Identification, Interpretability, and Bayesian Word Embeddings

Adam M. Lauretig
Department of Political Science
Ohio State University
Columbus, OH
`lauretig.1@osu.edu`

Abstract

Social scientists have recently turned to analyzing text using tools from natural language processing like word embeddings to measure concepts like ideology, bias, and affinity. However, word embeddings are difficult to use in the regression framework familiar to social scientists: embeddings are are neither identified, nor directly interpretable. I offer two advances on standard embedding models to remedy these problems. First, I develop Bayesian Word Embeddings with Automatic Relevance Determination priors, relaxing the assumption that all embedding dimensions have equal weight. Second, I apply work identifying latent variable models to anchor the dimensions of the resulting embeddings, identifying them, and making them interpretable and usable in a regression. I then apply this model and anchoring approach to two cases, the shift in internationalist rhetoric in the American presidents' inaugural addresses, and the relationship between bellicosity in American foreign policy decision-makers' deliberations. I find that inaugural addresses became less internationalist after 1945, which goes against the conventional wisdom, and that an increase in bellicosity is associated with an increase in hostile actions by the United States, showing that elite deliberations are not cheap talk, and helping confirm the validity of the model.

1 Introduction

Important questions in the social sciences turn on the meanings of words used to express ideas like language change, emotion, and ideological affinity (Hamilton et al., 2016; Rheault et al., 2016; Pomeroy et al., 2018). One increasingly popular way to represent meaning, originating in natural language processing, is through the use of word embeddings. This class of models learns a set of coefficients which encode meaning by predicting a word given the surrounding words (Mikolov et al., 2013a,b). These coefficients are the *embeddings*, which can then be used to analyze word meanings.

Unfortunately, existing embedding models are not always appropriate for answering social scientists' questions. Embeddings are not identified, and the dimensions are not directly interpretable, which makes it difficult to perform statistical inference on the embeddings produced by standard models, for example, using them as covariates in a regression model.[1]

To resolve these issues, I cast word embeddings as a Bayesian latent variable model. Identifying multidimensional latent variable models is a known problem, and I draw on solutions proposed in the ideal point modeling literature (Rivers, 2003; Clinton et al., 2004) to render embeddings interpretable and usable in a regression framework. I demonstrate these results on two corpora: a collection of inaugural addresses, and a selection of declassified diplomatic documents from the *Foreign Relation of the United States*. In the inaugural addresses, I find rhetoric became more domestically-focused after 1945, a shift which existing social science approaches cannot detect. This finding stands in contrast to what existing theories of international relations would have us expect. In the *FRUS* documents, I find that more bellicose rhetoric results in more aggressive American foreign policy behavior, helping confirm that elite deliberation matters for shaping foreign policy, and that the measurements I create correlate with existing datasets, helping to establish the validity of the model results.

---

[1]This is because in a regression setup, the coefficient is the change in the dependent variable for a 1-unit increase in the independent variable. With embedding dimensions, it is not clear what a 1-unit increase in the independent variable means, nor does direction have any clear meaning.

2 Social Science and Embedding Models of Language

Traditional approaches to creating variables from text in the social sciences involve human coders, who assign documents to categories based on pre-defined criteria. However, this approach is expensive, and does not scale. Text as data techniques attempt to solve this problem through the use of natural language processing techniques to convert a corpus of text into numeric objects which makes inference possible (Grimmer and Stewart, 2013; Gentzkow et al., 2017). These techniques allow scholars to create variables and operationalize concepts in corpora that are too large for human coding, and investigate ideas which cannot be measured directly (unlike indicators like Gross Domestic Product or population).

While a variety of models have been proposed to create variables from political text, including scaling models (Lowe et al., 2011), and topic models (Blei et al., 2003; Grimmer, 2010; Roberts et al., 2016), these approaches focus on the document as the unit of analysis. Word embeddings, which have a long history in the natural language processing literature (see Turney and Pantel (2010) for an extensive review of pre-neural network models), have recently been embraced by social scientists for their potential for inference at the word level. Modern neural word embedding models learn a low-dimensional representation of a word as a dense vector by either factorizing a word co-occurrence matrix or predicting the co-occurrence of a pair of words using a single-layer neural network. Among the best known of these models is word2vec (Mikolov et al., 2013a,b), which proposed an efficient model for learning embeddings, framing embedding learning as a prediction task, rather than a factorization task.

For social scientists, word embeddings are a powerful tool because they can represent the meanings of individual words. Embeddings can help isolate patterns in corpora that are expensive to label, and make apparent latent phenomena not observable through simple document-feature counts such as patterns of semantic change, (Hamilton et al., 2016), cultural assumptions and biases (Caliskan et al., 2017; Kozlowski et al., 2018; Garg et al., 2018), and ideological affinity in international organizations (Pomeroy et al., 2018).

However, these embeddings can be problematic for social science research, where scholars care about both model identification and interpretable results. Embeddings are multidimensional latent variable models, which are not, by default, identified: a known problem with this class of model, where multiple permutations of latent dimensions can result in the same observed data (Rivers, 2003; Clinton et al., 2004; Aldrich et al., 2014). However, by anchoring points on these dimensions, it is possible to present identified and interpretable dimensions. In the ideal-point literature, these anchors represent ideological "endpoints," with theory guiding the selection of which legislators are most liberal and conservative. Choosing words as anchors with a large number of dimensions is more difficult than choosing legislators, however, I offer a solution below.

There have been multiple efforts at developing Bayesian word embeddings (Rudolph et al., 2016; Barkan, 2017; Ji et al., 2017; Havrylov and Titov, 2018), however, none of these have exploited the key advantage of Bayesian inference: the ability to quantify the uncertainty in parameter estimates, and use prior information to inform parameter estimates. The one approach that has incorporated both uncertainty and hypothesis testing is Han et al. (2018), who offer both measures of uncertainty, and a way to test the effect of metadata on the similarity of embeddings, however, this approach does not account for identification problems in the learned embeddings.

3 Bayesian Word Embeddings

In this section, I develop word embeddings as Bayesian latent variable models estimated with variational inference, following similar work for probabilistic principal components analysis (Bishop, 1999) and ideal-point models (Imai et al., 2016). I first discuss the embedding model setup, add Automatic Relevance Determination priors to the model, and then, present the variational updates to estimate the model.

Word embeddings predict the probability of a word-context pair co-occurring, and because the co-occurrence is a binary variable ($Y_{ij} = 1$ if w_i and w_j co-occur, 0 otherwise), I use a probit link function to model the probability of co-occurrence.

$$p(Y_{ij} = 1) =$$
$$(\mathbf{1}[z_{ij} > 0]\mathbf{1}[y_{ij} = 1] + \mathbf{1}[z_{ij} < 0]\mathbf{1}[y_{ij} = 0]) \quad (1)$$
$$\mathcal{TN}(z_{ij}|\boldsymbol{x}_i^\top \boldsymbol{\beta}_j, 1).$$

$\boldsymbol{X}$ and $\boldsymbol{\beta}$ are $K \times I$ (or $K \times J$, respectively) -dimensional matrices, $\boldsymbol{Y}$ is an $I \times J$ co-occurrence matrix, the corpus contains I words and J context words. Each embedding vector ($\boldsymbol{x}_i$ or $\boldsymbol{\beta}_j$) has a K-dimensional multivariate normal prior.

Most existing approaches to word embeddings contain no measures of uncertainty, or the covariance between dimensions. This can be a problem during estimation, as the model attempts to put equal weight on all dimensions. To resolve this, I use Automatic Relevance Determination (ARD) priors, which place a separate gamma-distributed scalar (e.g. α_{X_k}) on the diagonal for each dimension of the covariance matrix (MacKay and Neal, 1994; Bishop, 1999). These priors penalize unnecessary model dimensions, improving model fit.

This specification results in the following likelihood:

$$
\begin{aligned}
p(\boldsymbol{Z}, & \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\alpha_X}, \boldsymbol{\alpha}_\beta | \boldsymbol{Y}) \propto \\
& (\mathbf{1}[z_{ij} > 0]\mathbf{1}[y_{ij} = 1] + \mathbf{1}[z_{ij} < 0]\mathbf{1}[y_{ij} = 0]) \\
& \mathcal{TN}(z_{ij} | \boldsymbol{x}_i^\top \boldsymbol{\beta}_j, 1) \times \\
& \prod_i \mathcal{MVN}(\boldsymbol{x}_i | 0, \boldsymbol{\alpha}_X^{-1}) \times \\
& \prod_i \mathcal{MVN}(\boldsymbol{\beta}_j | 0, \boldsymbol{\alpha}_\beta^{-1}) \times \\
& \prod_k \text{Gam}(\alpha_{X_k} | c_{X_0}, d_{X_0}) \times \\
& \prod_k \text{Gam}(\alpha_{\beta_k} | c_{\beta_0}, d_{\beta_0}).
\end{aligned}
\tag{2}
$$

For Bayesian models like this, the goal is to estimate posterior distributions of the parameters most likely to have produced the observed data. Given the joint density (probability of data and parameters), we want to calculate the conditional density of the parameters by evaluating the following integral (notation follows Bishop (1999)):

$$
P(\boldsymbol{Y}) = \int p(\boldsymbol{Y}, \boldsymbol{\theta}) d\boldsymbol{\theta}
\tag{3}
$$

where $\boldsymbol{\theta} = \{\boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_\beta\}$. This integral is analytically intractable, so we transform the integral using Jensen's inequality:

$$
\begin{aligned}
\ln P(\boldsymbol{Y}) &= \ln \int p(\boldsymbol{Y}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \ln \int \mathcal{Q}(\boldsymbol{\theta}) \frac{P(\boldsymbol{Y}, \boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&\geq \int \mathcal{Q}(\boldsymbol{\theta}) \ln \frac{P(\boldsymbol{Y}, \boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \mathcal{L}(Q)
\end{aligned}
\tag{4}
$$

where $\mathcal{L}(Q)$ is evidence lower bound (ELBO).

The difference between the true model $P(Y)$ and variational approximation $\mathcal{L}(Q)$ can be represented is the Kullbeck-Leibler divergence:

$$
KL(Q||P) = -\int \mathcal{Q}(\boldsymbol{\theta}) \ln \frac{P(Y|\boldsymbol{\theta})}{\mathcal{Q}(\boldsymbol{\theta})} d\boldsymbol{\theta}
\tag{5}
$$

so we turn to a mean-field variational approximation to estimate the model, minimizing the Kullbeck-Leibler divergence (Wainwright and Jordan, 2008; Blei et al., 2017). This requires assuming that the approximation to the posterior can be factorized:

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{Z}, &\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\alpha}_X, \boldsymbol{\alpha}_\beta) = \\
& \mathcal{Q}(\boldsymbol{Z}), \mathcal{Q}(\boldsymbol{X}), \mathcal{Q}(\boldsymbol{\beta}), \mathcal{Q}(\boldsymbol{\alpha}_X), \mathcal{Q}(\boldsymbol{\alpha}_\beta)
\end{aligned}
\tag{6}
$$

and that appropriate approximating distributions can be found. In this case, the requirement is met: z_{ij} is approximated with a truncated normal, $\boldsymbol{x}_i$ and $\boldsymbol{\beta}_j$ are approximated with multivariate normals, and α_{X_k} and α_{β_k} are approximated with gamma distributions. This factorization and approximation can be further factorized into the following parameter updates:

$$
\begin{aligned}
z_{ij}^* &= \mathbb{E}[\boldsymbol{x}_i^\top] \mathbb{E}[\boldsymbol{\beta}_j] \\
\mathbb{E}[q(z_{ij})] &= \begin{cases} z_{ij}^* + \frac{\phi(z_{ij}^*)}{\Phi(z_{ij}^*)} & \text{if } y_{ij} = 1 \\ z_{ij}^* - \frac{\phi(z_{ij}^*)}{1 - \Phi(z_{ij}^*)} & \text{if } y_{ij} = 0 \end{cases} \\
\boldsymbol{A} &= \left(\text{diag}(\mathbb{E}[\boldsymbol{\alpha}_X])^{-1} + \sum_j \mathbb{E}[\boldsymbol{\beta}_j \boldsymbol{\beta}_j^\top] \right) \\
\boldsymbol{a}_i &= \sum_j \mathbb{E}[\boldsymbol{\beta}_j] \mathbb{E}[z_{ij}] \\
\mathbb{E}[q(\boldsymbol{x}_i)] &= \boldsymbol{A}^{-1} \boldsymbol{a}_i \\
\boldsymbol{B} &= \left(\text{diag}(\mathbb{E}[\boldsymbol{\alpha}_\beta])^{-1} + \sum_i \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] \right) \\
\boldsymbol{b}_j &= \sum_i \mathbb{E}[\boldsymbol{x}_i] \mathbb{E}[z_{ij}] \\
\mathbb{E}[q(\boldsymbol{\beta}_j)] &= \boldsymbol{B}^{-1} \boldsymbol{b}_j \\
c_x &= c_{x_0} + \frac{I}{2} \\
d_{x_k} &= d_{x_0} + \frac{||\mathbb{E}[\boldsymbol{x}_k]||^2}{2} \\
\mathbb{E}[q(\alpha_{X_k})] &= \frac{c_x}{d_{x_k}} \\
c_\beta &= c_{\beta_0} + \frac{J}{2} \\
d_{\beta_k} &= d_{\beta_0} + \frac{||\mathbb{E}[\boldsymbol{\beta}_k]||^2}{2} \\
\mathbb{E}[q(\alpha_{\beta_k})] &= \frac{c_\beta}{d_{\beta_k}}
\end{aligned}
\tag{7}
$$

where $c_{x_0}, d_{x_0}, c_{\beta_0}, d_{\beta_0}$ are hyperparameters set by the user. Convergence is monitored via change in the ELBO, and when change drops below a user-specified threshold, the model is considered converged. This model is implemented in the R package `bwe`.[2]

4 Identifying Model Output

The output from multidimensional latent variable models is not identified, as many possible permutations of latent values can produce the same observed data (Rivers, 2003). However, by fixing $K(K+1)$ linearly independent values (anchors), users can guarantee the embedding matrix is identified (Rivers, 2003; Clinton et al., 2004; Bafumi et al., 2005). To determine these anchors in the ideal point modeling literature, theory drives the endpoint selection: Clinton et al. (2004) fix both points for Jesse Helms, Ted Kennedy, and Lincoln Chaffee as right, left, and center anchors, respectively, in a $K = 2$ model.

While theory should always motivate modeling choices, determining theoretically motivated anchors when K ranges from 50 to 300 can be difficult. I propose a solution: theory can motivate initial anchor selection, and then, for each additional anchor, the most cosine dis-similar word is chosen as the opposite anchor. This allows the analyst to specify theoretically motivated opposites as initial anchors, and then, resulting anchors are chosen from remaining words. I provide an implementation of this algorithm in the R package `bwe`.

5 Interpreting Model Output

Anchoring the embeddings ensures they are identified, however, they are still not in a format which allows for ready interpretability in the regression-based models social scientists are most familiar with. To transform embeddings so that they can be used in regression, I opt for a modification of the anchoring approach discussed above. For this approach, the user specifies a pair of endpoints for a dimension, where the endpoints of interest are set to 1 and -1. This can be applied to as many dimensions as necessary, and then the automatic, cosine-based anchoring is used for the rest of the dimensions. An affine transformation is then used to transform the embedding matrix relative to the chosen anchors.

A key advantage of this approach is that because two anchors are supplied, words are scaled on this dimension. For example, while simply choosing "war" as an anchor results in the results words scaled according to their similarity with "war", setting "war" and "peace" as opposite anchors (1 and -1, respectively) allows for a measure of bellicosity in a corpus.

This method can be applied to as many words/concepts as the user is interested in (as the automated cosine similarity will handle the other dimensions), and, of note for social scientists, each of these word scalings, which are $I \times 1$, can be multiplied by a $D \times I$ document-term matrix: $D \times I * I \times 1$, scaling the documents in a corpus according to dimensions of interest. These document values can then we used in a regression, and the coefficients can be interpreted in a straightforward way.

6 Inaugurals and Internationalism

In an initial test of this model, I investigate whether the United States saw itself in a new, global role after 1945, as perceived in presidents' inaugural addresses. After 1945, the United States was the global hegemon, and international relations theory argues that this resulted in a shift in American attitudes towards the world (Mearsheimer, 2001). It has been shown that the public takes elite cues on various issues (Zaller, 1992; Druckman and Jacobs, 2015), and since foreign-policy is generally viewed as an elite-led phenomenon (Aldrich et al., 2006), I explore whether, after the second World War, inaugural addresses were more internationally focused than those before the war.

I use the corpus of inaugural addresses available in the `quanteda` R package (Benoit et al., 2018), which contains 58 speeches. I keep words which occur with frequency > 5, and then lowercase and tokenize the texts, resulting in 2705 words. I estimate the model with a context window of 9, with 5 negative samples for every positive sample, and and the number of dimensions $K = 50$. After fitting the model, I compare three possible anchorings: an un-anchored embedding, an embedding anchored on "american," and an embedding scaled with the first dimension anchored on "international" and "domestic;" the results are visible in Table 1. We see that changing the anchoring points changes the most simi-

lar words, however, anchoring helps make these embeddings more interpretable. To test whether there was a statistically significant difference between American perceptions of global roles before and after 1945, I multiply the document-term matrix by the embedding dimension anchored on "international" and "domestic," creating an "internationalism" scale for documents. I test this hypothesis using a one-sided Kolmogorov-Smirnov test, and reject the null hypothesis, that pre-1945 inaugural addresses are less internationalist than the post-1945 addresses at $p < .05$. This means pre-1945 addresses are more "internationalist" than the post-1945 addresses. I plot the differing distributions in Figure 1.

What explains this finding? One possibility, building on Herring (2008, ch. 1), is that the United States was not isolationist prior to 1945, that isolationism was largely a product of the 1920s and 1930s, however, the United States was more unilateral before 1945. Because the multilateral world order was a fact of life after 1945, it is possible presidents were less likely to comment on international affairs, international action was the norm, rather than the exception. Furthermore, the public played a larger role in shaping foreign policy action, particularly during the Vietnam War era, than it had previously (Aldrich et al., 2006), and this could lead to a blurring of the lines between foreign and domestic politics when presidents address the public.

I compare the results from Bayesian Word Embeddings to the results from a standard model used in the social sciences to analyze text, the structural topic model (Roberts et al., 2016)[3]. I find that "domestic" and "international" topics are not linked, the structural topic model captures no relationship between these words. I then investigate the change in "domestic" and "international" topics before and after 1945, and find no effect. There is ample belief and qualitative evidence of a change in American views about the world after 1945, which is not captured in the structural topic model. These differing results suggest that the embeddings are capable of recovering patterns in language that document-based topic models cannot.

7 Diplomacy and the Onset of War

Natural language processing and text as data methods offer the opportunity to quantify decision-

---

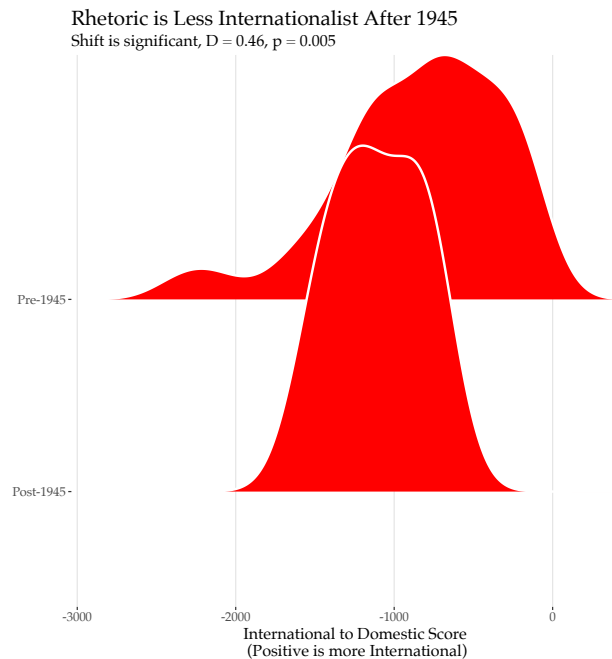[3]Results presented in the Appendix.

Figure 1: After 1945, rhetoric in inaugural addresses becomes less internationalist, and more domestic.

making and attitude among elites, which is notoriously difficult to measure, especially in times of conflict. Existing approaches to measuring elite attitudes often depend on survey or laboratory experiments (Feaver and Gelpi, 2006; LeVeck et al., 2014), however, I offer an alternative approach that allows us to examine elite decisions as they occur. Drawing on a novel corpus of recently digitized diplomatic cables, the *Foreign Relations of the United States* (FRUS), I investigate whether changes in the bellicosity of elite rhetoric precedes an escalation in US hostility. The FRUS dataset provides an exciting opportunity to investigate bellicosity among American foreign policy elites as events happened, as it contains primary source documents of private communications from the policymakers who develop and implement the United States' foreign policy. Among the sources for documents included in FRUS are "Presidential libraries, Departments of State and Defense, National Security Council, Central Intelligence Agency, Agency for International Development, and other foreign affairs agencies as well as the private papers of individuals involved in formulating U.S. foreign policy," with a focus on documents relevant to policy-making (State Department, 2017). When a FRUS volume is compiled, the compiler(s) first identify a set of themes, develop a list all relevant documents, and then select

Word of Interest	Anchor: International, Domestic							
war	large	declarations	pay	carefully	choice	equal	this	greater
peace	practices	meeting	strife	inspiring	confederacy	advance	temple	objections
american	engagements	soil	cultivate	by	heroes	goes	pride	she
international	declare	path	honor	expression	speaking	where	vision	ignorance
national	temple	subject	learned	demand	advance	objections	principle	guard
	Anchor: American							
war	abroad	remedies	violate	slaves	violence	declarations	proposition	sectional
peace	army	plenty	victory	effort	resumption	front	regulation	agreement
american	regards	brief	instrumentality	execute	able	friendly	hands	friendship
international	assembly	european	continent	capable	various	canal	differing	affected
national	now	recognition	corporations	monetary	south	more	character	diversity
	Anchor: None							
war	made	had	peace	force	never	after	still	place
peace	world	nations	war	strength	prosperity	progress	just	security
american	through	opportunity	america	life	justice	right	individual	equal
international	maintain	lasting	fixed	beneficial	settlement	likely	relationship	intercourse
national	most	necessity	common	given	free	first	an	power

Table 1: The most similar words to "war," "peace," "american," "international," and "national," according to each of the anchoring choices, measured via cosine similarity. Choosing appropriate anchors leads to more interpretable embeddings than the unanchored model.

those with the greatest historical import. These are then redacted or declassified, typeset, compared to the original document, and printed and bound (McAllister et al., 2015).

To explore elite bellicosity, I investigate behavior during 1964-1966, the leadup to the Vietnam War, and the breakdown of the "Cold War Consensus" (Krebs, 2015). The era is particularly interesting because, while the United States increased its commitment to Vietnam, it also engaged in several other interventions around the world (Herring, 2008, ch. 16). Thus, we would expect to see that an increase in bellicosity in the FRUS corpus would be correlated with an increase in hostile actions by the United States.

I measure hostility using the Cline Center Historical Event Data, coded from the *New York Times* (Althaus et al., 2017). These data take the form *(DATE, STATE A, ACTION, STATE B)*, where *(STATE A, STATE B)* are directed dyads, *DATE* is the day the event was observed, and *ACTION* is one of five categories of action: neutral, verbal cooperation, verbal conflict, material cooperation, or material conflict (Norris et al., 2017). I select only those events where *STATE A* is the United States, and sum events at the biweekly level. I measure hostility using counts of material conflict events, and display the hostile event counts in Figure 2.

To calculate bellicosity, I first estimate a Bayesian Word Embedding model, with context window of 9, $K = 50$, keeping any word that occurs at least 40 times. I then anchor the embeddings on a "war-peace" dimension. I summarize the results of the anchoring using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), which calculates a low dimensional number of components, similar to principal components analysis. Unlike PCA, UMAP calculates distance using cosine similarity, while balancing both global and local structure in the embeddings (Becht et al., 2019). I present results in Table 2, and the components reveal themes in the corpus, clustering by region and issue, helping highlight the face validity of the embeddings.

To estimate the bellicosity of a given document, I multiply the war-peace dimension by the document term matrix, averaging document bellicosity scores at the bi-weekly level. I plot the bi-weekly bellicosity scores in Figure 2.

To determine if there is a relationship between hostile events and bellicosity, I regress events on the lagged bellicosity (to account for a delay in policy implementation), using a Poisson generalized linear model, due to the count-distributed nature of the outcome.[4] I plot the regression line against the data in Figure 3, and find a positive and statistically significant effect.

This result suggests that bellicosity in elite deliberations, captured in diplomatic documents, results in an increase in conflictual events, which suggests that the documents in the *FRUS* corpus do not simply contain cheap talk, these deliberations ultimately shape policy. These findings also help establish the validity of the "bellicosity" scale, that is, it correlates with an entirely

---

[4]In the appendix, I remove outliers and high-leverage points from the dataset, and fit the model again. Results do not change.

iran	doubtful	communications	tam	bases	relatively	robertson	initials
iranian	blocked	relations	systematically	family	zambia	outflows	secretary
shah	sponsored	appreciably	north	leave	udi	payments	footnotes
aram	ultimatum	masses	hanoi	deployments	tran	liabilities	present
iranians	telecommunications	sites	recce	fixed	rhodesia	fowler	conflict
afghan	recommendation	overtures	drv	precondition	sr	banks	president
squadron	imminent	concurs	vinh	laotian	neighboring	tax	raymond
mnd	jet	harass	chau	reasons	continent	corporations	even

Table 2: The top words from a subset of components estimated from UMAP. Components include a variety of regional and substantive themes. These results help highlight the validity of the embeddings: semantically similar words are appearing near each other in cosine space.
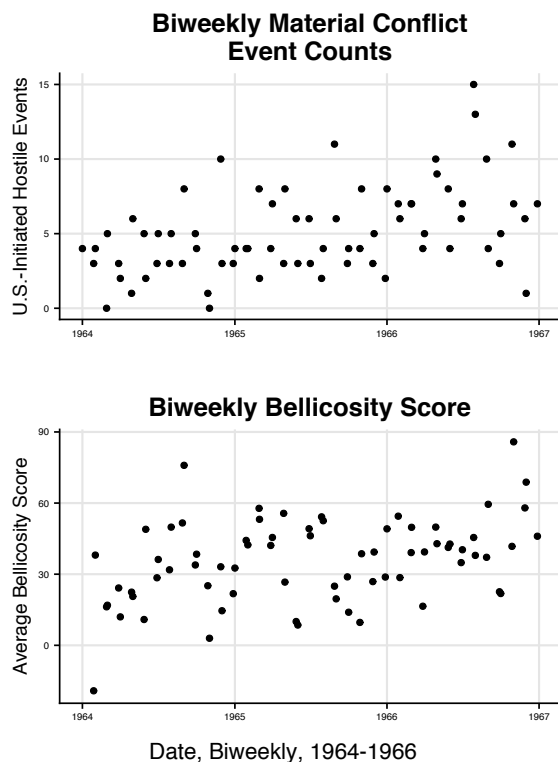

Figure 2: Plotting material conflict event counts and bellicosity scores over time aggregated at the bi-weekly level. Both bellicosity and the count of material conflict events increase with time, as the United States increased its involvement in the Vietnam War.

Figure 3: An increase in the previous bi-weekly period's bellicosity is associated with an increase in U.S.-initiated hostile events. The regression is from a Poisson generalized linear model, and uncertainty is displayed with 95% confidence intervals.

separate dataset, which captures a similar phenomena. All replication materials are available at https://github.com/adamlauretig/bwe_application_naacl_2019.

8 Conclusion

In this paper, I introduced Bayesian Word Embeddings, a method for estimating word embeddings which uses variational bayesian m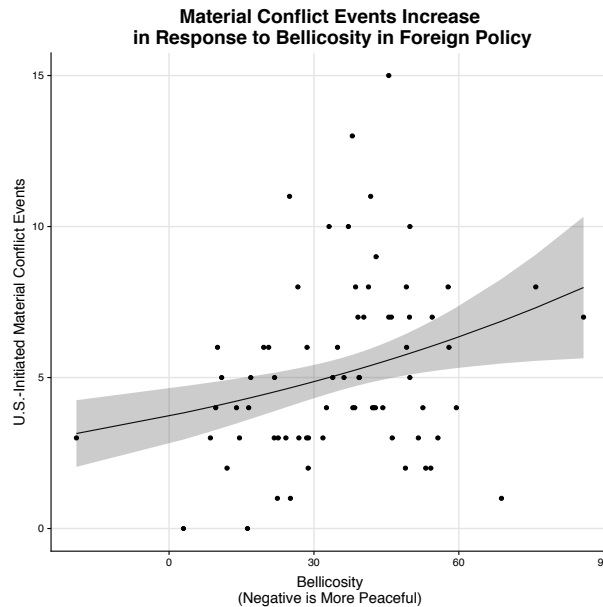ethods. I incorporated Automatic Relevance Determination priors on the embedding dimensions, relaxing the requirement that all dimensions have equal weight. Linking word embeddings to Bayesian latent variable models, I then discussed issues with identification, and solutions proposed in the ideal-point literature, as well as offering an alternative which allows for scaling along dimensions of interest, which creates model model that can then be used in a regression.

I applied Bayesian Word Embeddings to two cases: examining the change in American attitudes about the world before and after 1945 as captured in Presidential inaugural addresses, and then, testing whether an increase in the bellicosity of internal elite discussion (in diplomatic documents) results in an increase in American hostility. I found

that there was a statistically significant different in the views of the world expressed in inaugural addresses, and that this shift was the opposite of what hypotheses generated from international relations theory would expect. When testing the effect of bellicosity on the hostility of American foreign policy, I that an increase in bellicosity resulted in an increase in hostility.

Overall, I have contributed a tool which can serve many purposes for social scientists. By building a probabilistic embedding model, I have constructed a tool which moves beyond document-based inferential approaches to text as data, allowing inference on individual words. This promises new reaches for social scientists, in particular, the promise of crossover with interpretivist work, building on Nelson (2017). Concepts such as securitization theory (Wæver, 1995) draw on the idea that language and word choice by elites shape the attitude of the public, and through the methods introduced above provide the opportunity to generate statistical tests for hypotheses derived from theories like securitization theory.

Future methodological work will follow three tracks. The first will build on Rudolph et al. (2016) and Han et al. (2018), one goal is incorporating document-level metadata into embedding estimation, allowing embeddings to vary according to document-specific attributes, and then, identifying the resulting embeddings. The second will take advantage of stochastic variational inference (Hoffman et al., 2013) to enable Bayesian Word Embeddings to scale to massive corpora. Finally, the third track for future word will involve tying the anchoring approach discussed above with the emerging literature on making casual claims from text (Fong and Grimmer, 2016; Mozer et al., 2018), and taking advantage of the word similarities to identify appropriate linguistic counterfactuals.

Acknowledgments

References

John H. Aldrich, Christopher Gelpi, Peter Feaver, Jason Reifler, and Kristin Thompson Sharp. 2006. Foreign policy and the electoral connection. *Annual Review of Political Science*, 9:477–502.

John H. Aldrich, Jacob M. Montgomery, and David B. Sparks. 2014. Polarization and ideology: Partisan sources of low dimensionality in scaled roll call analyses. *Political Analysis*, 22(4):435–456.

Scott Althaus, Joseph Bajjalieh, John F. Carter, Buddy Peyton, and Dan A. Shalmon. Cline center historical phoenix event data [online]. 2017.

Joseph Bafumi, Andrew Gelman, David K. Park, and Noah Kaplan. 2005. Practical issues in implementing and understanding bayesian ideal point estimation. *Political Analysis*, 13(2):171–187.

Oren Barkan. 2017. Bayesian neural word embedding. In *AAAI*, pages 3135–3143.

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. 2019. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38.

Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

CM Bishop. 1999. Variational principal components. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 509–514. IET.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

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

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

James N. Druckman and Lawrence R. Jacobs. 2015. *Who Governs?: Presidents, Public Opinion, and Manipulation*. University of Chicago Press.

Peter D. Feaver and Christopher Gelpi. 2006. *Choosing your battles: American civil-military relations and the use of force*. Princeton University Press.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *ACL (1)*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Matthew Gentzkow, Bryan T. Kelly, and Matt Taddy. 2017. Text as data. Technical report, National Bureau of Economic Research.

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

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, page mps028.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1489–1501.

Rujun Han, Michael Gill, Arthur Spirling, and Kyunghyun Cho. 2018. Conditional word embedding and hypothesis testing via bayes-by-backprop. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4890–4895.

Serhii Havrylov and Ivan Titov. 2018. Embedding words as distributions with a bayesian skip-gram model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1775–1789.

George C. Herring. 2008. *From colony to superpower: US foreign relations since 1776*. Oxford University Press.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Kosuke Imai, James Lo, and Jonathan Olmsted. 2016. Fast estimation of ideal points with massive data. *American Political Science Review*, 110(4):631–656.

Geng Ji, Robert Bamler, Erik B. Sudderth, and Stephan Mandt. 2017. Bayesian paragraph vectors. *arXiv preprint arXiv:1711.03946*.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. *arXiv preprint arXiv:1803.09288*.

Ronald R. Krebs. 2015. How dominant narratives rise and fall: Military conflict, politics, and the cold war consensus. *International Organization*, 69(04):809–845.

Brad L LeVeck, D Alex Hughes, James H Fowler, Emilie Hafner-Burton, and David G Victor. 2014. The role of self-interest in elite bargaining. *Proceedings of the National Academy of Sciences*, 111(52):18536–18541.

Will Lowe, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. Scaling policy preferences from coded political texts. *Legislative studies quarterly*, 36(1):123–155.

David J.C. MacKay and Radford M. Neal. 1994. Automatic relevance determination for neural networks. In *Technical Report*. Cambridge University.

William B. McAllister, Joshua Botts, Peter Cozzens, and Aaron W. Marrs. 2015. *Toward "Thorough, Accurate, and Reliable": A History of the Foreign Relations of the United States Series*. Government Printing Office.

John J. Mearsheimer. 2001. *The tragedy of great power politics*. WW Norton & Company.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *NAACL-HLT*, volume 13, pages 746–751.

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L Jason Anastasopoulos. 2018. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *arXiv preprint arXiv:1801.00644*.

Laura K. Nelson. 2017. Computational grounded theory: A methodological framework. *Sociological Methods & Research*, page 0049124117729703.

Clayton Norris, Philip Schrodt, and John Beieler. 2017. Petrarch2: Another event coding program. *The Journal of Open Source Software*, 2(9).

Caleb Pomeroy, Niheer Dasandi, and Slava Jankin Mikhaylov. 2018. Multiplex communities and the emergence of international conflict. *arXiv preprint arXiv:1806.00615*.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Douglas Rivers. 2003. Identification of multidimensional spatial voting models. *Typescript. Stanford University*.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, (just-accepted):1–49.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.

State Department. About the foreign relations of the united states series [online]. 2017.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Ole Wæver. 1995. Securitization and desecuritization. In *On Security*, pages 46–86. Columbia University Press.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

John R. Zaller. 1992. *The nature and origins of mass opinion*. Cambridge university press.

A   Comparing BWE to STM

Unlike the results from the Bayesian Word Embedding, the structural topic model detects no difference in topics before and after 1945. The top words, as determined by FREX score, are visible in 3.

International Topic	Domestic Topic
representative	pacific
civilization	territory
making	question
international	whilst
tax	importance
popular	constitution
concern	slavery
supreme	domestic

Table 3: Top eight words from structural topic model for international and domestic topics, by FREX score.

B   Inauguration Robustness Check

One concern with the role of internationalism in inaugural addresses is that by splitting at 1945, the "internationalism" of the pre-1945 sample is due to World War Two, and the Roosevelt presidency. To account for this, I re-estimate the Kolmogorov-Smirnov test from above, excluding the Roosevelt
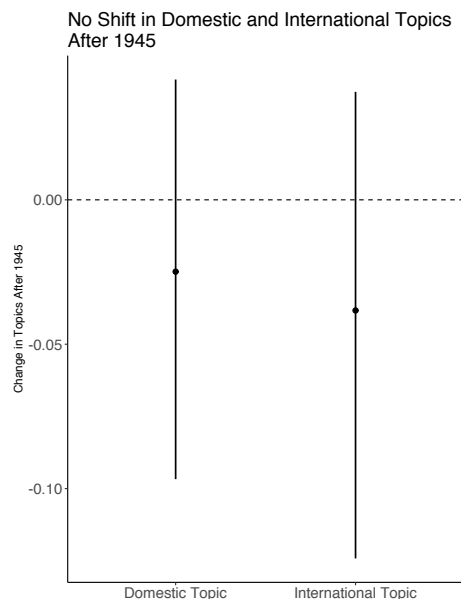

Figure 4: There is no significant difference between foreign and international topics before and after 1945, uncertainty is displayed with 95% confidence intervals.
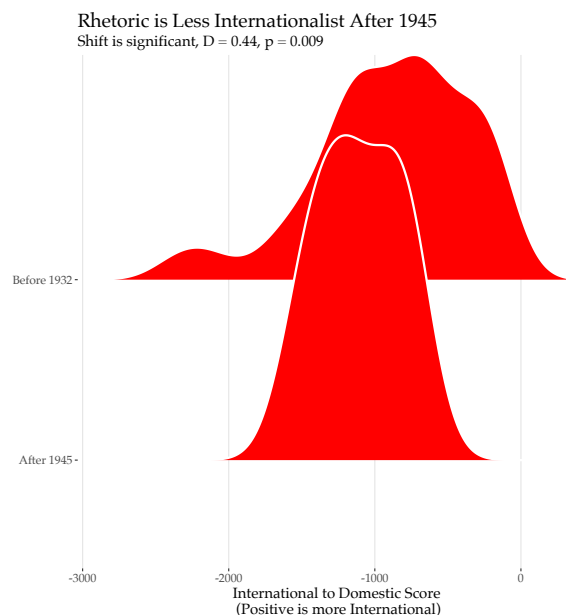

Figure 5: Even excluding the Roosevelt administration, when only examining inaugural addresses from before 1932 and after 1945, the pre-1932 inaugural addresses are more internationalist than the post-1945 addresses.

inaugural addresses, and present the results in 5. This compares inaugurals from before 1932 to those after 1945, and the substantive results do not change.
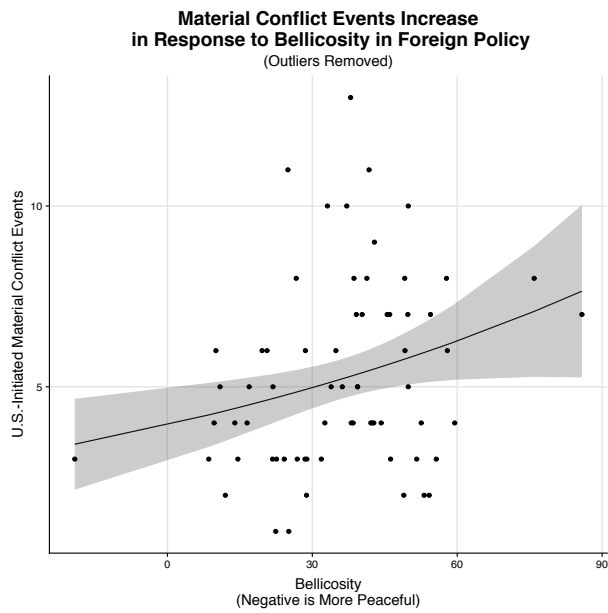
16

**Material Conflict Events Increase
in Response to Bellicosity in Foreign Policy**
(Outliers Removed)

Figure 6: An increase in bellicosity is associated with
an increase in U.S.-initiated hostile events. The regres-
sion is from a Poisson generalized linear model, and
uncertainty is displayed with 95% confidence intervals.

C GLM without Outliers

To ensure that the results presented in Figure 3
were not the results of outliers, I removed any
outliers and high-leverage points, and re-fit the
model. The results were the same, as visible in
Figure 6.