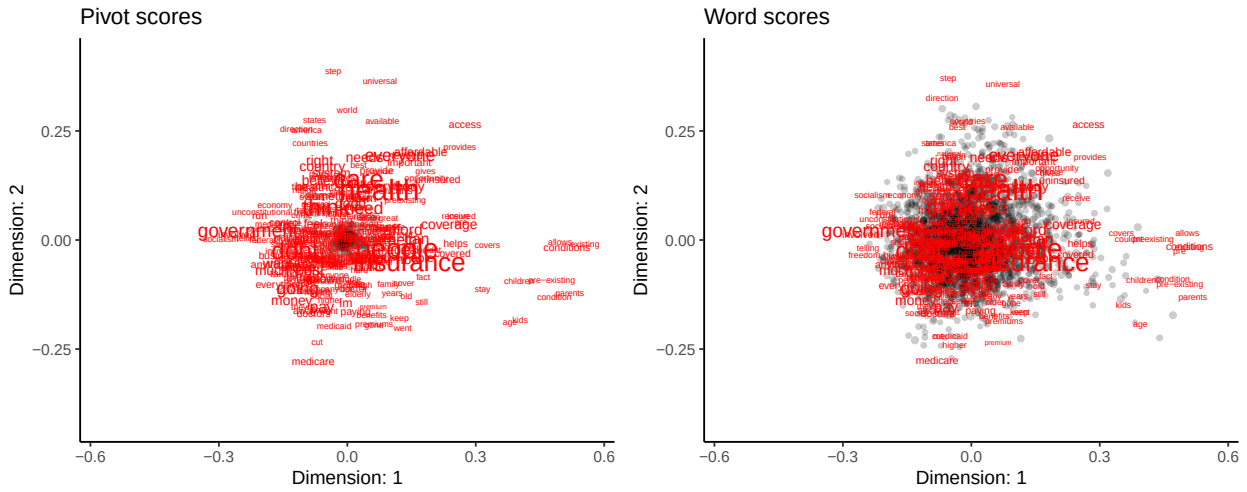


Figure 1: *Relationship between word scores and pivot scores.* This figure shows the linear correspondence between pivot scores (used for keywords) and overall word scores (used to score documents) when parameter b is large. The left panel shows pivot scores while the right shows word scores.

Care Act using a Lasso (Tibshirani, 1996). This is a penalized regression that selects the best independent predictors of ACA attitudes from our text dimensions.[14] To compare the coefficients, we first scale each dimension so that each variable in the regression has the same variance. The coefficients from this model are then the additive dimensions of attitudes toward the Affordable Care Act, and the size of the coefficient reflects a dimension's importance in prediction. All methods (both unsupervised and supervised components) were trained on 90% of data and tested on a 10% holdout set.

Figure 2 shows the performance of comparison methods in our data. For all but the topic models, the x axis denotes a regression on the first n dimensions of a method's output. For topic models, we asked the method to return the number of dimensions then predicted using that output. The topic model results are shown at the number of topics minus 1. For GloVe (Pennington, Socher and Manning, 2014) – a method for training word embeddings on large data sets –, we modeled on 50 dimensions, where "GloVe" uses these 50 dimensions (which do not have an order) and "Glove

[14] Our Lasso uses the defaults in the 'glmnet' package (Friedman, Hastie and Tibshirani, 2010). The glmnet Lasso function selects the regularization level using smallest cross-validated error.

+ PCA" uses the output of a principal component analysis on those 50 dimensions. The "PCA of X" comparison uses a principal component on the standardized word co-occurrence matrix. This outperforms the principal components of the term-document matrix. [15]

In addition to the unsupervised method, we also compare pivot scaling to a method supervised on the outcome from the start – one that trains thousands of predictors on the outcome, rather than a handful. We use multinomial inverse regression (Taddy, 2013) for this purpose.[16] This method produces a dimension that is closely related to ACA favorability by design, and that suggests a maximum accuracy we can expect for linear classifiers.

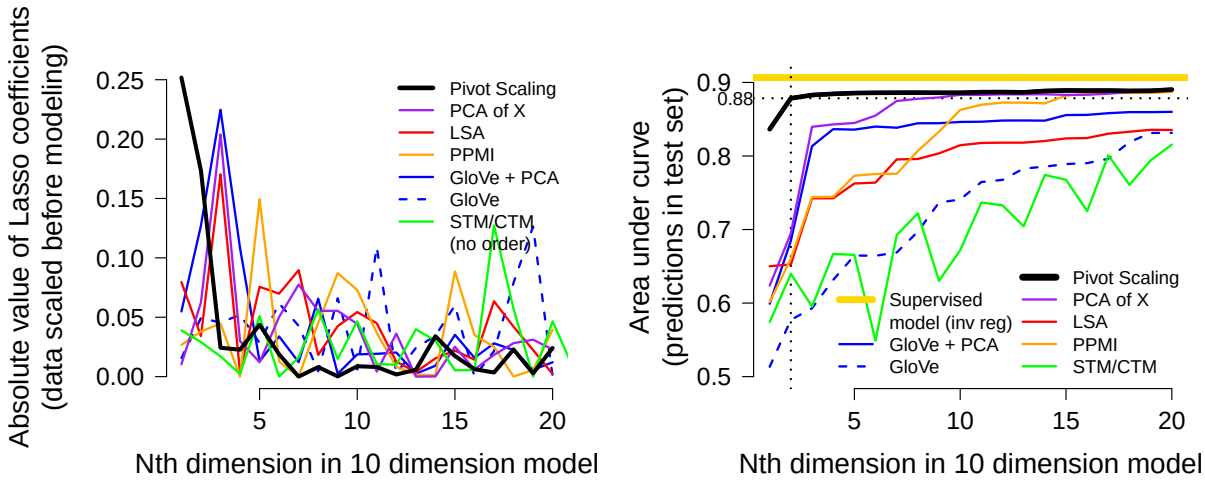


Figure 2: *Dimensionality of other methods.* The left panel of this figure shows the absolute value of coefficients from a 20 dimensional penalized regression predicting ACA favorability from each method's output. Predictive dimensions are concentrated in the first dimensions of pivot scaling, while predictive dimensions are spread across the output from other methods. The right panel of this figure shows area under the ROC curve for 20 penalized regressions predicting ACA favorability from each method's output.

Figure 2 shows that the Lasso is using substantially lower dimensional information in pivot scaling than in all other comparison methods. In the left panel, we show the absolute value of

[15] The comparisons for 'LSA' and 'PCA' are close matches to methods used in political science, such as Wordfish and correspondence analysis (Lowe, 2016). These methods happen to perform worse on our data than their typical performance, however, possibly due to some common words appearing with no other words. This lack in co-occurrences is washed out in our method, but is picked up as important variation in these other methods.

[16] We use gamma=1, however, different penalizations do not appreciably improve its performance, including the Lasso on all words.

coefficients from a 20 dimensional Lasso predicting ACA favorability in our data. This shows that a Lasso chooses the first dimensions of pivot scaling's output, but chooses higher dimensions from other methods. A researcher that uses these methods would need to justify their high dimensional choice in later analyses, while pivot scaling provides a hands off and useful ordering.

The right panel shows the area under the ROC curve for the first 10 regression models from each method in Figure 2. Pivot scaling is capable of condensing much of the information in the text responses into a small number of variables. Given this low dimensionality, a researcher will not need to select one out of many possible variables for later analyses.[17]

Topic models can recover similar dimensions as pivot scaling, but perform worse on both low and high dimensional prediction in our data. Selecting any number of 50 or fewer topics in a correlated topic model (see Figure 2 and Figure A3 in the appendix) will all return dimensions that predict ACA attitudes at lower predictive accuracy than the first 2 of our dimensions. Selecting the number of topics automatically using information criteria will give around 50 topics.

We show similar dimensions to pivot scaling from a correlated topic model in appendix Tables A5 and A6, where we have run several models and chosen 4 topics (3 dimensions) so that the results look somewhat like our output.

6.3 Correlates with Other Survey Responses

In addition to the open-ended survey responses collected by the Kaiser Family Foundation and Pew Research Center between 2009 and 2016, we also have a smaller collection of responses to the same question in a panel survey from Institute for the Study of Citizens and Politics (ISCAP) at the University of Pennsylvania. This survey has a larger variety of closed-ended responses and has these responses going back several years.

[17]Note that the topic models performed especially poorly. This is surprising because the models should more closely resemble the PCA output. We observed greater similarity in some analyses for this paper when using an older version of the "stm" package and, more importantly, a training set that included non-public, open-ended responses from political activists (who were more polarized and wrote longer responses). We do not use the non-public training data in this version to increase replicability and representativeness of the data.

We applied the word representations to this nationally representative sample to see whether the top dimensions were correlated with vote choice and change in ACA attitude after controlling for partisanship. Partisanship is by far the best predictor of both vote choice and ACA attitudes. Finding that a variable is significantly correlated with a political outcome after including partisanship as a control is a high standard.

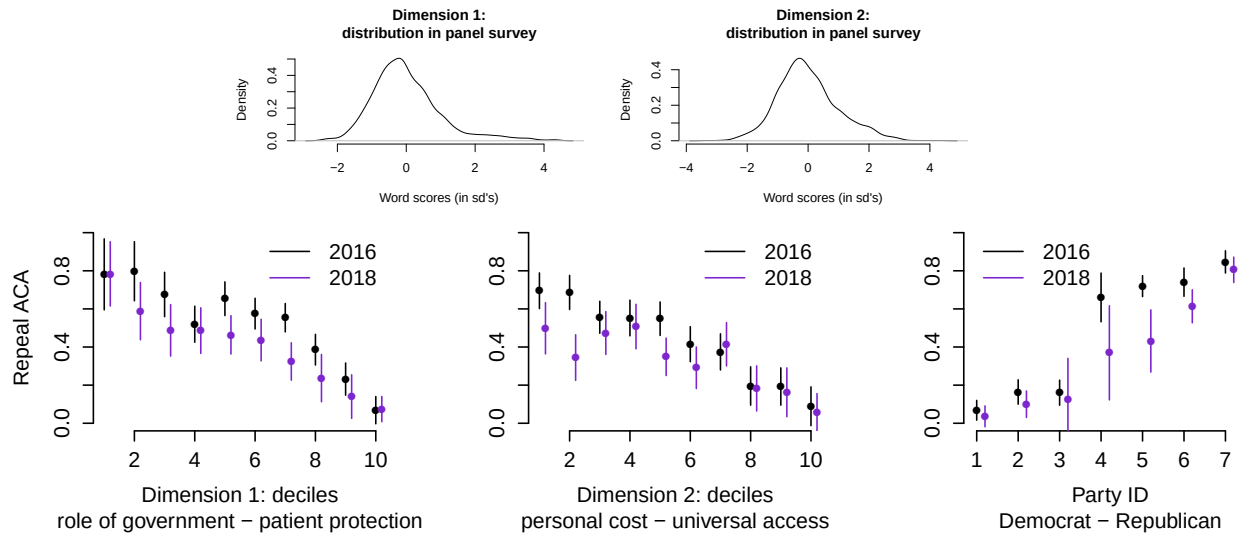
Table 2 shows that the first dimension of our output is correlated with 2016 vote choice, controlling for 2016 partisanship (model 3). The 2nd dimension but not the first is associated with whether a Republican voted for Donald Trump instead of Republican establishment candidates (4). Though the difference in these coefficients for Donald Trump vs Republican establishment is suggestive, and it is not altered even controlling for ACA attitudes, it is not statistically significant at conventional levels ($p = 0.19$ without ACA favorability control, 0.10 with ACA favorability control – see appendix Table A4).

The 2nd dimension is similarly associated with changes in ACA attitudes from 2012 to 2016 (1). People who talked about personal costs, rather than universal access, were more opposed to the law. This is an important change because major components of the ACA were only implemented after 2012. This association was partially reversed in 2018 (2). People who talked about costs in 2016 were less likely to favor repeal in 2018 compared to 2016.

6.4 ACA Attitudes Over Time

In Figure 3, we show changes in the mean of each of the top two dimensions over time based on the Kaiser Family Foundation and Pew Research data 2009 through 2015. We plot separate time series for respondents who stated favorable or unfavorable in the preceding closed-ended response. The error bars are bootstrapped 95% confidence intervals.

These changes over time align with 1) the ACA being signed into law in 2010 and 2) the implementation of major components of the law. In the top right panel, for example, we see that respondents who felt favorably about the law had not yet started to discuss specific policies prior to



	Dependent variable:			
	Repeal 2016 vs Repeal 2012	Repeal 2018 vs Repeal 2016	Donald Trump vs Hillary Clinton	Donald Trump vs Republican establishment
	(1)	(2)	(3)	(4)
Dimension 1	−0.10	0.10	−0.08	−0.04
(+ patient protection vs	(0.07)	(0.08)	(0.02)	(0.04)
− role of government)				
Dimension 2	−0.27	0.15	−0.03	−0.11
(+ universal access vs	(0.06)	(0.07)	(0.02)	(0.04)
− personal cost)				
Dimension 0	−0.02	−0.20	0.01	0.03
(+ general vs − specific)	(0.06)	(0.07)	(0.02)	(0.04)
Observations	865	575	1,023	428

Table 2: *Between and within party correlates in small panel survey.* All models in the table control for party identification – 2012 party identification for model 1 and 2016 for the other models. Independent variables are scaled so that a unit change correspondents to a one standard deviation change in a given subset of the data. The dependent variables in the vote choice models are coded 1 if the respondent voted for the first candidate (Donald Trump) and coded -1 if the respondent voted for the second candidate(s). Republican "establishment" candidates are: Jeb Bush, John Kasich, Marco Rubio, and Chris Christie. Vertical lines in the decile figure are 95% confidence intervals, and repeal ACA is coded 1 when a response exceeds 4 on a scale of 1 to 7.

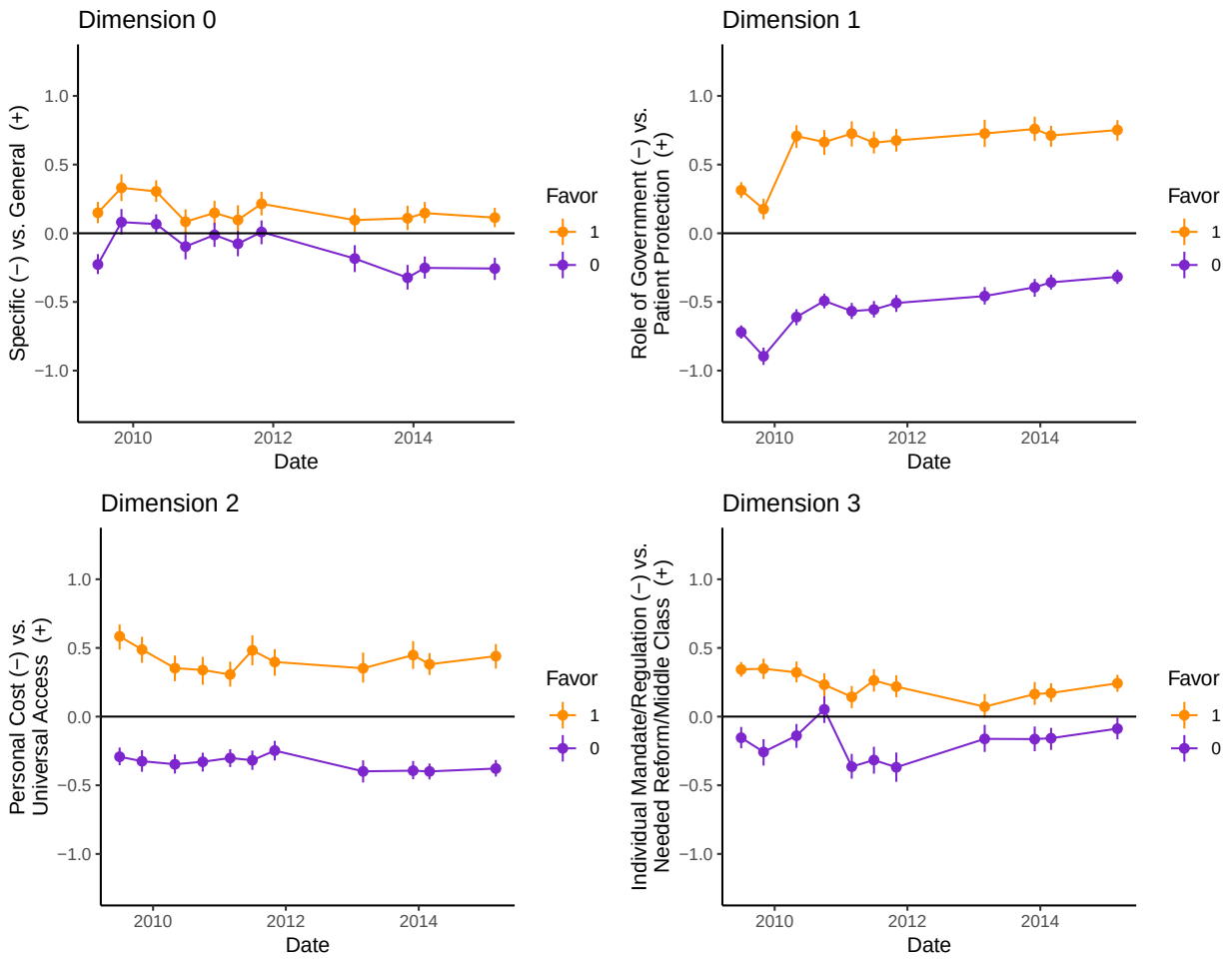


Figure 3: *Changes in ACA justifications over time.* This figure shows the 2009 through 2015 means of the 1st, 2nd, and 3rd dimensions of our text scaling (the 0 dimension is word frequency) separated by respondents who felt favorably or unfavorably about the ACA. The error bars are 95% confidence intervals based on bootstrapped standard errors.

its signing, including changes benefiting people with pre-existing conditions and allowing young people to stay on their parents' health insurance. In the top left panel, we see that respondents who oppose the law used more specific words after its implementation.The bottom left panel shows that an emphasis on personal cost vs universal access was relatively constant over time, with perhaps a very slightly greater emphasis on personal cost from 2013 on.

7 Applications to Other Data Sets

We covered the ACA responses in depth to illustrate our method in an unusually large data set of open-ended survey responses. We now show that the method can be applied more broadly using two public data sets that were released in late 2018.

7.1 Partisan Animus

The first application comes from the 2016 American National Election Study. We discovered this data from a tweet in October 2018.[18] The text is in response to a question "Is there anything you dislike about the Democratic/Republican Party?".

The goal of the application is to see if the method proposed here can pick up partisan animus. With this goal in mind, we removed partisans in the data who are talking about their own party or, in keeping with the tweet, who did not rate the opposing party below 40 on a thermometer scale from 0 to 100. We also required that words appeared more than twice for the results shown below, since this improved almost all methods, both unsupervised and supervised. We show the results including words appearing more than once – a default in the widely used "stm" R package – in the appendix. Keywords in that model are more specific, but the interpretations of dimensions do not change.

We repeat the analysis in Figure 2 to see if a Lasso chooses the top dimensions of pivot scaling to predict partisan identification. We show the results in Figure 4 and show the keywords for the top 2 dimensions in Table A9 in the appendix. The first dimension separates respondents talking about politics vs. policy and is not strongly related to partisanship, while the second separates partisans. This figure shows that pivot scaling again substantially outperforms all other unsupervised methods in low-dimensional predictions and is not much worse than even a supervised model.

[18] https://twitter.com/MattGrossmann/status/1050826155987795968

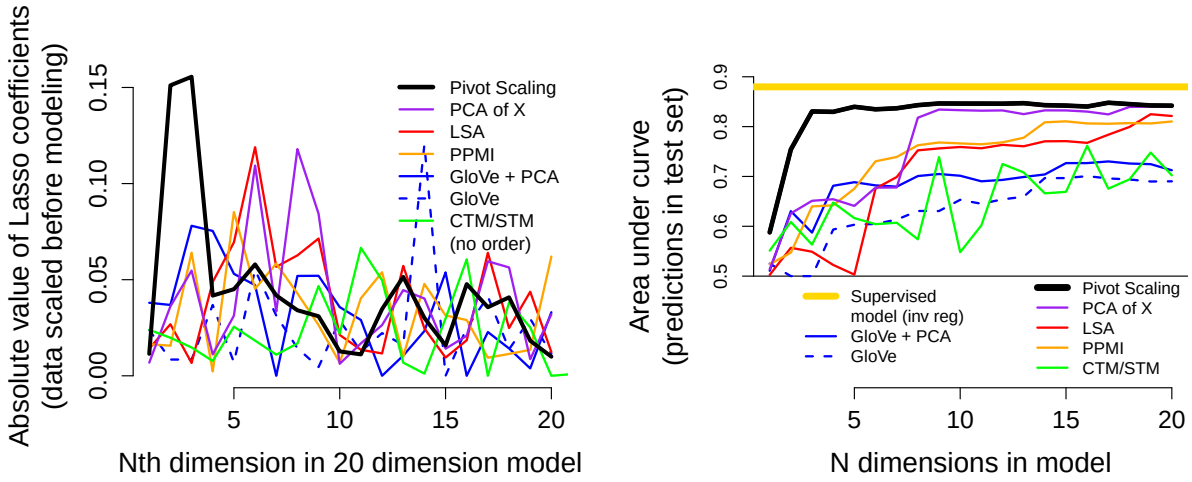


Figure 4: *Partisan animus.* The left panel of this figure shows the absolute value of coefficients from a 20 dimensional penalized regression predicting party identification in the ANES, while the right shows the AUC for performance in the 10% test set by the number of dimensions included in the Lassos (or multinomial inverse regression for the supervised comparison). Predictive dimensions are concentrated in the first dimensions of pivot scaling. The first dimension separates respondents talking about politics vs. policy, while the second separates partisans. We include the keywords from these top 2 dimensions in appendix Table A9.

7.2 Russian Trolls

The final application in this paper comes from data on Russian linked Twitter accounts released by Twitter in October 2018.[19] We include this application here to illustrate the capabilities (broad/high-level analysis, computational scalability) and limitations (pre-processing, sample selection required) of this method on social media data. The substantive goal is to simply document broad patterns in the campaign to influence the 2016 election, focusing on content originating from the trolls and avoiding sensational but extremely rare and/or non-representative messaging.

Given this focus, we exclude post-2016 election and non-English accounts from the analyses, in addition to control characters.[20] This removes many of the troll tweets, leaving one million tweets from 2016 of which only 700 thousand tweets using non-unique language were from English-only

[19] https://about.twitter.com/en_us/values/elections-integrity.html#data

[20] Accounts were considered non-English if their account language was set to a language other than English, or if their account description contained non-Emoji Unicode from the supplementary multilingual plane.

accounts. We also consider only those tweets that were not retweets of other accounts, and that did not come from news aggregators.[21] Removing the news aggregators accounts for extreme levels of bot-like activity from the news accounts, and we score those accounts based on activity from the remaining accounts.

Since much of the trolls' activity occurred in 2015[22] or after the 2016 election, was not in English, and/or were retweets of non-trolls, this leaves around 130 thousands tweets originating from just over 600 accounts in the analysis below.[23] With this data, we replicate a hand-coded typology (Linvill and Warren, 2018)[24], track the polarity of those clusters' gist scores over time (how much they conform to the keywords), and then point out some considerations when using this method on social media data.

Given the relatively large data set, we use the hard cutoff version of pivoted text scaling, which is simply a principal component analysis of the X matrix with rows truncated to only the top k words. Pivoted text scaling with a hard cutoff can be estimated in 8.5 seconds on this 130 thousand document data with a vocabulary size of 140 thousand words/tokens.[25]

In Figure 5, we show that the top two dimensions of this text scaling correspond to the hand labeled account descriptions from Linvill and Warren (2018). This correspondence is spread over two or more dimensions of the output, and combining two dimensions creates either a single left-right dimension or a hashtag vs news dimension. The panels in the top right show the overlap in the tweets from the left vs right troll accounts and the hashtag gamer vs news feed accounts.

The bottom panel of Figure 5 suggests that the "right troll" accounts became gradually more

[21] Linvill and Warren (2018) classify news aggregators in their publicly available data set. We have also replicated those classifications using the account descriptions in Twitter's own released data. There, the accounts use variants of "breaking news" (e.g. San Jose's breaking news) to describe themselves.

[22] Many of these tweets were follow requests sent from trolls to non-trolls.

[23] We undertake a more complete analysis of this in a separate substantively focused paper. The complete analysis is relatively complex and considers messaging, counts of activity over time, and network clustering.

[24] Three fourths of the accounts from their top categories appear in our 2016 only analysis.

[25] The cutoff k set to one half the number of words appearing more often than their accompanying words is likely too large in massive data sets, but we have not observed differences in our results for speculative, smaller k's (e.g. the square root of the vocabulary size times e) that grow more slowly with vocabulary size.

focused on the election and tweeted more support for Donald Trump as the election approaches, while the "left troll" accounts were largely consistent in their messaging, at least in their focus on police shootings instead of the 2016 election. The non-political accounts suddenly tweeted content consistent with interest in the 2016 election and support for Donald Trump only in the last week of the campaign. Overall, however, both categories of accounts were roughly stable in their messaging over time when regularizing rare word usage.

In addition to the above multi-dimensionality considerations, it is worth noting that although the dimensions of this method are approximately orthogonal, this does not mean that changes in these dimensions would not, for example, be correlated over time. In the first dimension of the Twitter data, we show that both news aggregators and left-leaning trolls talk about the police – but talk about the police in different ways, as can be seen on the second dimension. Changes in how much people talk about the police can affect both the first and second dimensions.

Further, as a cautionary point on over-interpretation of the largest variation in a social media data set, we note that social network data can be heavily skewed toward very active users, and that scaling tweets does not pick up the largest variation at the *user* or *follower* level. Removing the news accounts allows us to analyze variation in the larger number of accounts with many followers that are not spamming content. To an extent, this resembles the removal of uninformative votes from roll call data prior to scaling ideology in legislatures. At the same time, this pre-processing re-introduces research control over the output, much like sample selection in research more generally. Addressing this pre-processing and sample selection in a more principled way is beyond the scope of this single article, but we hope to address these and related issues with translating eclectic corpora into focused corpora in the near future.

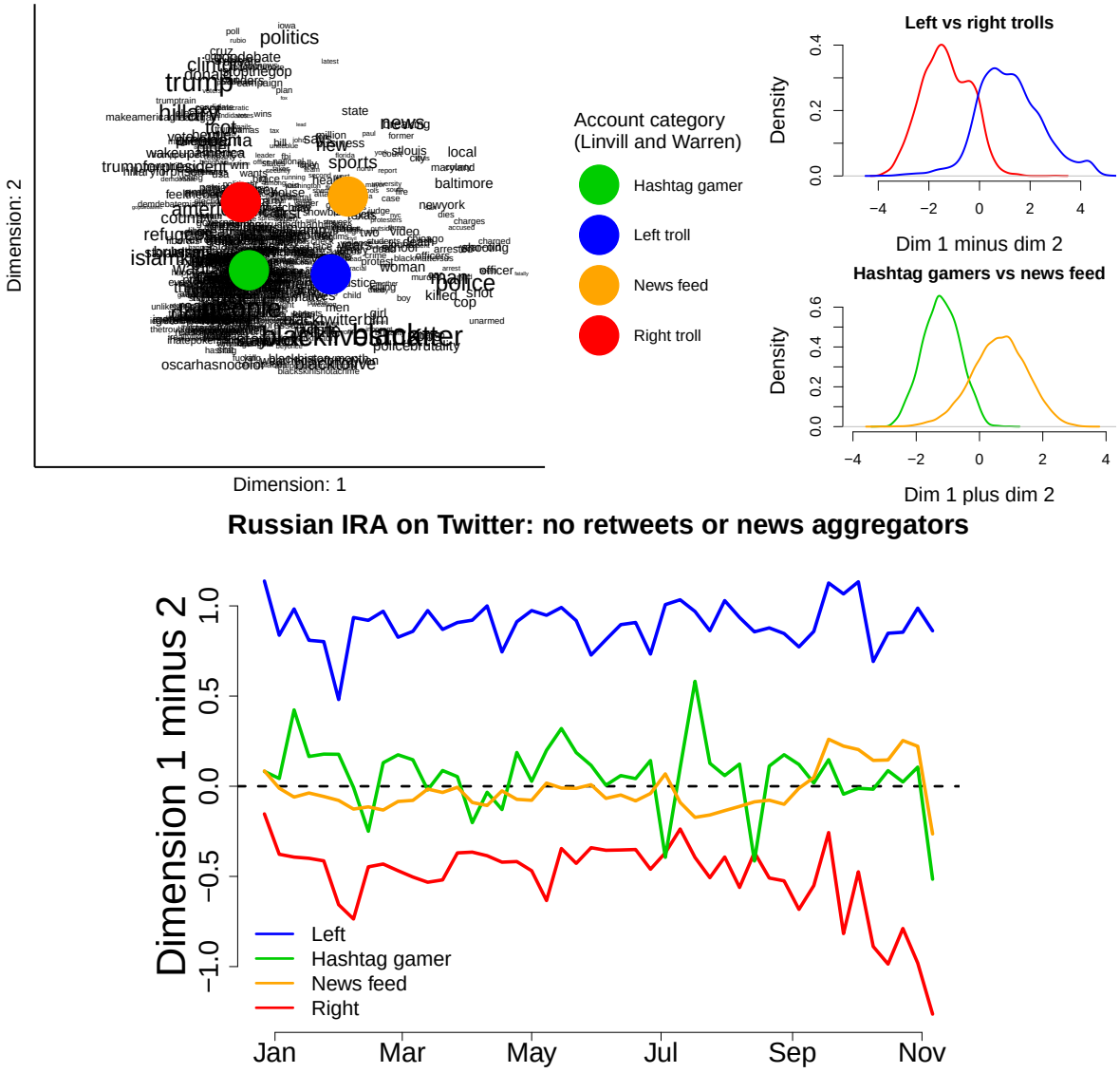


Figure 5: *Russian intelligence linked Twitter accounts.* The top-left panel of this figure shows the word scores and the average document scores in account categories of Linvill and Warren (2018). The top-right panel shows the tweet-level overlap in scores for left vs right trolls (dimension 1 minus dimension 2) and hashtag gamers vs news feed accounts (dimension 1 plus dimension 2). The bottom panel tracks the average document scores over time within each account category, showing that the right trolls, hashtag gamers, and news feed accounts tweeted more consistently about the 2016 election and in support of Donald Trump near election day, while the left trolls were consistent over time in talking about police shootings and Black Lives Matter rather than the 2016 election. Keywords for the dimensions are shown in Tables A11 and A12 in the appendix.

8 Discussion

Pivoted text scaling provides ordered and interpretable representations of short text data, along with keywords to help evaluate results. Its output substantially outperforms existing techniques on low-dimensional predictions. The top dimensions from pivot scaling further correspond to intuitive explanations for individuals' changing justifications for supporting or opposing the Affordable Care Act – pre and post implementation of the law– as well as some perhaps unexpected political cleavages among supporters of American political parties.

The agreement on the topic in an open-ended survey thus seems to allow respondents to repeat a small set of vocabulary in meaningful patterns. Our departure from prior work is due to our specific interest in representing short and focused texts in an interpretable way. This method, then, will not be well-suited to all texts and purposes. In particular, latent Dirichlet allocation (Blei, Ng and Jordan, 2003) and correlated topic models (Blei and Lafferty, 2007; Roberts et al., 2014) will likely outperform this method when text is particularly diverse, and for which very common words will not necessarily be useful starting points for scaling the text.

Existing unsupervised methods for long form political text (Slapin and Proksch, 2008; Lauderdale and Herzog, 2016) may also outperform this method when most speakers in a corpus use many specific words, as these provide confidence intervals for individuals. Other methods in political science, such as semi-supervised methods that use hand labels (Benoit et al., 2016), methods that include other indicators of political preferences (Kim, Londregan and Ratkovic, 2018), and supervised methods (Laver, Benoit and Garry, 2003; Lowe, 2007; Beauchamp, 2012; Taddy, 2013), will also be well-suited for tasks like measuring ideology in legislatures.

List of Figures

Acknowledgements

The author appreciates comments and feedback on this method from Adam Bonica, Nick Beauchamp, Chris Callison-Burch, James Fowler, Lisa Friedland, Seth Hill, Dan Hopkins, Gary King, Kokil Jaida, Kenny Joseph, Jacob Montgomery, Ani Nenkova, Molly Roberts, Brandon Stewart, and Lyle Ungar.

Funding

This project was generously supported by the Russell Sage Foundation (grant 94-17-01), as part of a project on attitudes toward the Affordable Care Act.

References

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data." *Political Analysis* . 6

Beauchamp, Nicholas. 2012. "Using Text to Scale Legislatures with Uninformative Voting." pp. 1–44. 25

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *The American Political Science Review* 110(2):278–295. 25

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022. 1, 3, 25

Blei, David M and John D Lafferty. 2007. "A correlated topic model of science." *The Annals of Applied Statistics* 1(1):17–35. 1, 3, 25

Blitzer, John, Ryan McDonald and Fernando Pereira. 2006. "Domain Adaptation with Structural Correspondence Learning ." *EMNLP* pp. 120–128. 2

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(01):62–78. 6

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science* 58(2):367–386. 9

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang and David M Blei. 2009. "Reading tea leaves: How humans interpret topic models." *NIPS* . 1

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6):391–407. 1, 9

Friedman, Jerome H, Trevor Hastie and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33(1). 14

Grimmer, Justin and Brandon M Stewart. 2017. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297. 1

Hopkins, Daniel J. 2017. "The Exaggerated Life of Death Panels: The Limited Influence of Elite Rhetoric in the 2009-2012 Health Care Debate." *Political Behavior* 40(3):681–709. 11

Johnstone, Iain M and Arthur Yu Lu. 2009. "On Consistency and Sparsity for Principal Components Analysis in High Dimensions." *Journal of the American Statistical Association* 104(486):682–693. 2, 7

Johnstone, Iain M and Debashis Paul. 2018. "PCA in High Dimensions: An Orientation." *Proceedings of the IEEE* 106(8):1277–1292. 2, 7

Kim, In Song, John Londregan and Marc Ratkovic. 2018. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* 26(2):210–229. 25

Krosnick, Jon A. 1990. "Government policy and citizen passion: A study of issue publics in contemporary America." *Political Behavior* 12(1):59–92. 5

Lauderdale, Benjamin E and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* . 5, 25

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331. 4, 25

Levy, Omer and Yoav Goldberg. 2014. "Neural word embedding as implicit matrix factorization." *NIPS* . 7

Linvill, Darren and Patrick Warren. 2018. "Troll factories: The internet research agency and state-sponsored agenda building.". 22, 24

Lowe, Will. 2007. "Understanding Wordscores." *Political Analysis* 16(04):356–371. 4, 9, 25

Lowe, Will. 2016. "Scaling Things We Can Count.". 2, 15

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality." *NIPS* . 7

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." *EMNLP* 14:1532–1543. 14

Poole, Keith T and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science* pp. 357–384. 4

Rheault, Ludovic and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* 78:1–22. 7

Roberts, Margaret, Brandon Stewart and Dustin Tingley. 2016. "stm: R Package for Structural Topic Models." *Journal of Statistical Software* . 9, 7

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082. 1, 4, 25

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 5:1–25. 7

Rosipal, Roman and Nicole Krämer. 2006. Overview and Recent Advances in Partial Least Squares. In *Subspace, Latent Structure and Feature Selection. SLSFS . Lecture Notes in Computer Science, Volume 3940*, ed. Saunders C, Grobelnik M, Gunn S and Shawe-Taylor J. Berlin, Heidelberg: . 10, 1, 3

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722. 1, 2, 4, 25

Taddy, Matt. 2013. "Multinomial Inverse Regression for Text Analysis." *Journal of the American Statistical Association* 108(503):755–770. 15, 25

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B (Methodological)* 58(1):267–288. 14

Wallach, Hanna M, David M Mimno and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." *NIPS* pp. 1973–1981. 1

Yin, Wenping and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: pp. 1351–1360. 12, 17, 19, 20

Zaller, John and Stanley Feldman. 1992. "A simple theory of the survey response: Answering questions versus revealing preferences." *American Journal of Political Science* pp. 579–616. 3

Zou, Hui, Trevor Hastie and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15(2):265–286. 2

Online Appendix for "Text Scaling for Open-Ended Survey Responses and Social Media Posts"

William R. Hobbs

A Online Appendix

Table of Contents

A.1 Implementation

We implement pivot scaling using a variant of canonical correlation analysis sometimes called orthonormalized partial least squares (Rosipal and Krämer, 2006). Our method in its simplest form – using a hard cutoff for common versus rare words – is simply a principal component analysis of a transformed word co-occurrence matrix, where we have subset the rows to leave only k common words and left the p columns as is, including both common and rare words. The CCA procedure below implements a smooth cutoff, permits the inclusion of out-of-sample word embeddings, and provides a framework for assigning both "gist" scores and "pivot" scores.

A.1.1 Overview of canonical correlation analysis

Before going through the details of pivot scaling, we first describe canonical correlation analysis. CCA finds the largest correlations between two sets of data and is usually estimated using singular value decomposition (SVD). [26] The dimensions with the largest correlations map two sets of data to a latent space that is a good representation of both data sets. In its estimation, the SVD optimizes Pearson correlations, or cosine similarity between centered matrices:

$$\max_{\phi_x, \phi_y} \frac{\phi_x^\top C_{xy} \phi_y}{\sqrt{\phi_x^\top C_{xx} \phi_x} \sqrt{\phi_y^\top C_{yy} \phi_y}} \tag{5}$$

In this formula, C_{xy} is the cross product of centered matrices X and Y, where X is one set of input variables and Y is another input, while C_{xx} is the covariance matrix for X alone and C_{yy} for Y alone. ϕ_x is an eigenvector of $C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx}$ and ϕ_y is an eigenvector of $C_{yy}^{-1} C_{yx} C_{xx}^{-1} C_{xy}$, where $^{-1}$ indicates an inverted matrix.

ϕ_x and ϕ_y are the matrices that project the X and Y matrices onto a shared latent space. These are the coefficients from the model, like βs from a linear regression. We multiply the singular vectors

[26] As an example of the use of CCA on text (and the primary inspiration for its use here), Dhillon et al. (2015) use CCA to take advantage of both the left (before) and right (after) contexts of a word in a sentence to train their embeddings to obtain two "views" of the data.

by either the left, X, or right, Y, centered input to the CCA to obtain the variables' locations in the shared space.

A.1.2 Overview of CCA in pivot scaling

In pivoted text scaling, we will scale our text based on a word co-occurrence matrix with word frequencies and document lengths removed and a truncated word co-occurrence matrix multiplied by word frequencies. Maximizing the similarity between these two views of out data will scale text based on reliable, common words. Projecting each side of the data to the shared latent space will give us the word or "gist" scores on one side (which we use to score documents) and keywords or "pivot" scores on the other.

INPUT	
M	Term-document matrix (in-sample data), typically binary in short texts (i.e. words not used more than once)
W	Word embedding matrix (optional, out-of-sample data)
b	The exponent on word frequencies – typically, set as high as possible (e.g. 2 or 4).
DERIVED DATA	
G	Word co-occurrence matrix - $(M^\top M)^{\circ\frac{1}{2}}$, where $\circ\frac{1}{2}$ is an element-wise square root
D_g	Diagonal of G matrix
X	Row standardized word co-occurrence matrix - $\|g_i\|^{-1}G$ - left input of CCA
Z	Word co-occurrence or embedding matrix with weights - $Z = D_g^b X$ or $Z = D_g^b W$ - right input of CCA. b is a power for the vector D_g
k	The truncation of the principal components of X – controls sparsification and the number of pivots in the accompanying R package, k is calculated from the standardized matrix X using the number of words with $\|x_j\| > 1$
P_j	Column sums of $X^{\circ 2}$
P_i	Row sum of $X^{\circ 2}$
C_{XZ}	Covariance matrix for X and Z
OUTPUT	
ϕ	Singular vector - ϕ_x is a left singular vector and ϕ_z is a right singular vector – used to create ϕ^{proj}
ϕ^{proj}	Projection - ϕ_x^{proj} projection from X to shared space with Z, ϕ_z^{proj} projection from Z
ϕ_x^{proj}	Word/"gist" scores - coefficients for all words
ϕ_z^{proj}	Keyword/"pivot" scores - coefficients for keywords
$M\phi_x^{proj}$	Document scores (divide by number of words in document)

Table 3: This is a reference table for the notation used below.

1. Account for differences in word frequency and document length – Standardize word co-occurrences G with Euclidean norm $\|g_i\|$ and element-wise square root	$X = \|g_i\|^{-1}G; \quad G = (M^\top M)^{\circ\frac{1}{2}}$
2. Create frequency-based regularizer, estimate number of pivot words – Estimate k principal components of Z for W or X multiplied by word counts	$Z_k = D_g^b X$ or $Z_k = D_g^b W$
For large b, calculate c, and associated k, such that rare word weights in regularizer Z's principal components approach zero:	if $\frac{P_j}{P_i} < 1$ then $\frac{1}{e^{-\lambda}+1} \to 0$ where $\lambda = 2b\left(\ln\left(\frac{P_j}{P_i}\right) - c\right)$
3. Use whitened Z_k (in CCA) to adjust decomposition of X:	$\max_{\phi_x,\phi_z} \frac{\phi_x^\top C_{xz}\phi_z}{\sqrt{\phi_x^\top I\phi_x}\sqrt{\phi_z^\top C_{zz}\phi_z}}$.
4. Apply projections to term-document matrix M: and divide by number of words in document	$M\phi_x^{proj}$

Table 4: *Summary of pivoted text scaling.* Notation for this table is introduced in Table 3. Projections are estimated using singular value decomposition. Larger bs induce the desired "pivot" behavior (i.e. upweight common words). We standardize the final document scores based on the number of words in a document.

A.1.3 Left input

In our CCA, one side of the input will be our in-sample, standardized word co-occurrence, X, described previously.

A few adjustments to the ordinary CCA and its input data create the common word regularization described in the "Approach" section. On the left side, we do not normalize the X matrix in our CCA. This is the version of CCA sometimes called orthonormalized partial least squares (Rosipal and Krämer, 2006). It will only adjust the decomposition of X. [27]

[27] Note that for computational reasons this "PCA" is singular value decomposition on X rather than a centered X for the "compress_fast=TRUE" option in the accompanying R package.

$$\max_{\phi_x,\phi_z} \frac{\phi_x^\top C_{xz} \phi_z}{\sqrt{\phi_x^\top (I)\, \phi_x} \sqrt{\phi_z^\top C_{zz} \phi_z}} \tag{6}$$

A.1.4 Right input

Next, on the right, we need a Z matrix that captures the reliability information in the inverse rank frequency distributed word counts, and that we can use to separate keywords from rare, noisy words.

We create this Z with a key ingredient in mind: the inverse-rank frequency of words in the original word co-occurrence matrix, and our X matrix multiplied by word counts, produces principal components that are also inverse-rank frequency distributed. Inverse-rank on the k dimension is driven by rank distance from the kth most common word in the corpus. Noise around that distribution on each dimension increases for smaller eigenvalues, and reflects rank uncertainty.

Given this, we create Z by raising b in:

$$Z = D_g^b X, \tag{7}$$

where D_g is the diagonal of G, or essentially the number of times a word was used in the corpus M. For large b, word frequencies are roughly proportional to the eigenvalues of this decomposition raised to $\frac{1}{b}$, and the number of words loading on a dimension/eigenvectors increases for smaller, noisier eigenvalues. The whitening process in CCA – decomposing and then standardizing eigenvalues to one – makes the eigenvalues' size irrelevant, except for their ordering. This removes the influence of word frequency above the cutoff.

This "distillation" process gives us a smooth, word rank cutoff for the right side of our CCA, to the extent that words' ranks can be separated given the covariance of similarly frequent words and accompanying words. When we truncate this Z matrix, we also smoothly truncate by word rank.

In the accompanying R package, the regularizer Z's dimensionality k is set by default to be the

number of pivot words divided by two. This k approximation targets a total pivot word weight of $2k$ (i.e. the number of words appearing more often than their accompanying words).

A.1.5 Scoring documents

Our last step after the canonical correlation analysis that estimates the word/"gist" scores is to return document scores based on our word location estimates. To do this, we simply multiply the projection, ϕ_x^{proj}, (i.e. the coefficients) by the original term document matrix M, then divide these document scores by the total number of words used in a document. We apply an exponent of 0.75 on the total number of words in a document so that the text dimensions are not correlated with document length.

A.1.6 Pivot cutoff

The above steps are sufficient to implement the CCA in pivot scaling. It is potentially helpful to describe the function produced by truncating the the Z matrix.

Activation over all dimensions (in text data) is approximately a logistic function:[28]

$$\frac{1}{e^{-\lambda}+1} \propto \|\phi_z^{proj}\| \tag{8}$$

where λ equals $2b\left(\ln\left(\frac{P_j}{P_i}\right)-c\right)$, with P_j and P_i the column and row sums, respectively, of the standardized word co-occurrence matrix $X^{\circ 2}$.

We show convergence to that functional form around the scalar c in the appendix Figure A2. The location of c is controlled by setting k and the form of the logistic function is controlled by b. Setting k equal to one half the number of words with column Euclidean norm greater than one in the standardized word co-occurrence matrix X gives words that appear less than their accompanying words little to no weight as pivot words.

[28] Having pivot scores equal to 0 for rare words is more important than this precise functional form.

A.1.7 Keywords

Once we induce pivot behavior with large b, we can multiply ϕ_z^{proj} by the corresponding canonical correlation to place the pivot scores on the same scale as the overall word scores. ϕ_x^{proj} and ϕ_z^{proj} will then be similar to equivalent for the pivot words, while relatively rare words in ϕ_z^{proj} will remain close to zero.

A.1.8 Pivot Scaling with Out-of-Sample Word Embeddings

The above sections describe the implementation of pivoted text scaling using only in-sample data. It is intended to make estimation of policy dimensions more reliable, and creates dimensions that estimate the gist of language. In very small sets (less than one thousand observations, for example), even the common words that tend to appear more than their accompanying words will be noisy. In these data sets, it can be helpful to bring in out-of-sample data to make the estimation somewhat more reliable.

In this, the assumption is that if two words are also closely related in general English, then their covariance is less likely to be noise and so we can place greater weight on their covariance in our own data.

To include out-of-sample word embeddings in our scaling, we simply 1) replace $Z = D_g^b X$ with $Z = D_g^b W$, where W is the word embedding matrix, and 2) run CCA with $b = 1$ (i.e. the sample word frequencies without alteration).[29]

With this new Z, the right side of our CCA captures variance in words that appear very frequently in our data, as well as words that both appear moderately frequently in our data *and* appear in similar contexts in general English.

[29] Publicly available pre-trained word embeddings are already truncated – they typically contain 200 to 300 dimensions. For k less than the maximum dimensionality of the word embeddings, we truncate to the top k principal components of this new Z.

A.1.9 Formatting and use of out-of-sample data in validations

Prior to calculating the word co-occurrence matrix for our applications, we process the text using the defaults in the "stm" R package (Roberts, Stewart and Tingley, 2016), the most commonly used software for text analysis in political science. We do not stem the text because word embedding data is typically not stemmed.

A.2 Affordable Care Act Results

	Oppose ACA (text)	Favor ACA (text)	
Oppose ACA (closed-ended)	0.80	0.20	1.00
Favor ACA (closed-ended)	0.18	0.82	1.00

Table A1: *Confusion matrix.* The first column of this table states that dimension 1 minus dimension 2 of the text is less than average (score approximately 0) for 80% of respondents who like the ACA. For the second column, dimension 1 minus dimension 2 is greater than average for 82% of the respondents who dislike the ACA. Overall, 81% of respondents can be correctly classified in only two dimensions without training a supervised model.

Left - Patient Protection - Dimension 1
pre existing conditions changes
it expands coverage doesn't allow for being dropped for pre- existing conditions.
i like how people can be covered up to 26 yrs old, and the no pre-existing conditions
coverage for pre existing conditions
because people can no longer be denied insurance because of pre-existing conditions
Right - Role of Government - Dimension 1
i dont want the government to be controlling it
i don't like the government telling me what i have to do. i don't think the government should be stepping into peoples business
too socialist, too much red tape
don't like the government telling me what to do and the way it was forced on us
its unconstitutional it infringes on my freedom of religion. government. doesn't have the right to force me to buy anything from any industry
Left - Universal Access - Dimension 2
because of america wants to be one of the great nations of the world it needs to provide health care for every american
because everyone gets health care
it is time we followed the rest of the world and provided coverage for everyone in the u.s.
because i think everyone needs health care
i believe health care is a basic right to all, and i believe this has gone farther than any administration to provide it
Right - Personal Cost - Dimension 2
she has friends that talk about and there not on medicare or medicaid and they lost there insurance and when they try to get some it's very expensive.
they're going to force people to take it. those on the borderline won't be able to pay for it.
my insurance is going to go away and be replaced by crappy insurace
it benefits me
because of the medicare and medicaid regulations

Table A2: *Examples of open-ended responses.* This table shows a random sample of transcribed responses about the ACA that score relatively highly (greater than 2 standard deviations) on one of the top 2 dimensions and low (lower than 1 standard deviation) on the other. The method scores these responses unambiguously. A total of 1,201 responses fit these criteria in the 2009 - 2015 training set.

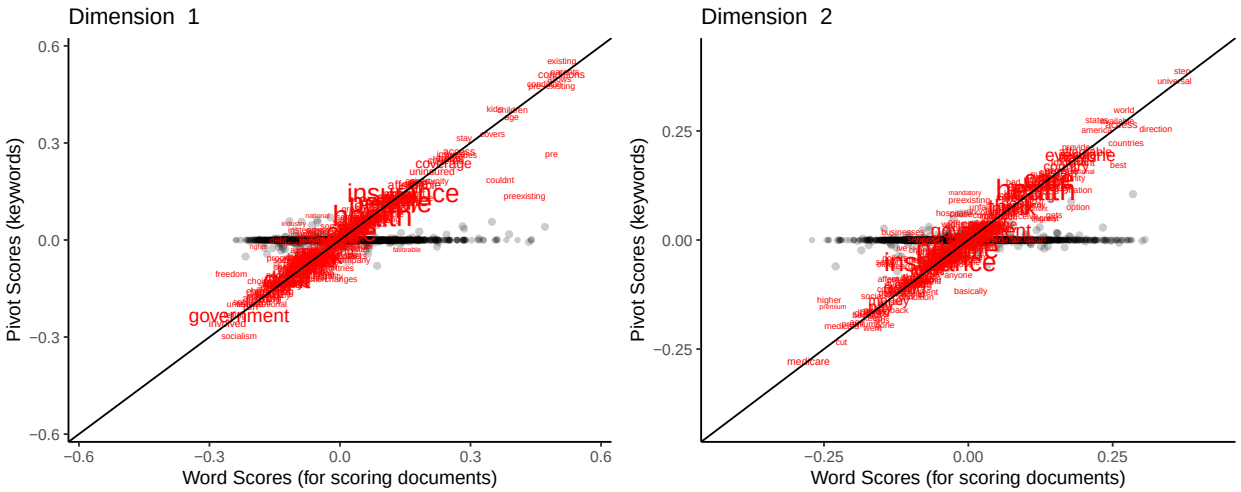


Figure A1: *Relationship between word scores and pivot scores.* Like Figure 1, this figure shows the linear correspondence between pivot scores (used for keywords) and overall word scores (used to score documents). The x axis for both panels is the word score and the y axis is the pivot score. The line resembling a line from a linear regression is a 45 degree (x=y) line. 'Pivot words' have close to the same coefficients in both views of the output.

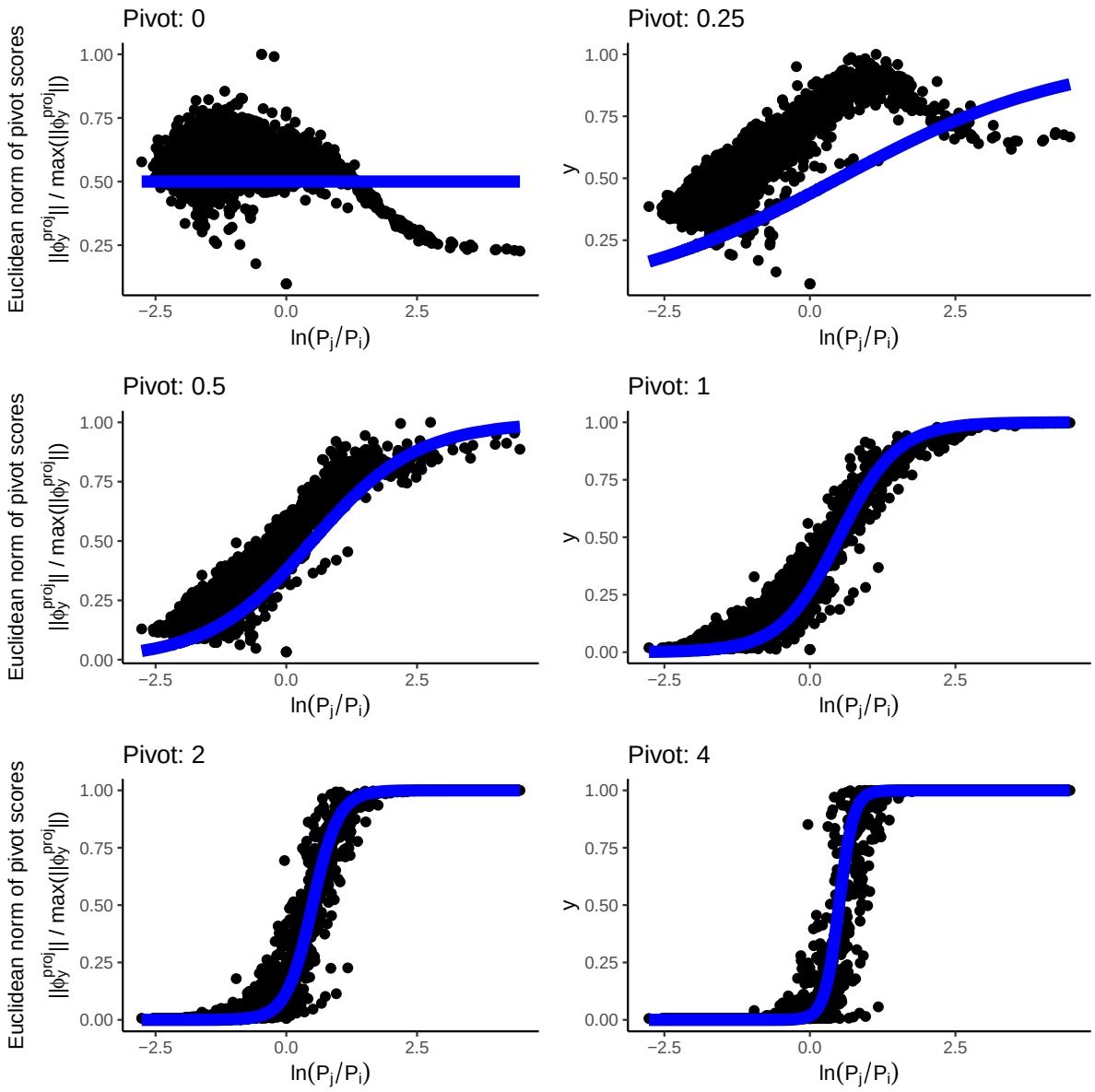


Figure A2: *Tuning* ***b*** *to induce pivots.* We use the b value in the bottom right panel, and set k so that the total weight of the words on the y-axis here is approximately equal to the number of words appearing more often than their accompanying words. In the ACA data, this helps limit the influence of Spanish language outliers.

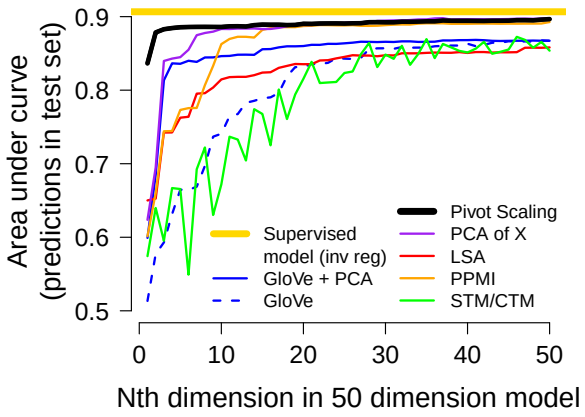


Figure A3: *Dimensionality of other methods, 50 dimensions.* This figure shows area under the ROC curve for 50 penalized regressions predicting ACA favorability from each method's output. It shows that pivot scaling achieves high AUC in low dimensions, while other methods converge to high accuracy more slowly (note that this panel shows multiple AUC's, rather than an ROC curve). The dimensionality for topic models here is the number of topics minus 1.

Keywords					
Dimension 1		*Dimension 2*		*Dimension 3*	
"role of government"	"patient protection"	"personal cost"	"universal access"	"individual mandate/ regulation"	"middle class/ poor/elderly"
Anti	Pro	Anti	Pro	Anti	Pro
involved	conditions	medicare	step	unconstitutional	class
socialism	existing	cut	universal	stay	income
government	parents	premiums	access	federal	middle
telling	pre-existing	went	available	parents	low
run	condition	age	world	private	poor
unconstitutional	allows	keep	direction	involved	help
federal	children	medicaid	america	telling	elderly
anything	age	benefits	states	government	helps
much	kids	doctors	provides	choice	lower
medicine	stay	wont	affordable	age	hard
socialized	covers	paying	everyone	pre-existing	working
want	access	away	needed	takes	helping
economy	insured	kids	needs	control	gone
business	provides	theyre	right	business	lot
everything	helps	pay	important	condition	new

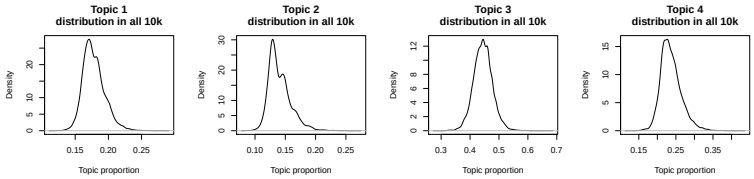
Table A3: *ACA Keywords.* This table shows the keywords in the top two dimensions of the open-ended responses on favoring or opposing the Affordable Care Act. We identify keywords by multiplying ϕ_z^{proj} by its Euclidean norm. The dimensions and keyword identification are unsupervised. We added the pro and anti labels after linking the scores to the preceding closed-ended survey response. Dimensions are approximately orthogonal, but words can appear in multiple dimensions, especially when they have meanings/connotations that are affected by accompanying words.

A.2.1 Affordable Care Act - other variables

	Dependent variable:				
	Repeal 2016 vs Repeal 2012	Repeal 2018 vs Repeal 2016	Donald Trump vs Hillary Clinton	Donald Trump vs Republican establishment	Hillary Clinton vs Bernie Sanders
	(1)	(2)	(3)	(4)	(5)
Dimension 1 (+ patient protection vs − role of government)	−0.09	0.10	−0.06	0.01	0.12
	(0.07)	(0.08)	(0.02)	(0.04)	(0.06)
Dimension 2 (+ universal access vs − personal cost)	−0.27	0.15	−0.02	−0.08	−0.01
	(0.06)	(0.07)	(0.02)	(0.04)	(0.05)
Dimension 3 (+ middle class vs − regulation)	0.01	−0.01	−0.05	0.004	0.17
	(0.07)	(0.07)	(0.02)	(0.04)	(0.06)
Dimension 0 (+ general vs − specific)	−0.02	−0.20	0.01	0.04	−0.02
	(0.06)	(0.07)	(0.02)	(0.04)	(0.05)
Favor ACA Repeal 2012			0.06	0.15	−0.02
			(0.01)	(0.03)	(0.04)
Observations	865	575	932	404	402

Table A4: *Between and within party correlates in small panel survey.* All models here control for partisan identification – 2012 PID in model (1) and 2016 PID in models 2 through 5. Independent variables are scaled so that a unit change correspondents to a one standard deviation change in a given subset of the data. The dependent variables in the vote choice models are coded 1 if the respondent voted for the first candidate (Donald Trump, Hillary Clinton) and coded -1 if the respondent voted for the second candidate(s). Republican "establishment" candidates are: Jeb Bush, John Kasich, Marco Rubio, and Chris Christie.

A.2.2 Affordable Care Act - topic models



Keywords (highest probability) - TOPIC MODEL			
Topic 1	Topic 2	Topic 3	Topic 4
Pro	Pro	Anti	No relationship
health	people	insurance	dont
going	will	care	get
everyone	need	think	like
cant	law	government	afford
much	feel	cost	pay
money	reform	lot	help
just	doesnt	want	can
good	better	us	coverage
able	medicare	work	everybody
something	buy	now	believe
go	away	way	think
know	control	healthcare	p
well	bill	many	right
companies	enough	im	needs
americans	costs	make	country

Table A5: *Topic model summary (4 topics).*

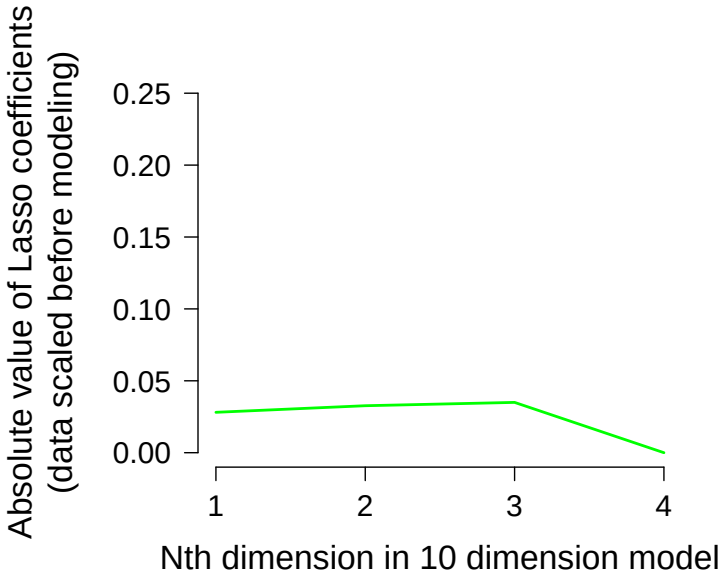


Figure A4: *Topic model summary (4 topics).* A coefficient of 0 for one topic is expected behavior.

Keywords (FREX) - TOPIC MODEL			
Topic 1	Topic 2	Topic 3	Topic 4
Pro	Pro	Anti	No relationship
going	law	care	get
money	better	government	like
know	buy	cost	can
getting	costs	way	country
everyone	people	healthcare	medical
makes	public	make	without
much	pre	business	expensive
already	every	anything	conditions
health	year	plan	poor
cover	given	obama	wont
choice	abortion	premiums	gives
price	new	high	believe
gone	said	now	whole
something	enough	work	system
able	will	lot	far

Table A6: We use the default FREX weight in the "stm" package when calculating the keywords in this table.

A.2.3 Affordable Care Act - word embeddings

The contribution of word embeddings is difficult to evaluate objectively, and a complete evaluation would require a substantial extension to this paper. Because of this, they will mostly be addressed in a separate paper that can dedicate space to guidance on their use. Here, we primarily point out that a regularization based on out-of-sample word meanings is a commonsense extension to pivot scaling. Given a basis of common words, we can add in additional word based on matching meanings across contexts, and de-emphasize other words without such correspondence – with relatively little impact on the overall orientation of the scaling, and keeping the output close to our own data.

That said, we do provide below an example of pivoted text scaling with word embeddings. With the word embeddings, dense word representations from a massive, out-of-sample training set provide more smooth connections among word meanings.

In this example, we rerun the ACA over time analysis in the main paper training only on data from the 2018 panel. This uses around 700 responses rather than 10,000. We replace the X matrix on the right side of the CCA with the word embedding matrix W, using pretrained word embeddings from (Yin and Schütze, 2016). The word embeddings allow us to use a small b in pivoting. With smaller b moderately common words are less likely to be removed from the pivots, especially if their associations in our data correspond to meanings in general data. We show the dimensions over time in Figure A6 for $b = 1$ with embeddings and for the usual $b = 4$ without embeddings in Figure A7. Keywords for $b = 1$ with embeddings are shown in Table A7.

The dimensions in Figure A6 resemble those from training on the full 2009 through 2015 data, despite the dramatically smaller sample size and the different context. For convenience, the results from the main paper are shown again here in Figure A5.

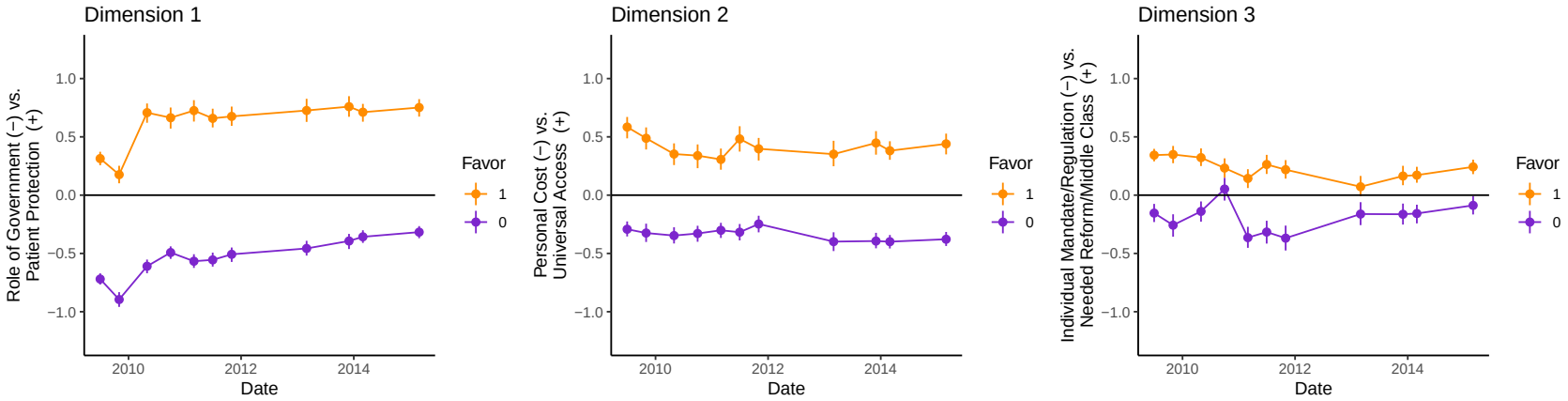


Figure A5: *Changes in ACA justifications over time, trained without word embeddings on 2009 through 2015.* This figure shows the 2009 through 2015 means of the 1st, 2nd, and 3rd dimensions of our text scaling separated by respondents who felt favorably or unfavorably about the ACA. This is the same as Figure 3 in the main text – it is included again as a convenient comparison to the word embedding results.

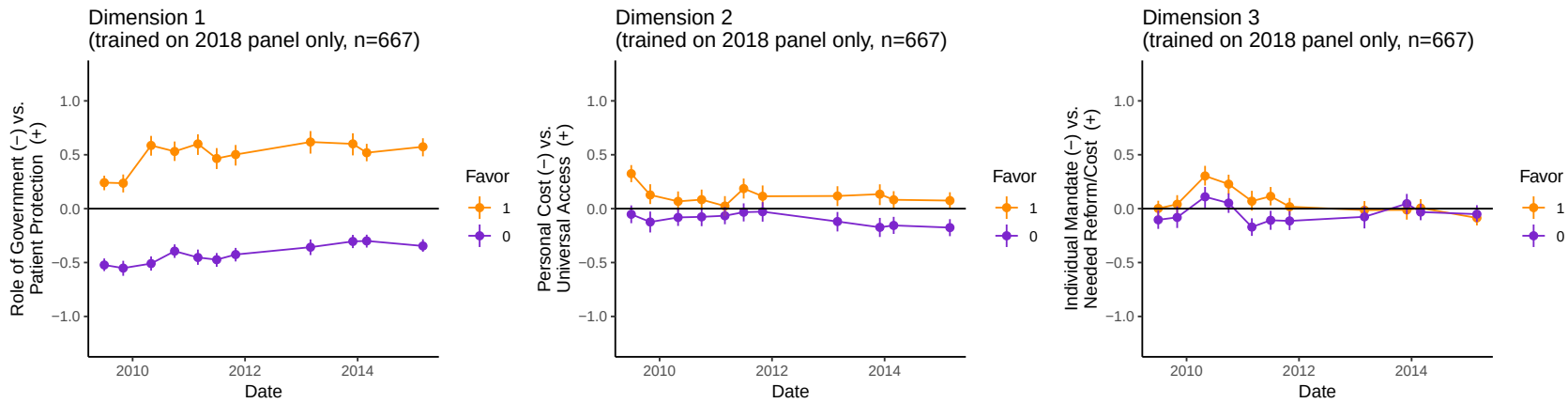


Figure A6: *Changes in ACA justifications over time, trained* with *word embeddings on 2018 only.* Word scores to produce this timeline were trained on 2018 data (n=667) and "meta" word embeddings (Yin and Schütze, 2016) and then applied to Pew and KFF data from 2009 through 2015. *b* was set to 1 to allow the influence of more words given the word embedding 'prior'. Keywords are shown in Table A7. Note that bootstrapped means and confidence intervals are at the document level for a given text dimension.

Keywords					
Dimension 1		*Dimension 2*		*Dimension 3*	
"cost/role of government"	"patient protection"	"personal cost"	"universal access"		
Anti	Pro	Anti	Pro	Anti	Pro
pay	conditions	high	access	universal	reform
government	pre-existing	pay	reform	government	lied
go	covers	costs	country	affordable	system
high	preexisting	premiums	think	control	keep
expensive	access	lower	right	much	many
much	covered	many	everyone	mandatory	helped
enough	pre	lot	need	needs	doctor
just	helps	insurance	care	increased	bill
right	children	much	needs	individual	plan
doctor	coverage	increased	expensive	provides	opinion
know	provides	doctors	one	access	know
control	allows	lied	affordable	allowed	still
dont	age	doctor	health	costs	premiums
think	existing	allowed	government	companies	think
money	insured	fine	system	right	will

Table A7: *ACA Keywords, trained on 2018 data and "meta" word embeddings (Yin and Schütze, 2016).*

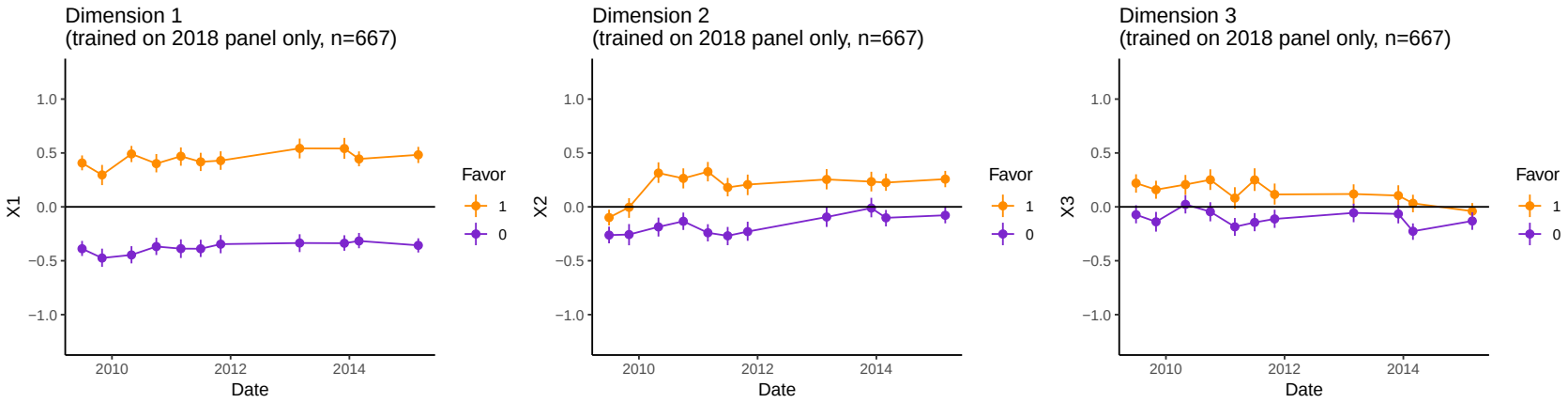


Figure A7: *Changes in ACA justifications over time.* Word scores to produce this timeline were trained on 2018 data (n=667) *without* word embeddings and then applied to Pew and KFF data from 2009 through 2015. *b* was set to 4.

Keywords					
Dimension 1		*Dimension 2*		*Dimension 3*	
"personal cost"	"patient protection"				
Anti	Pro	Anti	Pro	Anti	Pro
much	conditions	government	premiums	poor	reform
high	pre-existing	right	conditions	much	doctor
premiums	pre	just	keep	allows	keep
forced	needs	force	lot	covered	obama
pay	preexisting	enough	covers	way	us
working	access	dont	high	one	system
away	helped	think	pre	control	getting
good	americans	everyone	preexisting	fine	country
doctors	covers	costly	deductibles	aca	law
go	covered	expensive	existing	existing	free
fine	provide	cant	time	dont	cost
one	able	make	many	get	many
give	coverage	one	pre-existing	away	americans
us	provides	affordable	allows	buy	paying
buy	first	buy	covered	helps	lot

Table A8: *ACA Keywords, trained on 2018 data* without *word embeddings.* Distinctions about the role of government were no longer major talking points by 2018.

A.3 Partisan Animus Results

Keywords			
Dimension 1		*Dimension 2*	
"policy/ideology"	"politics"	"Democrats"	"Republicans"
gun	president	give	theyre
rights	anything	democratic	class
pro	middle	much	donald
stance	class	away	trump
womens	trump	socialism	will
control	need	free	need
views	donald	want	gay
health	pay	left	womens
conservative	get	work	healthcare
issues	poor	programs	racism
abortion	one	take	conservative
gay	seem	things	poor
marriage	theyre	believe	middle
religious	good	everything	immigration
policy	will	trying	president

Table A9: *Keywords for top 2 dimensions of ANES partisan animus analysis.* Preprocessing: words used more than twice.

Keywords			
Dimension 1		*Dimension 2*	
"policy/ideology"	"politics"	"Democrats"	"Republicans"
rights	anything	socialism	donald
stance	middle	give	president
gay	democrats	thing	trump
views	will	believe	taxes
womens	class	democratic	middle
gun	donald	can	everyone
pro	president	political	poor
health	one	giving	theyre
control	theyre	left	laws
life	everyone	things	rich
immigration	pay	much	class
abortion	getting	away	seem
religion	make	vote	richer
religious	get	take	done
issues	wealth	put	gay

Table A10: *Keywords for top 2 dimensions of ANES partisan animus analysis.* Preprocessing: words used more than once.

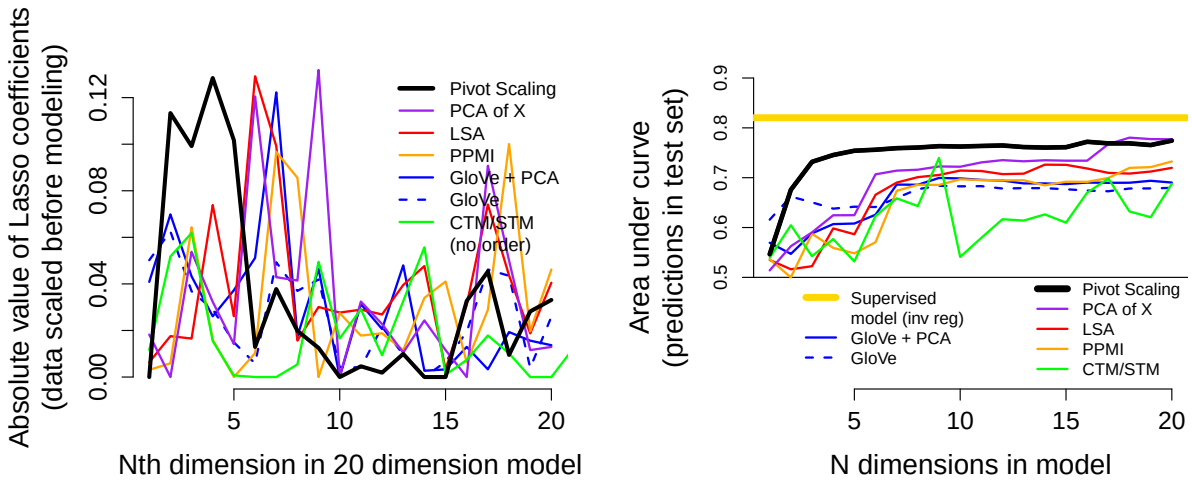


Figure A8: *Partisan animus.* The left panel of this figure shows the absolute value of coefficients from a 20 dimensional penalized regression predicting party identification in the ANES, while the right shows the AUC for performance in the 10% test set by the number of dimensions included in the Lassos (or multinomial inverse regression for the supervised comparison). Predictive dimensions are concentrated in the first dimensions of pivot scaling. The first dimension separates respondents talking about politics vs. policy, while the second separates partisans. We include the keywords from these top 2 dimensions in appendix Table A9. This is the same as Figure 4, but requires that all words be used more then twice in the data. The higher threshold improves the performance of both unsupervised and supervised methods, with the exception of the unordered GloVe. Preprocessing: words used more than once.

A.4 Russian IRA Activity on Twitter during U.S. Election Campaign

Keywords			
Dimension 1		*Dimension 2*	
islamkills	fatally	blacklivesmatter	iowa
stopislam	fatal	oscarhasnocolor	hampshire
brussels	officer	blackhistorymonth	rubio
refugees	unarmed	policebrutality	poll
trumpforpresident	charged	acab	cruz
votetrump	teen	unarmed	gopdebate
makeamericagreatagain	shot	blackskinisnotacrime	politics
trumptrain	shooting	btp	caucuses
hillaryforprison	suspect	cops	gop
ccot	local	americanhistoryisblack	ccot
trumppence	-duty	antipolicebrutalityday	clinton
neverhillary	-year-old	wearhoodiefortrayvon	fundraising
gopdebatesc	cop	magicbutreal	endorses
syrian	hit–run	amerikkka	abcnews
maga	sues	blacktolive	stopthegop

Table A11: *Keywords for top 2 dimensions of Russian intelligence linked Twitter account analysis.*

Keywords			
Right troll v Left troll		*Hashtag gamer v News feed*	
poll	fatally	islamkills	local
cruz	unarmed	stopislam	fatally
rubio	officer	oscarhasnocolor	fatal
gop	fatal	brussels	hit–run
trumptrain	shot	ireallylikeyoubut	county
hampshire	cop	highschooltaughtme	charged
ccot	teen	thetroublewithaddiction	baltimore
clinton	charged	refugees	maryland
makeamericagreatagain	acab	sometimes	alleged
donald	shooting	icantbeurfriendbecause	officer
trump	custody	hashtag	iowa
iowa	police	andthenishouldhavesaid	shooting
gopdebate	-duty	sleep	news
republican	suspect	shit	charges
caucuses	btp	ihatepokemongobecause	sues

Table A12: *Keywords for top 2 dimensions of Russian intelligence linked Twitter account analysis.* This table shows the keywords for dimension 1 minus dimension 2 (left vs right) and dimension 1 plus dimension 2 (hashtag gamer vs news feed).