Title
Supervised Classification of Political Text with Topic Models

Author
Holliday, Derek Edward

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Supervised Classification of Political Text with Topic Models

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Derek Edward Holliday

2023

ABSTRACT OF THE THESIS

Supervised Classification of Political Text

with Topic Models

by

Derek Edward Holliday

Master of Science in Statistics

University of California, Los Angeles, 2023

Professor Mark Stephen Handcock, Chair

Statistical classification of texts often use dimension-reduction techniques to reduce the number of features in the classification model. However, this often has the consequences of making inputs difficult for humans to decipher. In this thesis, I propose and algorithm using topic modeling as an interpretable dimension-reduction technique for text classification. I apply the algorithm in the context of nationalized campaign rhetoric amongst gubernatorial candidates in U.S. politics, finding such candidates largely speak about issues germane to their jurisdictions.

The thesis of Derek Edward Holliday is approved.

Chad J. Hazlett

Jeffrey B. Lewis

Christopher N. Tausanovitch

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2023

To my parents, Ann and Joe,

and my wife, Lisa,

whose love and support made all this possible

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

Classification of textual data is of increasing importance in both the statistical and social science literatures (Kowsari et al. 2019; Mirończuk and Protasiewicz 2018; Jurka et al. 2013). The goals of these literatures, however, are often different. The statistics literature is largely concerned with classification accuracy, while the social science literature is largely concerned with the mechanisms leading to certain classifications or the substantive implications of such classifications (Grimmer, Roberts, and Stewart 2022). These differences are particularly acute when categorizations are more abstract and boundaries between categories more permeable. The task of determining whether a news segment is about sports or health policy is very different from determining whether the same news segment as a liberal or conservative leaning or has a more positive or negative tone.

In this thesis, I propose a supervised classification approach for determining textual class using topic models. While classic text classification approaches use individual words or word pairings as model features, I content models using topic proportions as features as a form of dimension-reduction offers significant benefits to model intepretability in cases where categories are more abstract. Additionally, in contrast to many topic modeling approaches requiring post-hoc rationalizations regarding the substantive content of topics, my approach necessitates subject-matter expertise **prior** to model fitting in the selection of training documents for the model. This reduces the need for human coders for model-generated topics and increases transparency in the research process.

To demonstrate the utility of my approach, I present an application in a realm of growing theoretical importance in political science: the "nationalization" of U.S. politics. Generally, "nationalization" in political science refers to national political actors and issues influencing state- and local-level political behavior and outcomes (Jacobson 2015; Abramowitz and Webster 2016; Sievert and McKee 2018; Hopkins 2018). While a large portion of this literature considers how

electoral results across offices have become more correlated, more mainstream depictions reference the apparent nationalization of state and local **campaigns**. Anecdotally, the topics referenced by candidates for state and local office now have more to do with national political debates than ones more germane to their jurisdictions. If these anecdotes reflect a larger pattern, the potential consequences for political representation in state and local government are concerning; voters have little avenue for accountability when candidates campaign on issues over which they have little control. Methodologically, I consider the nationalization of political campaigns as a text classification problem; campaign statements can be classified as being either of national or state topical origin.

The thesis proceeds as follows. First, I review basic literature on text classification and topic models. I then synthesize the two with a technical explanation of my approach to text classification using topic proportions as features. In the next section, I detail the political science context for the application of my approach. After presenting results from the model, I discuss potential extensions to other fields.

2 Text Classification

Basic 2-category text classification seeks to determine the probability a given document D belongs to class Y given feature space $\mathbf{X}$. Formally, $P(Y_D = 1|\mathbf{X})$. Class is assigned at a particular cutoff, typically $P(Y_D = 1|\mathbf{X}) > 0.5$. Where text classification deviates from more classic classification problems, however, is the complexity and size of feature space $\mathbf{X}$. Text can be quantitatively represented in a variety of ways, and simple representations (such as word frequency) can yield parameter counts in the hundreds of thousands. This is further complicated by data sparsity across the feature space; words that exist in some documents may not exist in others.

The first step in text classification, pre-processing, attempts to assuage some of the complexity implicit within textual data (Vijayarani, Ilamathi, and Nithya 2015). Unitization and tokenization is the most basic of steps, defining the unit of analysis (or "token") in the text (Anandarajan, Hill, and Nolan 2018). In the simplest "bag of words" approach, every word is a token. In more complex representation, researchers can use n-grams, sequences of length n words (Robertson and Willett 1998). More advanced still us the use of word embeddings, or vector-representations of words meant to preserve contextual meaning (Schnabel et al. 2015).

Further pre-processing is achieved through standardization via stemming and lemmatization (Anandarajan, Hill, and Nolan 2018). Stemming aims to remove word suffixes to reduce the number of unique tokens in the dataset (Porter 1980), whereas lemmatization aims to slightly preserve distinct word meanings arising from different parts of speech (Korenius et al. 2004). This standardized set of textual data can then be stripped of certain words deemed irrelevant to classification through stopwording, which can make use of pre-existing stopword dictionaries as well

as researcher-defined dictionaries. This process is meant to yield a corpus of text with significantly reduced dimensionality and greater inter-document comparability, with research showing such pre-processing significantly improves classification accuracy (HaCohen-Kerner, Miller, and Yigal 2020; Uysal and Gunal 2014; Korenius et al. 2004).

Still, the dimensionality of the corpus even after pre-processing is likely to be high. This basic difficulty of textual data has been known in the statistical literature for many years, and researchers have developed many techniques to overcome the computational challenges posed by the complexity and size of $\mathbf{X}$. Mosteller and Wallace (1963), for example, use the disputed authorship of several of *The Federalist* papers as a context to test a supervised classification algorithm using a curated set of word frequencies. Working within the computational limits of their time, the authors start with a set of several thousand words and incrementally reduce the feature set down to just 30 relatively high-frequency, high-discrimination words.

While modern computational advances require less aggressive feature selection, large corpuses of text can still prove taxing on time and memory consumption (Kowsari et al. 2019). Certain feature manipulation techniques, such as inversely weighting term frequencies by commonality across documents, can help increase the discriminatory power of individual terms, but does not eliminate the broader issue of feature space. A number of modern algorithms, such as principle components analysis, latent semantic indexing, and linear discriminant analysis help reduce the feature space of text to more manageable levels, but with a focus on classification accuracy rather than feature interpretability (Singh et al. 2022; Kim, Howland, and Park 2005). So while useful for classification, vector-based text representations can have the unintended consequence of creating more opaque interpretations of classification models.

Regardless of feature representation or dimension-reduction techniques, text classification often proceeds using familiar classification algorithms from the machine-learning literature. Documents are split into training and testing sets, where (at minimum) documents in the training set have known class Y. Take the use of logistic regression as an example. Assuming a binary dependent variable $Y \in [0, 1]$, we can use a familiar linear model with a logistic link function:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.1}$$

where p is $P(Y = 1)$ and the model is fit using maximum likelihood. In the context of textual data, k can become quite large without aggressive dimension reduction, making estimation difficult and raising the possibility of overfitting. This often leads to the employment of more complicated classification algorithms such as penalized logistic regression (lasso), naive Bayes classifiers, tree-based classifiers, boosted gradient descent, and support vector machines. The specifics of each of these algorithms are beyond the scope of this thesis, but the general purposes of their mention here is to note their popularity in classification problems in text.

3 Topic Models

While many dimension-reducing strategies exist largely as a means to better classification, probabilistic topic models are a form of dimension-reduction for textual data used moreso to better understand features of the textual data. Specifically, rather than focusing on documents as simple collections of word frequencies, topic models understand documents as probabilistic mixtures of themes, technically referred to as **topics** (Blei 2012). While dozens of topic modeling algorithms exist today, and variations are plentiful, I focus my discussion on two of the most popular in this section: latent Dirichlet allocation (LDA) and structural topics models (STM). Note that full discussions of each model are beyond the scope of this thesis, but the major features are presented below.

3.1 LDA

At its core, LDA attempts to more efficiently represent textual data via short thematic descriptions that maintain distinguishing information about documents. LDA is a generative model advancement over latent semantic indexing (LSI) and probabilistic LSI (pLSI). LSI utilizes a singular value decomposition over the term frequency-inverse document frequency matrix $\mathbf{X}$ of documents in a corpus, such that $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\top}$. Σ here is a diagonal matrix of singular values meant to represent axes of greatest variation. While mathematically how terms are loaded into such dimensions is clear, the substantive interpretation of that loading is quite opaque. pLSI attempts to remedy this with a pseudo-generative model approach, where documents are modeled as mixtures of top-

ics, which themselves are probabilisticly determined by distributions of words. But why is it only "pseudo" generative? As Blei, Ng, and Jordan (2003) note, the joint probability of a document d and term w_n in pLSI are conditionally independent given topic z:

$$P(d, w_n) = P(d) \sum_z P(w_n|z)P(z|d) \tag{3.1}$$

However, d is just an indexing variable for the training set documents, making $P(z|d)$ defined only for observed documents. $P(d_{\text{new}})$ is therefore impossible to estimate in any straightforward way.

LDA remedies this issue by proposing a fully generative mixture model; documents are still mixtures of topics, but now represented as random variables with a Dirichlet prior. Assuming documents d being sequences (non-ordered) of words w_n, with topic mixture proportion θ for topics z_k, LDA assumes the following (abbreviated) form, with a full description is provided by Blei, Ng, and Jordan (2003):

$$\begin{aligned} N &\sim \text{Poisson}(\xi) \\ \theta &\sim \text{Dirichlet}(\alpha) \\ z_n &\sim \text{Multinomial}(\theta) \end{aligned} \tag{3.2}$$

In the algorithm, these are chosen sequentially across each document d, where words are chosen from $P(w_n|x_n, \beta)$. This makes α and β hyperparameters of the model, with topic proportions θ for topics z being the typical quantities of interest.

Estimation is performed via a convexity-based variational algorithm to compute the posterior for the variables of interest:

$$P(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{P(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{P(\mathbf{w}|\alpha, \beta)} \tag{3.3}$$

With the hyperparameters unknown, LDA must maximize the marginal log likelihood $\ell(\alpha, \beta) = \sum_{d=1}^{M} \log P(\mathbf{w}_d|\alpha, \beta)$. By introducing variational parameters γ and ϕ and maximizing the lower bound in respect to them, LDA is able to estimate such values through the following 2-step expecta-

tion maximization algorithm. First, for each document, optimize for $\gamma_d^\star$ and $\phi_d^\star$. With the resulting lower bound, maximize $\ell(\alpha, \beta)$. Repeat until convergence.

The benefits of LDA are numerous. Most importantly, words have probabilistic associations with topics in a highly interpretable way. This means topics are highly indicative of overall document themes, and with a cursory glance at the topic distributions and word associated with the topics, one can represent very long, complicated documents in a highly efficient, highly interpretable way.

3.2 STM

While LDA provides a straightforward, easy to interpret representation of text, it has little to contribute when discussing variation in the generative process of documents, words, and topics. That is, LDA is agnostic to document-level meta data that may influence topic and word prevalence. Perhaps certain topics are more common in certain circumstances, or some words more heavily present in topics under certain conditions. These associations must be done in a post-hoc manner instead of being explicitly modeled in the generative process. To that end, Roberts, Stewart, and Airoldi (2016), Roberts, Stewart, and Tingley (2019), and Roberts et al. (2014) developed the structural topic model (STM).

Starting with the same basic form previously described for LDA by Blei, Ng, and Jordan (2003), STM adds a number of elements. First, topic proportions θ are no longer governed exclusively by hyperparameter α. Instead, θ is integrated into a topic prevalence model, where it is a function of topic prevalence coefficients $\Gamma = \gamma_1 \cdots \gamma_k$ and $\Sigma = \sigma_1 \cdots \sigma_k$ with covariate matrix $\mathbf{X}$. Second, STM includes a topical content model, where a different set of covariates $\mathbf{J}$ controls term frequency β. This yields the following distributional form for documents d:

$$\begin{aligned} \gamma_k &\sim \text{Normal}(0, \sigma_k^2) \\ \theta &\sim \text{LogisticNormal}(\Gamma\mathbf{X}, \Sigma) \\ \mathbf{z}_n &\sim \text{Multinomial}(\theta) \\ \mathbf{w}_n &\sim \text{Multinomial}(\beta_{\mathbf{z}_n}) \end{aligned} \tag{3.4}$$

For identifiability, the logistic normal distribution for θ can be represented by $\eta \sim \text{Normal}(\mu, \Sigma)$

and fixing η_k to zero. With additional parameter κ providing rate-deviation information for words w in topic k with covariates c, the full posterior of interest is represented by:

$$P(\eta, \mathbf{z}, \kappa, \gamma, \Sigma | \mathbf{w}, \mathbf{X}, \mathbf{J}) \propto \Bigg(\prod_{d=1}^{D} \text{Normal}(\eta_d | \mathbf{X}_d \gamma, \Sigma) (\prod_{n=1}^{N} \text{Multinomial}(z_{n,d} | \theta_d) \cdot \text{Multinomial}(w_n | \beta_{d,k})) \Bigg) \cdot \prod P(\kappa) \prod P(\Gamma) \tag{3.5}$$

Similar to LDA, STM uses a variational expectation-maximization algorithm with variational parameter ϕ_d for parameter estimation. Again, the first step optimizes with respect to the variational parameters, while the second step maximies the lower bound for the log likelihood of the parameters of interest. Because of the complexity of the posterior, Roberts, Stewart, and Airoldi (2016) use an approximate posterior and maximize and approximate evidence lower bound instead.

In practice, users of STM input documents as bags of words with document-level covariates, declare the models for topic and term prevalence, and the model returns vectors of topic proportions and coefficients linking covariates to the different prevalence relationships. For example, Parthasarathy, Rao, and Palaniswamy (2019) examine the difference in topics discussed by men and women in rural village councils in India. Formally, they model topic prevalence as a function of speaker gender, finding women are much more likely to speak about topics related to loan programs and self-help financial groups, whereas men are more likely to discuss topics related to employment and wages.

4 The Problem of Intepretability

Both text classification and topic modeling are popular tools in text analysis, but both come with associated interpretability problems, which I discuss briefly here as a motivation for my algorithmic approach to classifying text.

The focus of text classification is maximizing out-of-sample predictive accuracy. To that end, the exact quantitative representation of text that enters the model is of secondary importance. As long as the prediction generated by the representation is accurate, few practitioners are concerned with the nature of the features themselves. This yields quantitative representations that are opaque with regard to the content of the textual data. Algorithms like k-nearest neighbors, LSI, or principle component analysis do not have clear probabilistic linkages of words to the reduced features they generate, making interpretation difficult.

Beyond the quantitative representation of text, text classification is also quite aggressive with feature pruning. If a word or topic is not highly discriminate between document classes, it is either removed during the pre-processing stage or pruned during the model fitting stage. What this omits from the final prediction, however, is a full understanding of the textual content of a document. The very fact that some features are poor predictors is important when discussing what the text is saying. In social science in particular, understanding the content of speech and how that content is or isn't related to particular classes is the inferential focus.

Topic models, alternatively, provide a very natural inferential tool to social scientists. Because documents are probabilistic collections of topics and topics probabilistic collections of words, there is a clear interpretation of the estimates yielded from such models. However, many social scientists

fall victim to poor inferential habits when utilizing topic models. Specifically, many practitioners assign labels to particular topics after model fitting, then use those labels as the variables in subsequent analyses. The problem lies in the post-hoc rationalization of certain word associations within topics. Meaning is only assigned **after** model fitting, meaning less scrupulous researchers could fit a model, determine which topics are associated with positive results, then assign a label to those particular topics that supports their hypothesis. It is important to remember that topical meanin is not baked into the topics; they remain simple probabilistic collections of words.

To some extent, this problem can be alleviated with simple validation exercises. Ying, Montgomery, and Stewart (2021) recommend crowdsourcing human coders to validate topic quality and content. These validation approaches are, of course, labor intensive and costly. Additionally, the framework still assumes the topic themselves are the primary quantity of interest, but often it is the topics' relation to other concepts that is important to social scientists. For example, how are topics related to the class of a document?

These interpretability and inferential problems are non-trivial, but also present an opportunity to merge text classification and topic modeling, as their strengths and weaknesses complement each other. In the next chapter, I propose such a merge.

5 My Approach

Text classification and topic modeling are both general tools providing specific insights into textual data. Both have relative strengths or weaknesses in particular data contexts. My algorithmic approach unifies the strength of the two with a focus on interpretability and including researcher expertise at the training stage rather than in post-hoc rationalizations. The full algorithm is given in Algorithm 1, and I detail the individual steps below.

Algorithm 1 Supervised text classification algorithm with topic models

---

Given text corpuses C_{train} and C_{test}, with documents $D_{i,\text{train}}$ and $D_{i,\text{test}}$ and number of topics k, this algorithm assigns estimated class $\hat{y}_i \in [0, 1]$ to $D_{i,\text{test}}$. y for C_{train} are known.

1. Pre-process C
    a) Lemmatize with words as tokens (removing punctuation and symbols)
    b) Remove lemmas with character lengths of less than researcher-specified amount (default 3)
    c) Remove common English stopwords
    d) Remove custom stopwords
    e) Remove rare words (default 3 or fewer occurrences in C_{train})
2. Compute word and topic probabilities θ for C_{train} using a structural topic model with k topics, yielding $\theta_{1\ldots k}$ for each $D_{i,\text{train}}$
    a) Include STM topic prevalence covariates to account for over-time and geography-specific topical trends
3. Train a classification algorithm predicting y for $D_{i,\text{train}}$ using $\theta_{1\ldots k}$
4. Apply trained STM to C_{test}, yielding $\theta_{1\ldots k}$ for each $D_{i,\text{test}}$
5. Apply trained classification algorithm to C_{test} using $\theta_{1\ldots k}$ for each $D_{i,\text{test}}$, yielding $\hat{y}_i \in [0, 1]$

5.1 Pre-processing

Before the text can be classified, a number of standard pre-processing steps must occur to yield a comparable set of tokens across texts. Following the findings of Korenius et al. (2004) that lemmatization leads to greater precision in text classification than stemming, the first pre-processing step is lemmatization. Specifically, I use the spacyr wrapper for the Python spaCy package, which tokenizes, lemmatizes, and tags texts with parts of speech. I then remove lemmas tagged as either punctuation or symbols. Additionally, I remove tokens with less than three characters, as these shorter words often have limited topical importance independent of other words.

The next step is the removal of stopwords. Research suggests stopword removal can have a significant impact on classification performance, especially when stopwords occur with high frequency (Méndez et al. 2006; Yu 2008). Because the purpose of this algorithm is the classify text through meaningful but abstract content, it is important here to remove words which have little topical content. For this reason, I remove both a generic set of stopwords using a list from Snowball as well as a custom set of words set by the researcher. These words should be frequent words in a corpus that may give meaningful indications of class without giving a meaningful indication of the underlying abstract concept of interest. For example, when performing speaker identification on a conversation between two speakers, those speakers may rarely use their own names when talking. While names, then, would be highly discriminatory in classification, they wouldn't necessarily give insight into the content of the speech itself.

Finally, after lemmatization and stopword removal, I remove words that are particularly uncommon in the training corpus. In my application, I set this cutoff to three or fewer occurrences. This means lower frequency words will have less of an impact on future classification.

5.2 Topic Modeling

Once the corpus has been processed, the algorithm needs to quantitatively represent each of the documents in the training corpus as a k-length vector of topic proportions. To do so, I leverage a

structural topic model (hereafter STM). As previously discussed, STM associates words with topics and topics to documents with certain probabilities (Roberts, Stewart, and Airoldi 2016). The textual data is ingested into the STM as a "bag of words," similar to the latent Dirichlet allocation (LDA). Importantly, researchers using this algorithm should specify a set of topic prevalence covariates; that is, what features of the documents may influence the likelihood that certain topics exist or don't exist? If researchers are evaluating topics that change over time or across geographies, such effects should be accounted for in the STM. This includes the class covariate of interest, as topics should be allowed to vary as a function of class. When the model is eventually applied to the testing set, this covariate will be missing, but the others will give greater accuracy to the specification of topic prevalence.

5.3 Train Classification Model

At this point in the algorithm, it may also become necessary to account for class imbalance in the training data. In some instances, the classification model may take the frequent occurrence of some topic (simply as a function of there being more of one class than another) as evidence for the associated class being omnipresent. There are a number of algorithms available for alleviating class imbalance, including minority oversampling, majority undersampling, and synthetic sample creation. Due to performance advantages found with the synthetic sample creation technique, I prefer the use of ROSE (random over-sampling examples) suggested by Lunardon, Menardi, and Torelli (2014) and Menardi and Torelli (2012). Specifically, ROSE is used to created synthetic documents represented by topic proportions.

The next step in the algorithm requires the training of a classification model, using only the STM-generated vectors of topic proportions for each document as features for the prediction of the class of the document. Ideally, these models should be trained using either a held-out validation set or cross-validation to confirm correct classification is occurring at a high rate out of sample. Any classification model preferred by the researcher can be used at this point, but a preference should be

given to models with easily interpretable coefficients, such as logistic regression. Given the fairly small feature set size k, more complicated models may prove unnecessary once the dimension reduction provided by STM has occurred. Additionally, with the algorithm's focus on interpretability of effects, obscuring such effects at this stage would be counterproductive. In my application, I test five classification models: logistic regression (non-penalized), Naive Bayes, penalized logistic regression (Lasso), boosted gradient descent (XGBoost), and support vector machine.

If multiple classification models are used in the training process, the researcher must either ensemble the model predictions or pre-specify a metric with which to evaluate which model will be chosen as the final classification model. This makes evaluation of the model using either a held-out validation set or cross-validation necessary when evaluating multiple models. While simple prediction accuracy is entirely acceptable as an evaluation statistic, I prefer using AUC (area under the receiver operating characteristic curve) due to its scale and classification-threshold invariance.

5.4 Applying Trained Model

Once the model has been trained and selected, the last step of the algorithm is to apply the trained models to the test corpus. First, the test corpus must be aligned with the training corpus before the trained structural topic model can be applied. Tokens in the test corpus not in the training corpus are dropped. Then the trained structural topic model is applied with word and topic probabilities pre-specified to yield estimated topic proportions for each of the testing documents (researcher-specified covariates, other than class, are also included).

With the test data now represented by topic probabilities, the trained classification algorithm can then be applied to the testing data. This application will yield class predictions for each document given predicted probabilities crossing some threshold (likely 0.5). These classifications will help researchers understand the prevalence of the abstract concept in the testing corpus.

Overall, this algorithm merges the strengths of text classification and topic modeling. Predictive accuracy is gained from the classification process while preserving the interpretability of the inputs.

Furthermore, researcher expertise comes **prior** to model fitting, as class is assigned to the training set documents before the topic model is fit. This means researchers are no longer tempted to give post-hoc interpretations to topics generated by topic models.

6 Context

I apply my text classification approach to an increasingly important literature in political science surrounding the "nationalization" of U.S. politics. Scholars of nationalization use the term to refer to national political actors and issues influencing state and local political activity. Broadly, the concern amongst such scholars is the pressure nationalization puts on expectations for political accountability. If features of state and local politics are defined only in terms of national politics, the ability for voters to hold officials accountable for their actions or functions relevant to their jurisdictions is limited.

One facet of nationalization research is the nationalization of election results. Since the 1970s, the correlation between Presidential and down-ballot vote-shares has increased sharply for candidates of the same party. This patterns extends from higher-salience statewide elections such as for governor and U.S. Congress to lower-level elections such as State Supreme Court or Superintendent of Public Instruction.

Another facet of nationalization left understudied, however, is the nationalization of the political campaigns of candidates for state and local offices. Multiple media outlets characterized a number of gubernatorial races in 2019 as nationalized, with Donald Trump being personally involved in many of the races and his impeachment being seen as a motivating issue for voters, even as governors have no jurisdiction over the issue. Such characterization extend beyond Trump as well, with other examples in West Virginia and Texas in 2011 and 2010, respectively.

There is some empirical evidence to support these more anecdotal claims of nationalized rhetoric. Butler and Sutherland (2023), for example, analyze gubernatorial state of the state addresses from

1960 to 2016 to determine if they have become more "nationalized" in both their similarity to other state of the state addresses and in their similarity to the national State of the Union Address. Using a topic model approach, they find evidence for both; topics are much more likely to be universally covered by all governors and by both governors and presidents. Additionally, Das et al. (2022) analyze the Twitter behavior of incumbent Members of Congress, governors, and mayors in 2018 to similarly determine the amount of topical overlap in their online posting. Again using a topic modeling approach, they find a tight coherence between the rhetorical behavior of governors and Members of Congress, with the topical distribution of the two sets of actors being largely indistinguishable from each other. Mayors, however, maintain a distinct set of discussion topics online.

These studies are not without their shortcomings. Neither is specific to campaigning itself, nor do they reach beyond incumbent politicians. For the Twitter context in particular, politicians often Tweet about non-political topics (commenting on current events, sports, etc.) and engage in the sort of political hobbyism we expect out of any political-attuned citizen (Hersh 2020).

Why might politicians "nationalize" their campaigns? Political science research repeatedly finds voters are significantly motivated by candidate partisanship (Orr and Huber 2019). National policy positions offer very clear signals of partisan type to voters, potentially incentivizing campaigns to promote their candidate's positions on such issues if they believe it appeals to a majority of voters (Vavreck 2009). Furthermore, as state and national policy dimensions become more correlated with each other, national policy positions offer insight into the state policy positions candidates may hold (Shor and McCarty 2011; Caughey and Warshaw 2015). Given the steep decline in both access and attention to state and local media, "nationalized" portions of state campaigns may be the only portions voters ever see (Moskowitz 2020; Martin and McCrain 2019; Hayes and Lawless 2018).

Countervailing pressures also exist to not nationalize campaigns. Most obviously, for state candidate running in partisan geographies hostile to their national partisan counterparts, curating a partisan brand may be a losing proposition. Furthermore, voters may recognize nationalized ap-

peals as being irrelevant to the office being contested, making the candidate seem disingenuous or avoiding accountability for campaign promises. The extent to which voters can successfully assign functional responsibility of offices to candidates is unclear in the literature. Arceneaux (2006) finds some evidence from surveys that voters are able to successfully attribute functional responsibility at high rates, whereas Brown (2010) suggests this assignment is moderated by partisanship. Additionally, using time-series cross-sectional data, de Benedictis-Kessner and Warshaw (2020) find that while governors are held accountable for local economic conditions, members of the president's party are penalized for more national economic trends.

A candidate's partisan standing within a district isn't the only factor that may moderate the nationalization of campaign appeals. Research exists documenting how context changes rhetorical content and style. Specifically, context changes perceptions about audience and introduces constraints on the message itself (Stier et al. 2018; Owen 2014; Bossetta 2018). Candidates are also able to control which audience receives a message. As funding networks become more nationalized, so too then may the messages to outside donors (Reckhow et al. 2016). This makes it critical for studies of nationalization to consider multiple mediums through which candidates communicate their messages.

In summary, campaign rhetoric gives a signal of candidate type. Given national policies cues may be strong signals of type, candidates may be motivated to use them in low-information environments. Countervailing pressures come, however, from the potential lack of accountability created by this dynamic.

7 Application

I apply Algorithm 1 in the context of the nationalization of U.S. politics with the following question: is the rhetoric used by candidates for state office primarily national or state in content? To do so, I define a training corpus of rhetoric with known national or state content and a testing set of campaign rhetoric with unknown content. I fit multiple classification models, choosing one for illustration purposes for the remainder of this thesis.

7.1 Training Data

The definition of a training corpus for problems of abstract classes is in a trivial problem. The training corpus must fulfill two criteria: the classes of the data must be known indisputably, and the content of the data must be substantively related to the abstract concept being measured. For interpretability, it is not enough for just the class to be known, the the predictions yielded by the model are plausibly nonsensical. In the context of nationalization, this proves to be an especially difficult problem. If we believe the hypothesis that politics is nationalized, it is likely that most facets of political speech follow such nationalization, making the distinction between state and national topics more difficult to parse. To alleviate such concerns, I use a training corpus of speeches that fulfill institutional requirements of state and national executive offices: State of the Union and State of the State Addresses. These addresses are given by both Presidents and Governors in front of joint sessions of their respective legislatures, with topics covering the state of current policy decisions, political priorities, and budgeting. They often address current events germane to their

jurisdictions and introduce policy goals for the future. While Butler and Sutherland (2023) do find the topics covered by Presidents and Governors have become more similar, they also find *most* of the topics are distinct to their jurisdictions.

Specifically, my training data includes an original corpus of 1,038 speeches and documents (227 national and 811 state) spanning 2000-2018. For national speeches, I includes all State of the Union addresses made by sitting presidents in the time period as well as opposition party responses to the addresses, inaugural addresses, official Presidential statements, and national party platforms. For state speeches, I use all state of the states addresses and state budget addresses given by governors during the time period.

While the training corpus is meant to include textual data able to distinguish between state and national policy debates based on the policy discussions themselves, there is still some vulnerability to words with non-policy content entering the model and allowing classification to "cheat." By "cheat" I simply mean identify state or national documents by non-policy content that happens to be highly discriminatory between the two classes. For example, references to state names, office titles, or references to state resident nicknames ("Hoosiers", for example) would be highly discriminatory but devoid of policy content, which is the true abstract concept I am trying to classify. To avoid this problem, I compile a list of custom stopwords that are eliminated from the corpus during pre-processing including state names, nicknames, level of office, certain transcription tags that could be specific to context (laughter, applause), and other common non-policy words in speeches (year, will, thank, for example).

7.2 Testing Data

With the training data defined, I now look to the focus of the thesis: has campaign rhetoric nationalized? I investigate this question in three rhetorical areas of ambiguous levels of nationalization where candidates communicate directly with voters: televised debates, televised advertisements, and Twitter.

7.2.1 Televised Debates

One of the most candid forms of political campaigning is through televised debates between candidates for office. While candidates invariably prepare canned responses to expected questions, they must still react to questions as they are asked and challenges from their opponents while staying "on brand" with the rest of their campaign. In this thesis, I analyze an original corpus of 397 debates (86 presidential and 311 gubernatorial) over the same timespan as the training corpus (2000-2018). The transcripts of these debates were retrieved from C-SPAN using a headless web-browsing. These debates cover all U.S. states.

Methodologically, the televised debate context offers a number of important advantages over other rhetorical contexts. Televised debates tend to last at least one hour, providing ample space for candidates to discuss policy details in greater depth than in other forums. While moderators can dictate the broader area of a discussion, candidates are (in)famously able to respond more to the question they wished was asked rather than the one actually asked. Additionally, the context helps control for candidate-level differences that may confound the level of nationalized rhetoric used by their campaigns. Candidates with vastly different amounts of financial support or ideologies have basically equal speaking time.

7.2.2 TV Advertisements

Perhaps the most frequently referenced and visually obvious form of campaigning comes in the form of televised advertisements for political candidates. In this thesis, I limit my analysis to 2,334 advertisements provided by the Wisconsin Media Project (WMP). Of these advertisements, 1,528 are for presidential campaigns and 806 are from gubernatorial campaigns, spanning the 2004 and 2008 election cycles. The text of these ads are scraped from PDF storyboards provided by WMP using the embedded text in the documents. Storyboards are available for 2000 and 2002, but without embedded text, so I exclude them from this analysis. Ads include those run by the candidates themselves as well as affiliated interest groups during both the primary and general election periods.

Of course, the built-in controls of the debate context are eliminate when analyzing rhetoric in advertisements. Candidates with larger financial resources have the ability to run both higher quantities of ads as well as a greater variety of ads regarding both topics and tone. Additionally, financially advantaged candidates can micro-target certain constituencies, perhaps shifting the topical focus (and, therefore, level of nationalization) of the advertisements. The ads themselves are also much shorter than debates, which may induce higher uncertainty in the estimates of topic proportions. While I will conduct the classification agnostic of these concerns in order to give a summary picture of rhetorical nationalization, future analysis should take into consideration how these individual-level characteristics shift incentives and propensity to nationalize.

7.2.3 Twitter

Lastly, I analyze a corpus of over one million tweets from sitting Members of Congress and Governors from 2018. These data are provided by Das et al. (2022) and include 952,425 tweets from Members of Congress and 101,546 tweets from incumbent Governors.

Of all the testing data, the Twitter data are the most unique. The national reference point in these data is no longer the President, but instead Members of Congress. Due to the more parochial nature of Congressional politics, I expect the rhetorical behavior of Members of Congress to include more references to state-specific policy items than Presidential rhetorical behavior. Additionally, only incumbents are included in the data, so there is an imbalance in the partisan composition of the tweets. Futhermore, the analysis period isn't strictly restricted to a "campaign" period. Rather, tweets from the entire year are considered. Lastly, Twitter is unique in that it doesn't represent a purely "political" forum. Politicians can and routinely do tweet about non-political topics, just as their constituents due. I expect this yield more uncertainty in the classification.

7.3 Model Fitting

I the structural topic model to the training corpus according to Algorithm 1. In doing so, I select $k = 40$ topics. Figure 7.1 shows a number of diagnostic test to determine the optimal k balancing likelihood, minimization of residuals, and semantic coherence. While a range of candidate k exist, $k = 40$ provides a good balance of fit and coherence. While $k = 50$ is also a good candidate, I ultimately prefer fewer parameters to avoid overfitting.

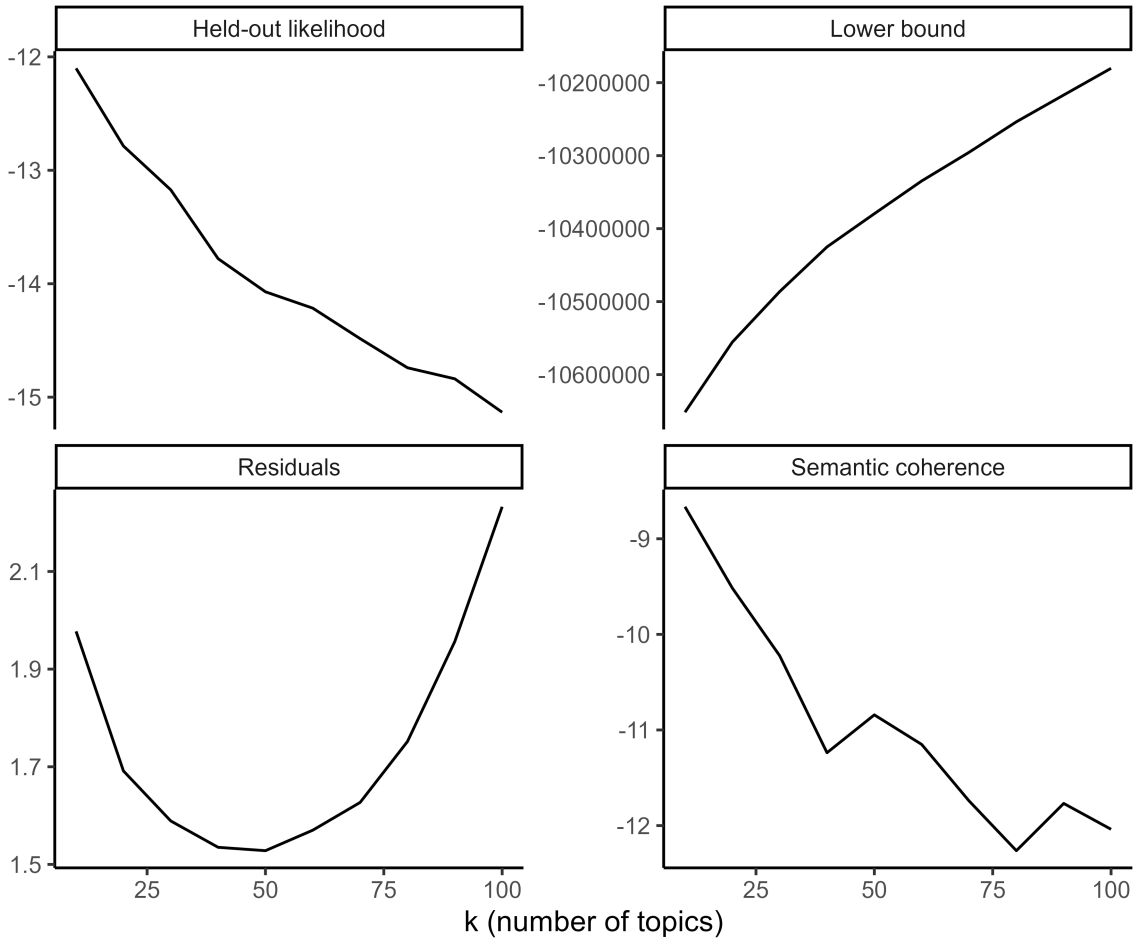

Figure 7.1: Structural Topic Model Diagnostics by Number of Topics

With k chosen, the training corpus of state of the state and state of the union speeches is now represented as k-length vectors of topic proportions. Using those topic proportions, I proceed to fitting five different classification models: logistic regression (non-penalized), naive Bayes, penalized logistic regression (Lasso), boosted gradient descent (XGBoost), and support vector machine.

In doing so, I hold out 20% of the training data as a validation set to determine out-of-sample fit, balancing across state and national classes. The initial results are given in Table 7.1, showing the number of speeches and documents correctly classified as being from their jurisdiction of origin.

Table 7.1: Classification Model Performance on Heldout Documents: Correct vs. Incorrect

Model	Correct (National)	Incorrect	Correct (State)	Incorrect
Logistic Regression (Nonpenalized)	39	1	167	1
Naive Bayes	36	4	162	6
Penalized Logistic Regression (Lasso)	39	1	166	2
Boosted Gradient Descent (XGBoost)	28	12	168	0
Support Vector Machine	40	0	167	1

Across all specifications, classification appears fairly accurate. To get a better statistical sense of how well the models do, Table 7.2 displays both the accuracy of the models as well as the AUC, which I will use to determine the best model to apply to the testing data. Note that this is not simply a function of k being set to 40. As Figure 7.2 shows, non-penalized logistic regression performs competitively with all other models regardless of k specification.

Table 7.2: Classification Model Performance on Heldout Documents: Accuracy and AUC

Model	Accuracy	AUC
Logistic Regression (Nonpenalized)	0.990	0.999
Naive Bayes	0.952	0.984
Penalized Logistic Regression (Lasso)	0.986	0.999
Boosted Gradient Descent (XGBoost)	0.942	0.994
Support Vector Machine	0.995	0.999

Again, all models perform remarkably well. Three models (logit, Lasso, and SVM) have an AUC of 0.999, a near perfect fit. All else being equal, because the algorithm prioritizes interpretability, I use simple logistic regression as the classification model for the remainder of the thesis. As was

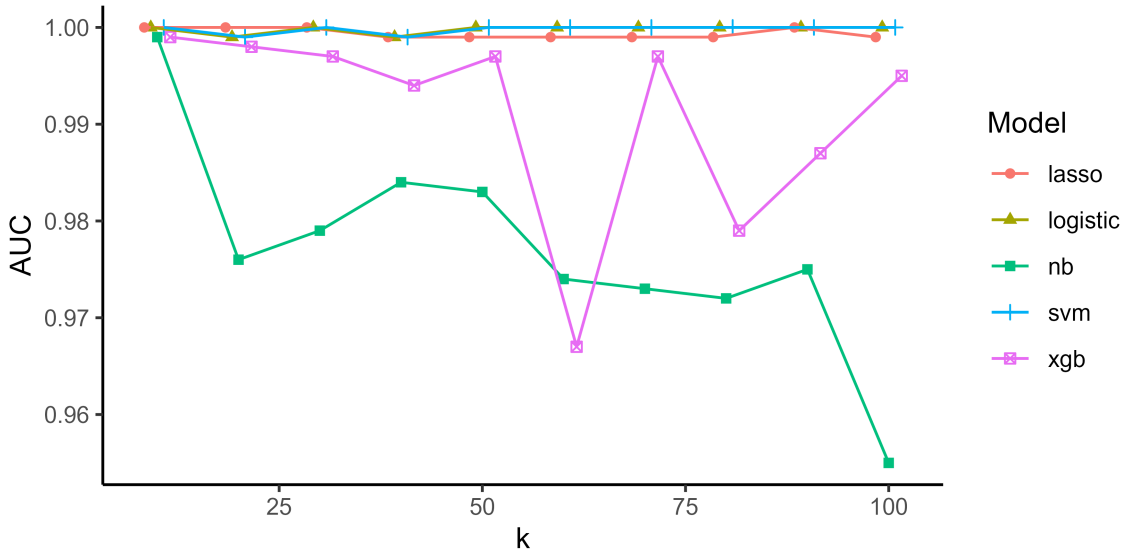

Figure 7.2: Classification Model Performance by Number of Topics

mentioned in the review of text classification, unpenalized logistic regression takes the following form:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{7.1}$$

where the confidence intervals for our probability estimates can be similarly computed via end-point transformation by applying the logistic transformation to $\mathbf{x}\beta \pm z \cdot SE(\mathbf{x}\beta)$:

$$\frac{e^{\mathbf{x}\beta \pm z \cdot SE(\mathbf{x}\beta)}}{1 + e^{\mathbf{x}\beta \pm z \cdot SE(\mathbf{x}\beta)}} \tag{7.2}$$

The final coefficient estimates for the trained model are shown in Figure 7.3. Negative coefficients are more highly associated with national topics, whereas positive coefficients are associated with state topics. Note the smaller number of national topics compared to states topics suggesting a greater variety of policy debates occurring at the state level. This interpretation is further supported by the lesser uncertainty associated with the national topics. Conditional on a document being of national origin, there is a high likelihood it will contain most of the national topics. The same is not true of state documents, which contain only subsets of topics for any given instance.

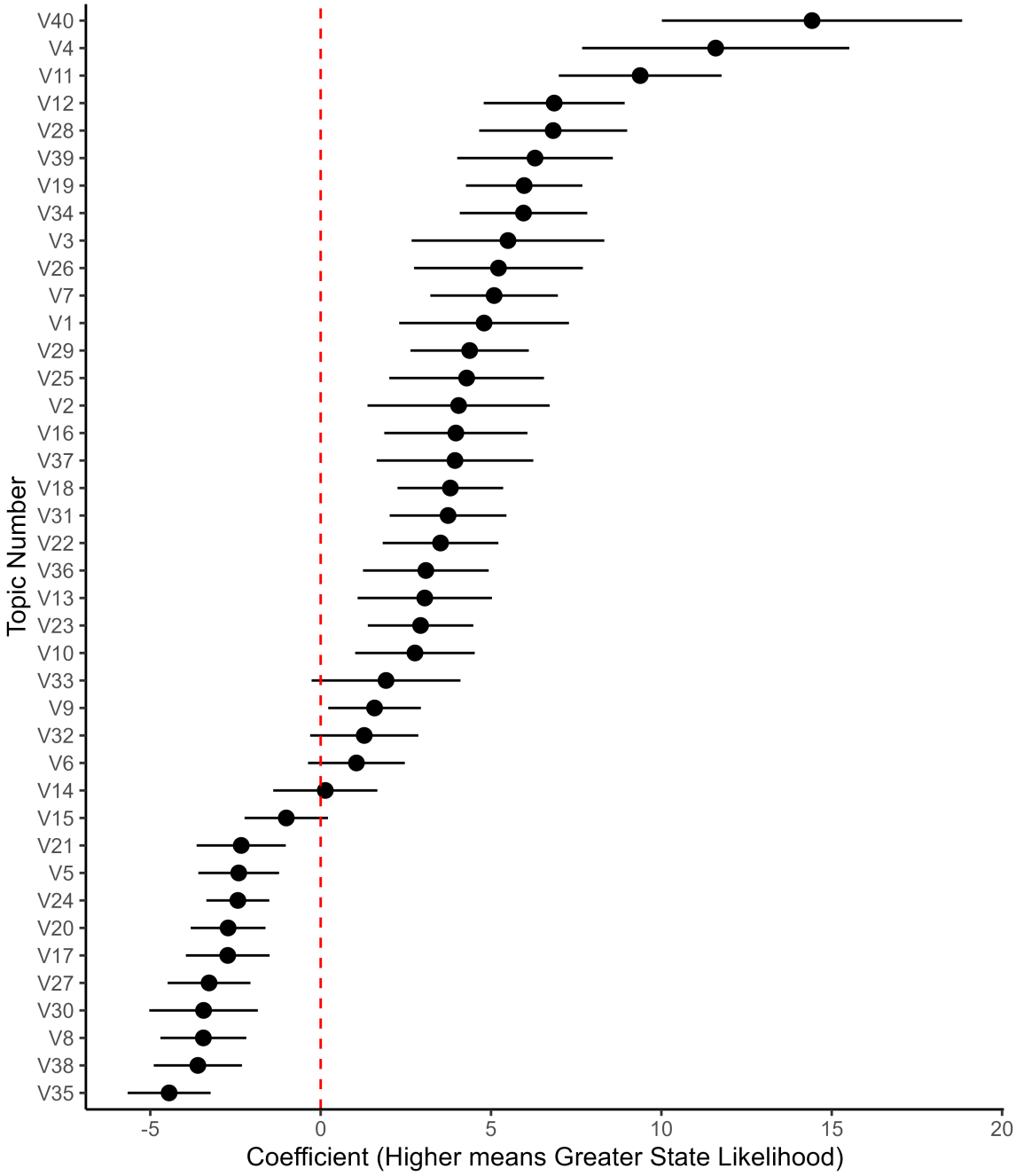

Figure 7.3: Trained Logistic Regression Coefficient Estimates

8 Results

With the classification model fitting, I apply the model to the different documents in the test corpus. The primary quantity of interest in these results is the prediction accuracy for the gubernatorial documents. A higher national classification rate suggests a higher level of national topics being discussed, whereas a higher state classification rate suggests governors are largely discussing topics germane to their jurisdictions in their campaigns. Across all different rhetorical contexts, a general trend emerges: governors largely campaign on topics germane to their jurisdictions. However, there is variance in the level of nationalization within context, which I explore below.

8.1 Televised Debates

Figure 8.1 shows the prediction results of the trained model when applied to the 397 televised debates in the testing corpus, with the box color shaded by percentage. When pooled over the 2000 to 2018 timespan, over 99% of the gubernatorial debates are correctly classified as being of predominantly state content. Presidential debates show a similar result: over 96% of presidential debates are classified as being predominantly national in nature. This suggests that, at least in the context of televised debates, candidates largely discuss policy topics germane to their own jurisdictions, casting into doubt whether such facets of politics have "nationalized." Indeed, only 5 of the 397 debates are "incorrectly" categorized as being primarily composed of topics of the other jurisdiction.

Has this lack of nationalization in televised debates changed over time? I consider the possibility

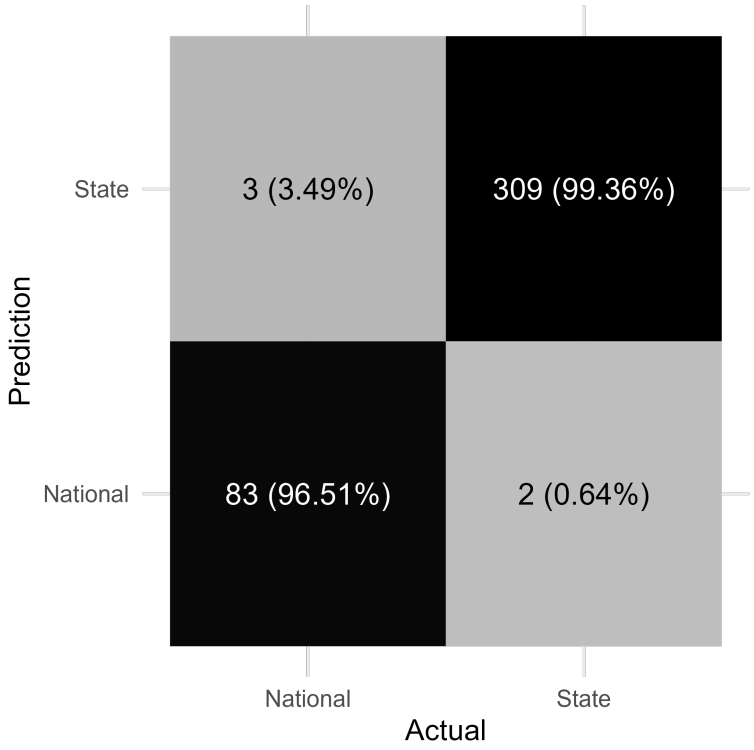

Figure 8.1: C-SPAN Predictions: Confusion Matrix

in Figure 8.2, plotting the predicted probabilities of the debates being national in nature over time, pooling in 2-year intervals. Each debate is represented by a point, colored by whether it was a national (Presidential) or state (gubernatorial) debate. Even when disaggregated in this way, there is very little movement over time in the average level of predicted nationalization. Figure 8.2 also demonstrates how many of the "incorrectly" classified debates are fairly close to the 0.5 cutoff.

It is important to note that these points show predicted probabilities, **not** the estimated proportions of certain topics. This is because certain topics are stronger signals of state or national origin than others, so the predicted probability acts as a sort of weighted average when making classifications.

As an example, a gubernatorial debates was held in New Mexico in 2002 between a Democrat, a Republican, and a Green Party candidate. While the model gave this debate a predicted national probability of 16.9%, the total estimated topic proportion associated with "national" topics was 23.4%, with 6.8% associated with topics having neither a state or national lean. This isn't surpris-

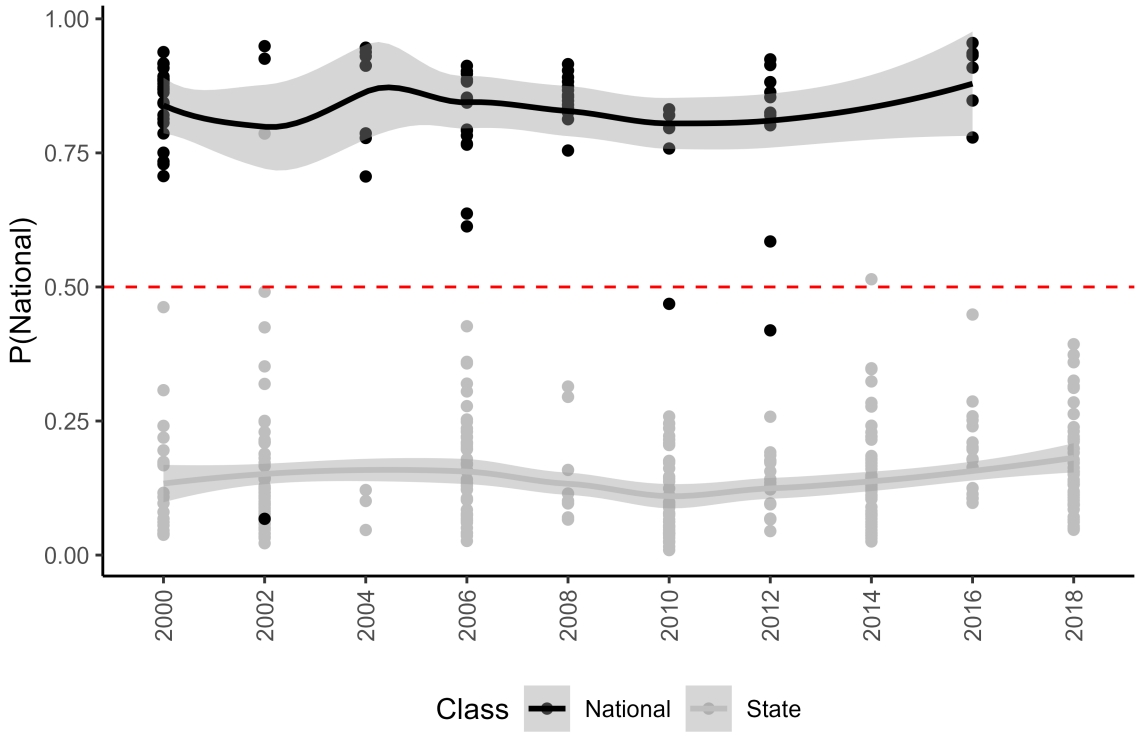

Figure 8.2: C-SPAN Predictions: Over Time

ing, as a large portion of the debate dealt with topics related to the North American Free Trade Agreement and the Iraq war, both topics where governors have no functional jurisdiction.

8.2 Advertisements

Next, I turn to applying the classification model to the 2,334 televised advertisement transcripts from the Wisconsin Media Project. The classification results are given in Figure 8.3. While presidential advertisements maintain a very high national classification rate of of 92%, the state classification rate for gubernatorial advertisements drops to just over 71%. This means, compared to televised debates, a higher proportion of advertisements of gubernatorial campaigns (28%) were classified as being predominantly of national content.

Figure 8.4 shows the distribution of predicted probabilities of being national for both presidential (national) and gubernatorial (state) advertisements. The distributions are both heavily skewed,

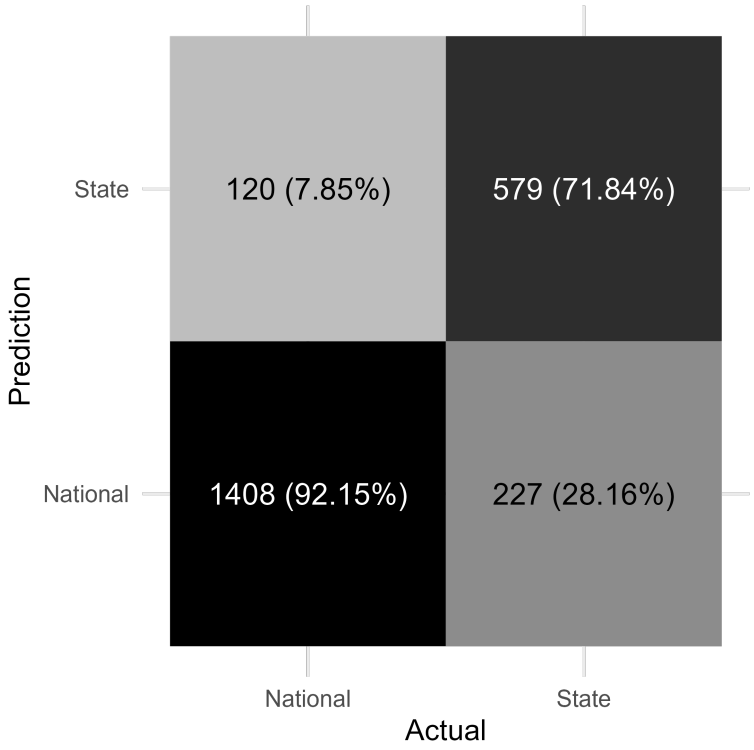

Figure 8.3: Advertisements Predictions: Confusion Matrix

with a very high proportions of advertisements yielding prediction rates close to 100% or 0% for presidential and gubernatorial advertisements (respectively). When ads are "incorrectly" classified, it doesn't seem like such classification occurs with high confidence. That is, the number of cases around 50% predicted probability is very similar to the number of cases at the 100% incorrect side.

What does a "nationalized" advertisement look like in this context? One gubernatorial ad assigned a predicted national probability of 87% reads as follows:

> The big developers, energy companies, and the banking industry just love Pat McCrory and George Bush. Why? Because McCrory and Bush have the same economic philosophy. Less regulation and less oversight to help these companies make even more profit. The result, economic collapse and a Wall Street Bailout. Who ends up paying? You the middle-class. Pat McCrory, stop supporting Bush economics and start supporting more regulation and oversight of big business.

References to then-president George W. Bush and the regulation of Wall Street are clearly

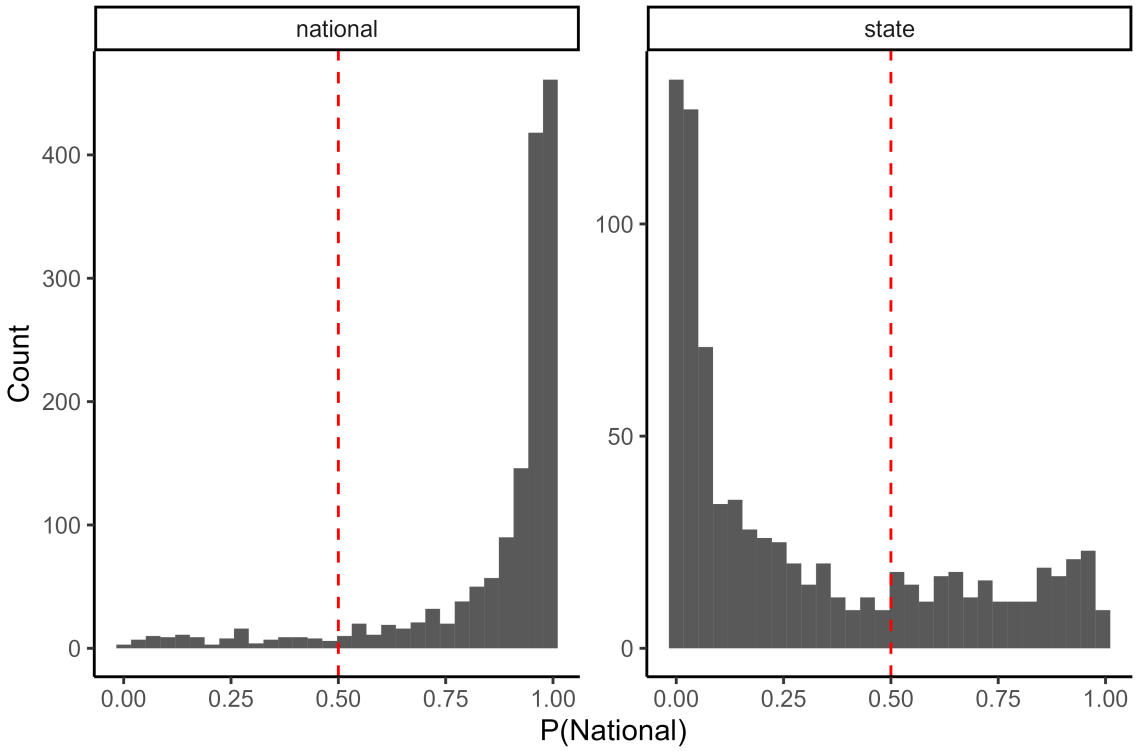

Figure 8.4: Advertisements Predictions: Predicted Probabilities

national-level topics that the governor of North Carolina has little to do with. However, positions on such issues may prove informative to voters when making their decisions between candidates.

8.3 Twitter

Lastly, I apply the classification algorithm to the approximately one million tweets from Das et al. (2022). The prediction results are given in Figure 8.5. Tweets sent by sitting governors in 2018 have a similar state classification rate as gubernatorial advertisements at over 73%, meaning 26% were classified as being predominantly national in nature. For Members of Congress, the content of the tweets is classified fairly evenly across state and national categories, with slighlty more being classified as being of predominantly state content. Given the previous discussion about the parochial and local nature of Congressional districts, this divide is expected.

There are similaries between my results and those of Das et al. (2022). In their analysis, the

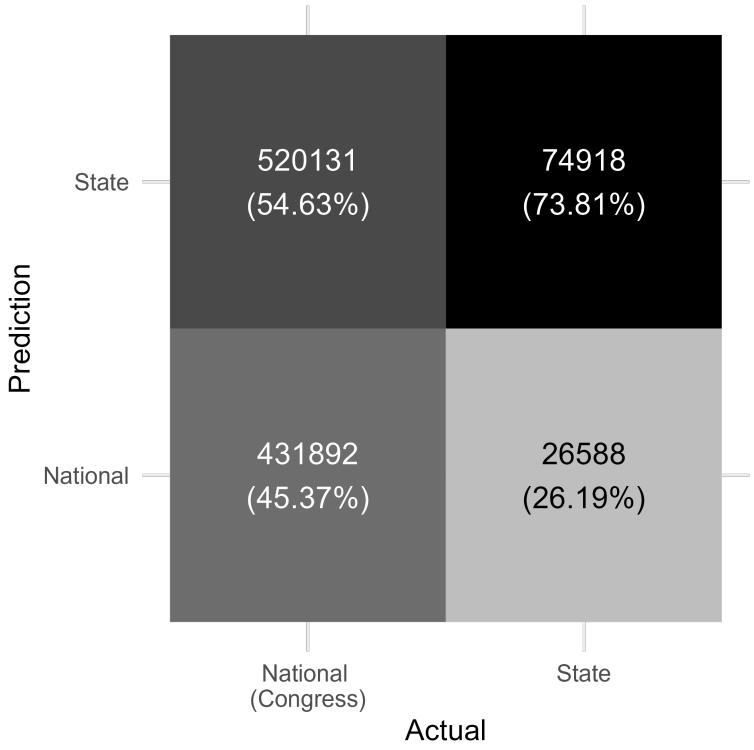

Figure 8.5: Twitter Predictions: Confusion Matrix

authors found the median topic distance between Governors and Members of Congress was 14% greater than the intra-governor distance. In my analysis, I find a difference in the state classification rates between the two offices to be around 19%. While the methods are clearly different, this coherence is reassuring for model performance.

The full distributions of predicted national probabilities are given in Figure 8.6. While there is similar skewedness to the distribution found in the advertisements classifications, the Twitter results show a more pronounced bimodality. Given the relatively short length of tweets (only 280 characters), this is unsurprising, as such short statements are likely to only contain references to one or two topics (or perhaps more correlated topics).

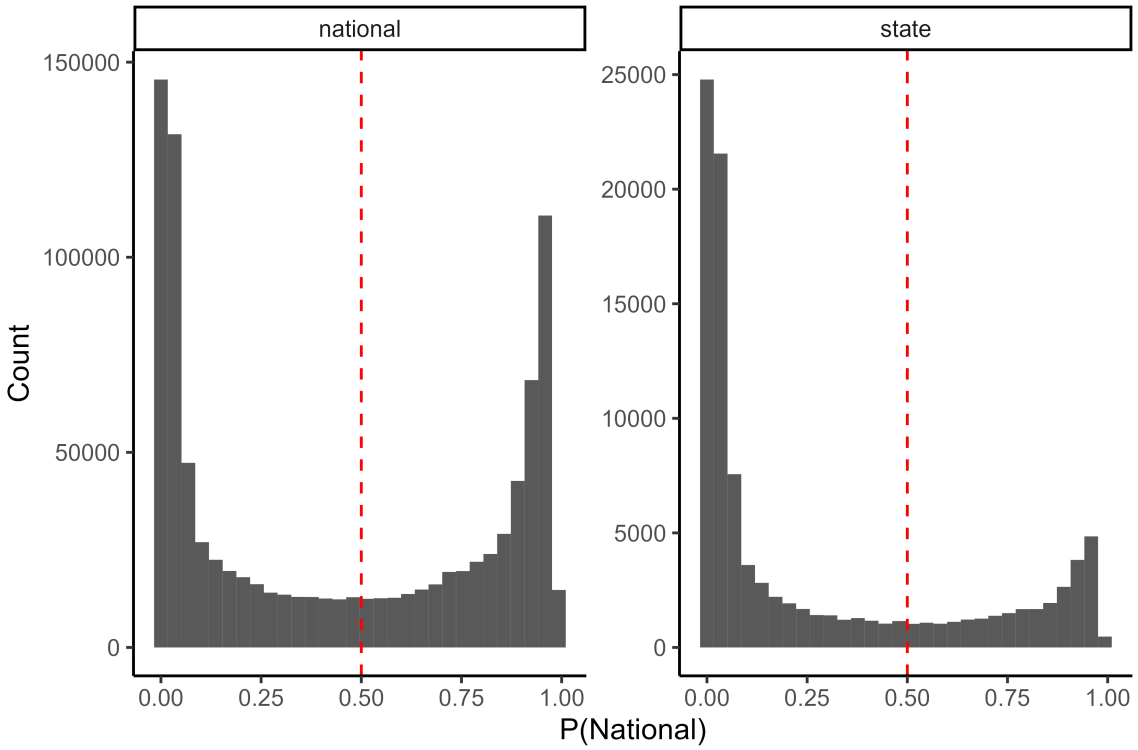

Figure 8.6: Twitter Predictions: Predicted Probabilities

9 Concluding Remarks

In this thesis, I have proposed and implemented a text classification algorithm using a structural topic model for feature reduction. I showed the model performs well in providing interpretable results for the classification of documents into abstract classes in the context of political science. In the substantive application, I found that campaign rhetoric amongst gubernatorial candidates in U.S. state elections predominantly references topics germane to state jurisdictions, casting doubt on the popular notion that "all politics is national."

Future research is needed to further understand the uncertainty built-in to quantitative representations of text. This is not a problem unique to topic model representations, as many other representations assume no uncertainty in the translation of word frequency into lower-dimensional space. This uncertainty may have consequences for the confidence in predictions yielded from the model.

One of the strengths of the algorithm proposed in this thesis is its flexibility in application. Additional research should be done to find best practices for certain defaults within the algorithm. Furthermore, exciting possibilities exist with the generalization of the model to multi-class classification problems. Lastly, future work should perform validation on how well the model performs relative to human coding of testing documents (similar to the topic validation proposed by Ying, Montgomery, and Stewart (2021)).

References

Abramowitz, Alan I., and Steven Webster. 2016. "The Rise of Negative Partisanship and the Nationalization of u.s. Elections in the 21st Century." *Electoral Studies* 41 (March): 12–22. https://doi.org/10.1016/j.electstud.2015.11.001.

Anandarajan, Murugan, Chelsey Hill, and Thomas Nolan. 2018. "Text Preprocessing." In *Practical Text Analytics*, 45–59. Springer International Publishing. https://doi.org/10.1007/978-3-319-95663-3/_4.

Arceneaux, Kevin. 2006. "The Federal Face of Voting: Are Elected Officials Held Accountable for the Functions Relevant to Their Office?" *Political Psychology* 27 (5): 731–54. https://doi.org/10.1111/j.1467-9221.2006.00530.x.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

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

Bossetta, Michael. 2018. "The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 u.s. Election." *Journalism & Mass Communication Quarterly* 95 (2): 471–96. https://doi.org/10.1177/1077699018763307.

Brown, Adam R. 2010. "Are Governors Responsible for the State Economy? Partisanship, Blame, and Divided Federalism." *The Journal of Politics* 72 (3): 605–15. https://doi.org/10.1017/s0022381610000046.

Butler, Daniel M., and Joseph L. Sutherland. 2023. "Have State Policy Agendas Become More Nationalized?" *The Journal of Politics* 85 (1): 351–55. https://doi.org/10.1086/720792.

Caughey, Devin, and Christopher Warshaw. 2015. "The Dynamics of State Policy Liberalism, 1936-2014." *American Journal of Political Science* 60 (4): 899–913. https://doi.org/10.1111/ajps.12219.

Das, Sanmay, Betsy Sinclair, Steven W. Webster, and Hao Yan. 2022. "All (Mayoral) Politics Is Local?" *The Journal of Politics* 84 (2): 1021–34. https://doi.org/10.1086/716945.

de Benedictis-Kessner, Justin, and Christopher Warshaw. 2020. "Accountability for the Local Economy at All Levels of Government in United States Elections." *American Political Science Review* 114 (3): 660–76. https://doi.org/10.1017/s0003055420000027.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data*. Princeton University Press.

HaCohen-Kerner, Yaakov, Daniel Miller, and Yair Yigal. 2020. "The Influence of Preprocessing on Text Classification Using a Bag-of-Words Representation." Edited by Weinan Zhang. *PLOS ONE* 15 (5): e0232525. https://doi.org/10.1371/journal.pone.0232525.

Hayes, Danny, and Jennifer L. Lawless. 2018. "The Decline of Local News and Its Effects: New Evidence from Longitudinal Data." *The Journal of Politics* 80 (1): 332–36. https://doi.org/10.1086/694105.

Hersh, Eitan. 2020. *Politics Is for Power*. Scribner.

Hopkins, Daniel J. 2018. *The Increasingly United States*. University of Chicago Press.

Jacobson, Gary C. 2015. "It's Nothing Personal: The Decline of the Incumbency Advantage in US House Elections." *The Journal of Politics* 77 (3): 861–73. https://doi.org/10.1086/681670.

Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt. 2013. "RTextTools: A Supervised Learning Package for Text Classification." *The R Journal* 5 (1): 6–12. https://doi.org/10.32614/RJ-2013-001.

Kim, Hyunsoo, Peg Howland, and Haesun Park. 2005. "Dimension Reduction in Text Classification with Support Vector Machines." *Journal of Machine Learning Research* 6 (2): 37–53.

http://jmlr.org/papers/v6/kim05a.html.

Korenius, Tuomo, Jorma Laurikkala, Kalervo Järvelin, and Martti Juhola. 2004. "Stemming and Lemmatization in the Clustering of Finnish Text Documents." In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. ACM. https://doi.org/10.1145/1031171.1031285.

Kowsari, Jafari Meimandi, Heidarysafa, Mendu, Barnes, and Brown. 2019. "Text Classification Algorithms: A Survey." *Information* 10 (4): 150. https://doi.org/10.3390/info10040150.

Lunardon, Nicola, Giovanna Menardi, and Nicola Torelli. 2014. "ROSE: A Package for Binary Imbalanced Learning." *The R Journal* 6 (1): 79. https://doi.org/10.32614/rj-2014-008.

Martin, Gregory J., and Joshua McCrain. 2019. "Local News and National Politics." *American Political Science Review* 113 (2): 372–84. https://doi.org/10.1017/s0003055418000965.

Menardi, Giovanna, and Nicola Torelli. 2012. "Training and Assessing Classification Rules with Imbalanced Data." *Data Mining and Knowledge Discovery* 28 (1): 92–122. https://doi.org/10.1007/s10618-012-0295-5.

Méndez, J. R., E. L. Iglesias, F. Fdez-Riverola, F. Díaz, and J. M. Corchado. 2006. "Tokenising, Stemming and Stopword Removal on Anti-Spam Filtering Domain." In *Current Topics in Artificial Intelligence*, 449–58. Springer Berlin Heidelberg. https://doi.org/10.1007/11881216_47.

Mirończuk, Marcin Michał, and Jarosław Protasiewicz. 2018. "A Recent Overview of the State-of-the-Art Elements of Text Classification." *Expert Systems with Applications* 106 (September): 36–54. https://doi.org/10.1016/j.eswa.2018.03.058.

Moskowitz, Daniel J. 2020. "Local News, Information, and the Nationalization of u.s. Elections." *American Political Science Review* 115 (1): 114–29. https://doi.org/10.1017/s0003055420000829.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." *Journal of the American Statistical Association* 58 (302): 275–309. https://doi.org/10.1080/01621459.1963.10500849.

Orr, Lilla V., and Gregory A. Huber. 2019. "The Policy Basis of Measured Partisan Animosity in

the United States." *American Journal of Political Science* 64 (3): 569–86. https://doi.org/10.1111/ajps.12498.

Owen, Diana. 2014. *New Media and Political Campaigns*. Edited by Kate Kenski and Kathleen Hall Jamieson. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.016_update_001.

Parthasarathy, Ramya, Vijayendra Rao, and Nethra Palaniswamy. 2019. "Deliberative Democracy in an Unequal World: A Text-as-Data Study of South India's Village Assemblies." *American Political Science Review* 113 (3): 623–40. https://doi.org/10.1017/s0003055419000182.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14 (3): 130–37. https://doi.org/10.1108/eb046814.

Reckhow, Sarah, Jeffrey R. Henig, Rebecca Jacobsen, and Jamie Alter Litt. 2016. ""Outsiders with Deep Pockets": The Nationalization of Local School Board Elections." *Urban Affairs Review* 53 (5): 783–811. https://doi.org/10.1177/1078087416663004.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003. https://doi.org/10.1080/01621459.2016.1141684.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "Stm: An r Package for Structural Topic Models." *Journal of Statistical Software* 91 (2). https://doi.org/10.18637/jss.v091.i02.

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): 1064–82. https://doi.org/10.1111/ajps.12103.

Robertson, Alexander M., and Peter Willett. 1998. "Applications of n-Grams in Textual Information Systems." *Journal of Documentation* 54 (1): 48–67. https://doi.org/10.1108/eum0000000007161.

Schnabel, Tobias, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. "Evaluation Meth-

ods for Unsupervised Word Embeddings." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1036.

Shor, Boris, and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105 (3): 530–51. https://doi.org/10.1017/s0003055411000153.

Sievert, Joel, and Seth C. McKee. 2018. "Nationalization in u.s. Senate and Gubernatorial Elections." *American Politics Research* 47 (5): 1055–80. https://doi.org/10.1177/1532673x18792694.

Singh, Ksh. Nareshkumar, S. Dickeeta Devi, H. Mamata Devi, and Anjana Kakoti Mahanta. 2022. "A Novel Approach for Dimension Reduction Using Word Embedding: An Enhanced Text Classification Approach." *International Journal of Information Management Data Insights* 2 (1): 100061. https://doi.org/10.1016/j.jjimei.2022.100061.

Stier, Sebastian, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. 2018. "Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter." *Political Communication* 35 (1): 50–74. https://doi.org/10.1080/10584609.2017.1334728.

Uysal, Alper Kursat, and Serkan Gunal. 2014. "The Impact of Preprocessing on Text Classification." *Information Processing &Amp; Management* 50 (1): 104–12. https://doi.org/10.1016/j.ipm.2013.08.006.

Vavreck, Lynn. 2009. *The Message Matters*. Princeton University Press.

Vijayarani, S., J. Ilamathi, and M. Nithya. 2015. "Preprocessing Techniques for Text Mining - an Overview." *International Journal of Computer Science & Communication Networks* 5 (1): 7–16.

Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart. 2021. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* 30 (4): 570–89. https://doi.org/10.1017/pan.2021.33.

Yu, B. 2008. "An Evaluation of Text Classification Methods for Literary Study." *Literary and Linguistic Computing* 23 (3): 327–43. https://doi.org/10.1093/llc/fqn015.