Advance Access publication February 16, 2009

Political Analysis (2008) 16:372–403
doi:10.1093/pan/mpn018

Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict

Burt L. Monroe
Department of Political Science, Quantitative Social Science Initiative, The Pennsylvania State University, e-mail: burtmonroe@psu.edu (corresponding author)

Michael P. Colaresi
Department of Political Science, Michigan State University, e-mail: colaresi@msu.edu

Kevin M. Quinn
Department of Government and Institute for Quantitative Social Science, Harvard University, e-mail: kevin_quinn@harvard.edu

Entries in the burgeoning ''text-as-data'' movement are often accompanied by lists or visualizations of how word (or other lexical feature) usage differs across some pair or set of documents. These are intended either to establish some target semantic concept (like the content of partisan frames) to estimate word-specific measures that feed forward into another analysis (like locating parties in ideological space) or both. We discuss a variety of techniques for selecting words that capture partisan, or other, differences in political speech and for evaluating the relative importance of those words. We introduce and emphasize several new approaches based on Bayesian shrinkage and regularization. We illustrate the relative utility of these approaches with analyses of partisan, gender, and distributive speech in the U.S. Senate.

1 Introduction

As new approaches to, and applications of, the ''text-as-data'' movement emerge, we find ourselves presented with many collections of disembodied words. Newspaper articles, blogs, and academic papers burst with lists of words that vary or discriminate across groups of documents (Gentzkow and Shapiro 2006; Quinn et al. 2006; Diermeier et al. 2007; Sacerdote and Zidar 2008; Yu et al. 2008), pictures of words scaled in some political space (Schonhardt-Bailey 2008), or both (Monroe and Maeda 2004; Slapin and Proksch 2008). These word or feature lists and graphics are one of the most intuitive ways to convey the

Author's note: We would like to thank Mike Crespin, Jim Dillard, Jeff Lewis, Will Lowe, Mike MacKuen, Andrew Martin, Prasenjit Mitra, Phil Schrodt, Corwin Smidt, Denise Solomon, Jim Stimson, Anton Westveld, Chris Zorn, and participants in seminars at the University of North Carolina, Washington University, and Pennsylvania State University for helpful comments on earlier and related efforts. Any opinions, findings, and conclusions or recommendations expressed in the paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

© The Author 2009. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

key insight from such analyses—the content of a politically interesting difference. Accordingly, we notice when such analyses 1) produce key word lists with odd conceptual matches,[1] 2) remove words from lists before presenting them to the reader (Diermeier et al. 2007) or 3) produce no word lists or pictures at all (Hillard et al. 2007; Hopkins and King 2007).

These word visualizations and lists are common because they serve three important roles. First, they are often intended (implicitly) to offer semantic validity to an automated content analysis—to ensure that substantive meaning is being captured by some text-as-data measure (Krippendorff 2004). That is, the visualizations reflect either a selection of words or a word-specific measure that is intended to characterize some semantic political concept of direct interest, for example, topic (Quinn et al. 2006), ideology (Diermeier et al. 2007), or competitive frame (Schonhardt-Bailey 2008). Second, the visualizations reflect either a selection of words or a word-specific measure that is intended to feed forward to some other analysis. For example, these word estimates might be used to train a classifier for uncoded documents (Yu et al. 2008), to scale future party manifestos against past ones (Laver et al. 2003), to scale individual legislators relative to their parties (Monroe and Maeda 2004), or to evaluate partisan bias in media sources (Gentzkow and Shapiro 2006). Third and more broadly, the political content of the words themselves—words tell us what they mean—allows word lists and visualizations of words to compactly present the very political content and cleavages that justify the enterprise. If we cannot find principled ways to meaningfully sort, organize, and summarize the substantial and at times overwhelming information captured in speech, the promise of using speeches and words as observations to be statistically analyzed is severely compromised.

In this paper, we take a closer look at methods for identifying and weighting words and features that are used distinctly by one or more political groups or actors. We demonstrate some fundamental problems with common methods used for this purpose. Major problems include failure to account for sampling variation and overfitting of idiosyncratic differences. Having diagnosed these problems, we offer two new approaches to the problem of identifying political content. Our proposed solution relies on a model-based approach to avoid inefficiency and shrinkage and regularization to avoid both infinite estimates and overfitting the sample data. We illustrate the usefulness of these methods by examining partisan framing, the content of representation and polarization over time, and the dimensionality of politics in the U.S. Senate.

2 The Objectives: Feature Selection and Evaluation

There are two slightly different goals to be considered here: *feature selection* and *feature evaluation*. With feature selection, the primary goal is a binary decision—in or out—for the inclusion of features in some subsequent analysis. We might, for example, want to know *which words* are reliably used differently by two political parties. This might be for the purpose of using these features in a subsequent model of individual speaker positioning or for a qualitative evaluation of the ideological content of party competition to give two examples. In the former case, fear of overfitting leads us to prefer parsimonious specifications, as with variable selection in regression. In the latter case, computers can estimate but

[1] *Motorway* is a Labour word? (Laver et al. 2003). The single most Republican word is *meth*? (Yu et al. 2008). The word that best defines the ideological left in Germany is *pornographie*? (Slapin and Proksch 2008). Martin Luther King and Ronald Reagan's speeches were distinguished by the use of *him* (MLK) and *she* (Reagan)? (Sacerdote and Zidar 2008).

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

not interpret our models; data reduction is necessary to reduce the quantity of information that must be processed by the analyst. Further, the scale of speech data is immense. Feature selection is useful because we necessarily need a lower dimensional summary of the sample data. Additionally, we know a priori that different speech patterns across groups can flow from both partisan and nonpartisan sources, including idiosyncrasies of dialect and style.

With feature evaluation, the goal is to quantify our information about different features. We want to know, for example, *the extent to which each word* is used differently by two political parties. This might be used to tell us how to weight features in a subsequent model or allow us, in the qualitative case, to have some impression of the relative importance of each word for defining the content of the partisan conflict. The question is not which of these terms are partisan and which are not, but which are the most partisan, on which side, and by how much. Again, in all these cases, parsimony and clarity are virtues. Overfitting should be guarded against because 1) we are interested, not solely in the sample data, but inferring externally valid regularities from that sample data and 2) a list of thousands of words, all of which are approximately equal in weight, is less useful than a list that is winnowed and summarized by some useful criterion.

3 Methods for Lexical Feature Selection and Evaluation

To fix ideas, we use a running example, identifying and evaluating the linguistic differences between Democrats and Republicans in U.S. Senate speeches on a given topic. The topics we use, like ''Defense'' or ''Judicial Nominations'' or ''Taxes'' or ''Abortion'', are those that emerge from the topic model of Senate speech from 1997 to 2004 discussed in Quinn et al. (2006).[2] In this section, our running example is an analysis of speeches delivered on the topic of ''Abortion'' during the 106th Congress (1999–2000), often the subject of frame analysis (Adams 1997; McCaffrey and Keys 2008; Schonhardt-Bailey 2008).[3] The familiarity of this context makes clear the presence, or absence, of semantic validity under different methods.

The lessons are easily generalized and applied to other contexts (gender, electoral districts, opposition status, multiparty systems, etc.), which we demonstrate in the final section. We take the set of potentially relevant lexical features to be the counts of word stems (or ''lemmas'') produced in aggregate across speeches by those of each party. In the interest of direct communication, we will simply say ''words'' throughout. This generalizes trivially to other feature lists: nonstemmed words, *n*-grams, part-of-speech-tagged words, and so on.[4] So, in short, we wish to select partisan words, evaluate the partisanship of words, or both.

[2]Our data are constructed from the Congressional Speech Corpus under the Dynamics of Rhetoric and Political Representation Project. The raw data are the electronic version of the (public domain) U.S. Congressional Record maintained by the Library of Congress on its THOMAS system. The Congressional corpus includes both the House and the Senate for the period beginning with the 101st Congress (1988–) to the present. For this analysis, we truncate the data at December 31, 2004. We then parse these raw html documents to into tagged XML versions. Each paragraph is tagged as speech or nonspeech, with speeches further tagged by speaker. For the topic coding of the speeches, the unit of analysis is the speech document. That is, all words spoken by the same speaker within a single html document. The speeches are then further parsed to filter out capitalization and punctuation using a set of rules developed by Monroe et al. (2006). Finally, the words are stemmed using the Porter Snowball II stemmer (Porter 2001). This is simply a set of rules that groups words with similar roots (e.g., speed and speeding). These stems are then summed within each speech document and the sums used within the Dynamic Multinomial Topic Coding model (Quinn et al. 2006).

[3]For our particular substantive and theoretical interests, pooling all speech together has several undesirable consequences and topic modeling is a crucial prior step.

[4]The computational demands for preprocessing, storage, and estimation time can vary, as can the statistical fit and semantic interpretability of the results.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

In the sections that follow, we consider several sets of approaches. The first are ''classification'' methods from machine learning that are often applied to such problems. The idea would be, in our example, to try to identify the party of a speaker based on the words she used. We provide a brief discussion about why this is inappropriate to the task. Second, we discuss several commonly used nonmodel-based approaches. These use a variety of simple and not-so-simple statistics but share the common feature that there is no stated model of the data-generating process and no implied statements of confidence about the conclusions. This set of approaches includes techniques often used in journalistic and informal applications, as well as techniques associated with more elaborate scoring algorithms. Here, we also discuss a pair of ad hoc techniques that are commonly used to try to correct the problems that appear in such approaches. Third, we discuss some basic model-based approaches that allow for more systematic information accounting. Fourth and finally, we discuss a variety of techniques that use shrinkage, or regularization, to improve results further. Roughly speaking, our judgment of the usefulness of the techniques increases as the discussion progresses, and the two methods in the fourth section are the ones we recommend.

We start with some general definitions of notation that we use throughout. Let $w = 1, \ldots, W$ index words. Let $\mathbf{y}$ denote the W-vector of word frequencies in the corpus, and $\mathbf{y}_k$ the W-vector of word frequencies within any given topic k. We further partition the documents across speakers/authors. Let $i \in I$ index a partition of the documents in the corpus. In some applications, i may correspond to a single document; in other applications, it may correspond to an aggregation, like all speeches by a particular person, by all members of a given party, or by all members of a state delegation, depending on the variable of interest. So, let $\mathbf{y}_k^{(i)}$ denote the W-vector of word frequencies from documents of class i in topic k.

In our running example of this section, we focus on the lexical differences induced by party, so we assume that $I = \{D, R\}$—Democratic and Republican—so that $y_{kw}^{(D)}$ represents the number of times Democratic Senators used word w on topic k, which in this section is exclusively abortion. $y_{kw}^{(R)}$ is analogously defined.

3.1 *Classification*

One approach, standard in the machine learning literature, is to treat this as a classification problem. In our example, we would attempt to find the words (w) that significantly predict partisanship (p). A variety of established machine learning methods could be used: Support Vector Machines (Vapnik 2001), AdaBoost (Freund and Schapire 1997), random forests (Breiman 2001) among many other possibilities.[5] These approaches would attempt to find some classifier function, c, that mapped words to some unknown party label, $c : w \rightarrow p$.

The primary problem of this approach, for our purposes, is that it gets the data generation process backwards. Party is not plausibly a function of word choice. Word choice is (plausibly) a function of party. Any model of some substantive process of political language production—such as the strategic choice of language for heresthetic purposes (Riker 1986)—would need to build the other way.

Nor does this correctly match the inferential problem, implying we (a) observe partisanship and word choice for some subset of our data, (b) observe only word choice for another subset, and (c) wish to develop a model for accurately inferring partisanship for the second subset. Partisanship is perfectly observed. While it is possible to infer future unobserved

[5]A detailed explanation of these approaches is beyond the scope of the paper. See Hastie et al. (2001) for a introduction.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

word choice with a model that treated this information as given ($P(p|w)$) by inverting the probability using Bayes' rule, this would entail knowing something about the underlying distribution of words ($P(w)$). More problematically, in many applications we would be using an uncertain probabilistic statement ($P(p|w)$), in place of what we know—the partisan membership of an individual. This effectively discards information unnecessarily.

Hand (2006) has noted several other relevant problems with these, nicely supplemented for political scientists by Hopkins and King (2007). Yu et al. (2008) apply such an approach in the Congressional setting, using language in Congressional speech to classify speakers by party. Their Tables 6–8 list features (words) detected for Democratic-Republican classification by their method, one way of using classification techniques like support vector machines or Naive Bayes for feature selection.

3.2 *Nonmodel-Based Approaches*

Many of the most familiar approaches to the problems of feature selection and evaluation are not based on probabilistic models. These include a variety of simple statistics, as well as two slightly more elaborate algorithmic methods for ''scoring'' words. The latter include the *tf.idf* measure in broad use in computational linguistics and the WordScores method (Laver et al. 2003) prominent in political science. In this section, we demonstrate these methods for our running example, discuss how they are related, and evaluate potential problems in the semantic validity of their results.

3.2.1 Difference of frequencies

We start with a variety of simple statistics. One very naive index is simply the difference in the absolute word-use frequency by parties: $y_{kw}^{(D)} - y_{kw}^{(R)}$. If Republicans say *freedom* more than Democrats, then it is a Republican word. The problem with this, of course, is that it is overwhelmed by whoever speaks more. So, in our running example of speech on abortion, we find that the most common words (*the*, *of*, *and*, *is*, *to*, *a*) are considered Republican. The mistake here is obvious in our data, but perhaps less obvious when one does something like compare the number of database hits for *freedom* in a year of the *New York Times* and the *Washington Post*, without making some effort to ascertain how big each database is. This is common in journalistic accounts[6] and common enough in linguistics that Geoffrey Nunberg devotes a methodological appendix in *Talking Right* to explaining why it is a mistake (Nunberg 2006, 209–10).

3.2.2 Difference of proportions

Taking the obvious step of normalizing the word vectors to reflect word proportions rather than word counts, a better measure is the difference of proportions on each word. Defining the observed proportions by $f_{kw}^{(i)} = y_{kw}^{(i)}/n_k^{(i)}$. Now, the evaluation measure becomes $f_{kw}^{(D)} - f_{kw}^{(R)}$.

Figure 1 shows the results of applying this measure to evaluate partisanship of words on the topic of abortion during the 106th (1999–2000) Senate. The scatter cloud plots these values for each word against the (logged) total frequency of the word in this collection of

[6]For example, a recent widely circulated graphic in *The New York Times* (available at http://www.nytimes.com/interactive/2008/09/04/us/politics/20080905_WORDS_GRAPHIC.html) provided a comparative table of the absolute frequencies of selected words and phrases in the presidential convention speeches of four Republicans and four Democrats (Ericson 2008).

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

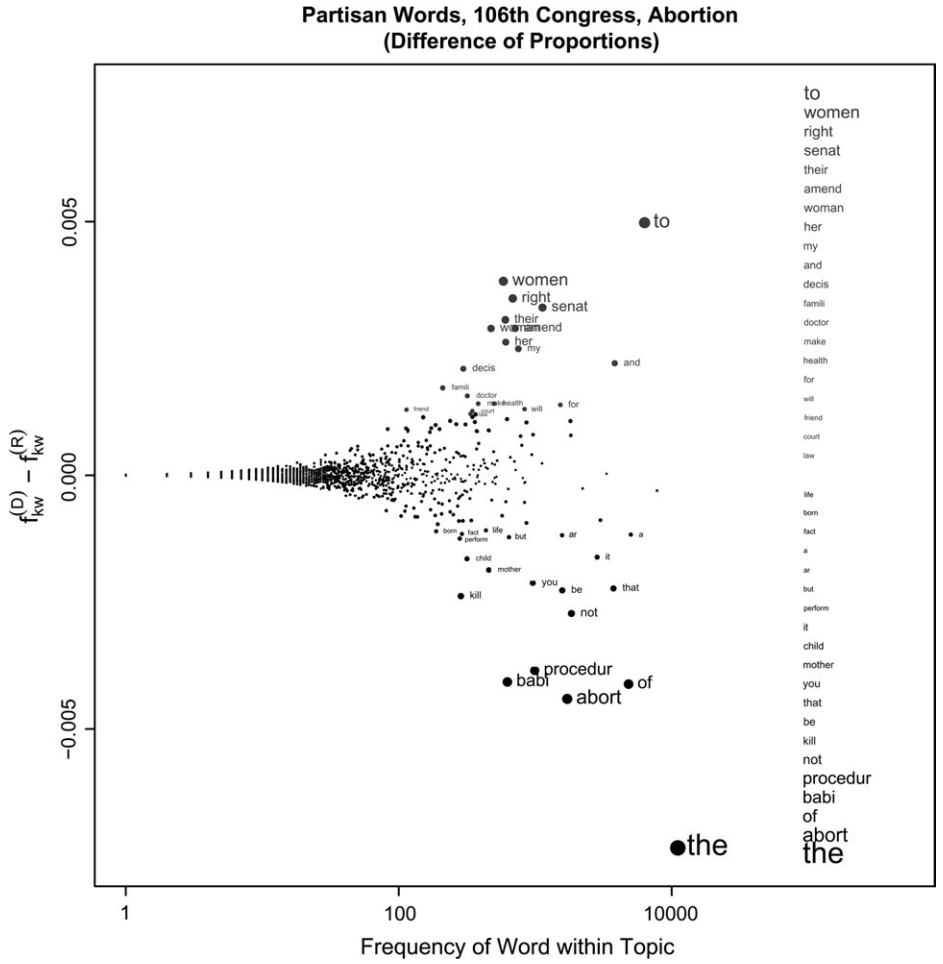

Fig. 1 Feature evaluation and selection using $f_{kw}^{(D)} - f_{kw}^{(R)}$. Plot size is proportional to evaluation weight, $|f_{kw}^{(D)} - f_{kw}^{(R)}|$. The top 20 Democratic and Republican words are labeled and listed in rank order to the right. The results are almost identical for two other measures discussed in the text: unlogged *tf.idf* and frequency-weighted WordScores.

speeches. In this and subsequent figures for the running example, the *y*-axis, size of point, and size of text all reflects the evaluation measure under consideration, in this case $f_{kw}^{(D)} - f_{kw}^{(R)}$. For the top 20 most Democratic and most Republican words, the dots have been labeled with the word, again plotted proportionally to the evaluation measure. These 40 words are repeated, from most Democratic to most Republican, down the right-hand side of the plot.

This is an improvement. There is no generic partisan bias based on volume of speech, and we see several of the key frames that capture known differences in how the parties frame the issue of abortion. For example, Republicans encourage thinking about the issue from the point of view of *babies*, whereas Democrats encourage thinking about the issue from the point of view of *women*. But the lack of overall semantic validity is clear in the overemphasis on high-frequency words. The top Democratic word list is dominated by *to* and includes *my*, *and*, and *for*; the top Republican word list is dominated by *the* and

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

includes *of*, *not*, *be*, *that*, *you*, *it*, and *a*. As is obvious in Figure 1, the sampling variation in difference of proportions is greatest in high-frequency words. These are not partisan words; they are just common ones.

The difference of proportions statistic is ubiquitous. It appears in journalistic (e.g., the top half of the convention speech graphic discussed in footnote 6, which notes, e.g., that ''Republican speakers have talked about reform and character far more frequently than the Democrats'') and academic accounts.[7] The problem with high-frequency words is often masked by referring to only selected words of interest in isolation.

3.2.3 Correction: removing stop words

A common response to this problem in many natural language processing applications is to eliminate ''function'' or ''stop'' words that are deemed unlikely to contain meaning. This is also true in this particular instance, as many related applications are based on difference of proportions on non–stop words only. This is the algorithm underlying several ''word cloud'' applications increasingly familiar in journalistic and blog settings,[8] as well as more formal academic applications like Sacerdote and Zidar (2008).

We note, however, the practice of stop word elimination has been found generally to create more problems than it solves, across natural language processing applications. Manning et al. (2008) observe: ''The general trend . . . over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever'' (p. 27). They give particular emphasis to the problems of searching for phrases that might disappear or change meaning without stop words (e.g., ''to be or not to be''). In our example, a stop list would eliminate a word like *her*, which almost definitely has political content in the context of abortion,[9] and a word like *their*, which might (e.g., *women and their doctors*).

More to the point, this ad hoc solution diagnoses the problem incorrectly. Function words are not dominant in the partisan word lists here because they are function words, but because they are frequent. They are more likely to give extreme values in differences of proportions just from sampling variability. Eliminating function words not only eliminates words like *her* inappropriately but it also elevates high-frequency non–stop words inappropriately. The best example here is *Senate*, which is deemed Democratic by difference of proportions, but is, in this context, simply a high-frequency word with high sampling variability.

3.2.4 Odds

Alternatively, we can examine the proportions in ratio form, through odds. The observed ''odds'' (we assert no probability model yet) of word w in group i's usage are defined as $O_{kw}^{(i)} = f_{kw}^{(i)}/(1 - f_{kw}^{(i)})$. The odds ratio between the two parties is $\theta_{kw}^{(D-R)} = O_{kw}^{(D)}/O_{kw}^{(R)}$.[10]

[7] For example, ''Hillary Clinton uses the word security 8.8 times per 10,000 words while Obama . . .uses the word about 6.8 times per 10,000 words'' (Sacerdote and Zidar 2008).

[8] Examples (in which the exact algorithm is proprietary) include Wordle (http://wordle.net) and tag clouds from IBM's Many Eyes visualization site (http://services.alphaworks.ibm.com/manyeyes/page/Tag_Cloud.html).

[9] Try making a statement about the Democratic position on abortion without using the word *her*.

[10] We could also work with risk ratios, $f_{kw}^{(D)}/f_{kw}^{(R)}$, which function more or less identically with low probability events, like the use of any particular word.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

This is generally presented for single words in isolation or as a metric for ranking words. Examples can be found across the social sciences, including psychology[11] and sociology.[12] In our data, the odds of a Republican using a variant of *babi* are 5.4 times those of a Democratic when talking about abortion, which seems informative. But, the odds of a Republican using the word *April* are 7.4 times those of a Democrat when talking about abortion, which seems less so.

3.2.5 Log-odds-ratio

Lack of symmetry also makes odds difficult to interpret. Logging the odds ratio provides a measure that is symmetric between the two parties. Working naively, it is unclear what we are to do with words that are spoken by only one party and therefore have infinite odds ratios. If we let the infinite values have infinite weight, the partisan word list consists of only those spoken by a single party. The most Democratic words are then *bankruptci*, *Snow*[*e*], *ratifi*, *confidenti*, and *church*, and the most Republican words are *infant*, *admit*, *Chines*, *industri*, and *40*. If we instead throw out the words with zero counts in one party, the most Democratic words are *treati*, *discrim*, *abroad*, *domest*, and *privacy*, and the most Republican words are *perfect*, *subsid*, *percent*, *overrid*, and *cell*.

One compromise between these two comes from the common advice to ''add a little bit to the zeroes'' (say, 0.5, as in Agresti [2002, 70–1]). If we calculate a smoothed log-odds-ratio from such supplemented frequencies, $\tilde{f}_{kw}^{(i)} = f_{kw}^{(i)} + \varepsilon$, we get the results as shown in Figure 2.

Note that regardless of the zero treatment, the most extreme words are obscure ones. These word lists are strange in a way opposite from that produced by the difference of proportions shown in Figure 1. These do not seem like words that most fundamentally define the partisan division on abortion. Words with plausible partisan content on abortion (*infant*, *church*) are overwhelmed by oddities that require quite a bit more investigation to interpret (*Chines*, *bankruptci*). In short, the semantic validity of this measure is limited.[13]

The problem is again the failure to account for sampling variability. With log-odds-ratios, the sampling variation goes down with increased frequency, as is clear in Figure 2. So, this measure will be inappropriately dominated by obscure words.

3.2.6 Correction: eliminating low-frequency words

Although this is a very basic statistical idea, it is commonly unacknowledged in simple feature selection and related ranking exercises. A common response is to set some frequency ''threshold'' for features to ''qualify'' for consideration. Generally, this simply removes the most problematic features without resolving the issue.

For an example of the latter, consider the list of Levitt and Dubner (2005; 194–8), in their freakonomic discussion of baby names, of ''The Twenty White Girl [Boy] Names That Best Signify High-Education Parents.'' They identify these, from California records,

[11] One study in which preschoolers were observed as they talked found that girls were six times more likely to use the word *love* and twice as likely to use the word *sad* (Senay, Newberger, and Waters 2004, 68).

[12] Americans used the word *frontier* in business names . . . more than 4 times more often than France (Kleinfeld and Kleinfeld 2004).

[13] There *is* abortion partisanship in these words. For example, Democrat Charles Schumer introduced an amendment to a bankruptcy reform bill designed to prevent abortion clinic protesters from avoiding fines via bankruptcy protections. We argue only that these do not have face validity as the *most important* words.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

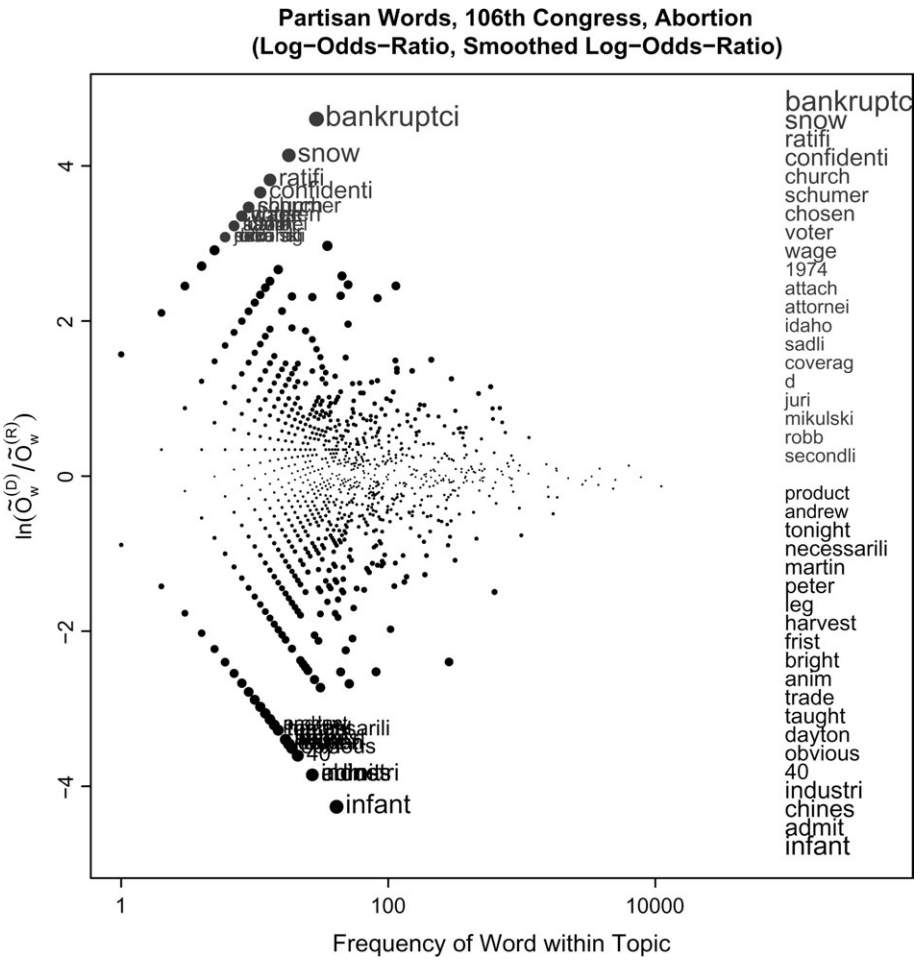

Fig. 2 Feature evaluation and selection using $\hat{\delta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $\left|\hat{\delta}_{kw}^{(D-R)}\right|$. Top 20 Democratic and Republican words are labeled and listed in rank order. The results are identical to another measure discussed in the text: the log-odds-ratio with uninformative Dirichlet prior.

with a list ranked by average mother's education. The top five girl names are *Lucienne*, *Marie-Claire*, *Glynnis*, *Adair*, and *Meira*; the top five boy names are *Dov*, *Akiva*, *Sander*, *Yannick*, and *Sacha*. A footnote clarifies that a name must appear at least 10 times to make the list (presumably because a list that allowed names used only once or twice might have looked *ridiculous*). It seems likely that each of these words was used exactly, or not many more than, 10 times in the sample. We would get a similar effect by drawing a line at 10 or 100 minimum uses of a word in Figure 2, with the method then selecting the most extreme examples from the least frequent qualifying words.

This mistake is also common in many of the emergent text-as-data discussions in political science. One example is given by Slapin and Proksch (2008), made more striking because they provide a plot that clearly demonstrates the heteroskedasticity. Their Figure 2 looks like Figure 2 turned on its side. The resulting lists of their Table 1 contain many obscurities. This is probably only an issue of interpretation. The item response model they describe, like that of Monroe and Maeda (2004) (see also Lowe [2008]), accounts for

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

variance when the word parameters are used to estimate speaker/author positions. That is, despite their use of captions like ''word weights'' and ''Top Ten Words placing parties on the left and right,'' these are really the words with the 10 leftmost and rightmost point estimates, not the words that have the most influence in the estimates of actor positions.

3.2.7 tf.idf (Computer Science)

It is common practice in the computational linguistics applications of classification (e.g., Which bin does this document belong in?) and search (e.g., Which document(s) should this set of terms be matched to?) to model documents not by their words but by words that have been weighted by their *tf.idf*, or *term frequency—inverse document frequency*. Term frequency refers to the relative frequency (proportion) with which a word appears in the document; document frequency refers to the relative frequency with which a word appears, at all, in documents across the collection. The logic of *tf.idf* is that the words containing the greatest information about a particular document are the words that appear many times in that document, but in relatively few others. *tf.idf* is recommended in standard textbooks (Jurafsky and Martin (2000, 651–4) (Manning and Schütze 1999, 541–4) and is widely used in document search and information retrieval tasks.[14] To the extent *tf.idf* reliably captures what is distinctive about a particular document, it could be interpreted as a feature evaluation technique.

The most common variant of *tf.idf* logs the *idf* term—this is the ''ntn'' variant (natural *tf* term, logged *df* term, no normalization, see Manning and Schütze [1999, 544]). So, letting df_{kw} denote the fraction of groups that use the word w on topic k at least once, then:

$$tf.idf^{(i)}_{kw}(\text{ntn}) = f^{i}_{kw} \ln\left(1/df_{kw}\right). \tag{1}$$

Qualitatively, the results from this approach are identical to the infinite log-odds-ratio results given earlier. The most partisan words are the words spoken the most by one party, while spoken not once by the other (*bankruptci*, *infant*).[15] Clearly, the logic of logging the document frequency[16] breaks down in a collection of two documents.

Alternatively, we can use an unlogged document frequency term—the ''nnn'' (natural tf term, natural *df* term, no normalization; see Manning and Schütze [1999, 544]) variant of *tf.idf*.

$$tf.idf^{(i)}_{kw}(\text{nnn}) = f^{i}_{kw} / df_{kw}. \tag{2}$$

The results for our running example are nearly identical, qualitatively and quantitatively with those from raw difference of proportions, shown in Figure 1. The weights are correlated (at +0.997 in this case) and differ only in doubling the very low weights of the relatively low-frequency words used by only one party.[17] In any case, for our purposes, neither version of *tf.idf* has clear value. See Hiemstra (2000) and Aizawa (2003) for efforts to put *tf.idf* on a probabilistic or information-theoretic footing.

[14] We note over 15,000 hits for the term in Google Scholar.
[15] The degenerate graphic of this result is omitted for space reasons, but available in the web appendix.
[16] Due to the large number of documents in many collections, this measure is usually squashed with a log function (Jurafsky and Martin 2000, 653).
[17] We omit this mostly redundant graphic here. It is available in the web appendix.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

3.2.8 WordScores (Political Science)

Perhaps the most prominent text-as-data approach in political science is the WordScores procedure (Laver et al. 2003), which embeds a feature evaluation technique in its algorithm. The first step of the algorithm establishes scores for words based on their frequencies within ''reference'' texts, which are then used to scale other ''virgin'' texts (for further detail, see Lowe [2008]).

In our running example, we calculate these by setting the Democrats at $+1$ and the Republicans at -1. Then the raw WordScore for each word is:

$$W_{kw}^{(D-R)} = \frac{y_{kw}^{(D)}/n_k^{(D)} - y_{kw}^{(R)}/n_k^{(R)}}{y_{kw}^{(D)}/n_k^{(D)} + y_{kw}^{(R)}/n_k^{(R)}}. \tag{3}$$

Figure 3 shows the results of applying this stage of the WordScores algorithm to our running example. The results bear qualitative resemblance to those with the smoothed log-odds-ratios shown in Figure 2. As shown in Figure 3, the extreme W_{kw} are received by the words spoken by only one party. As with several previous measures, the maximal words are obscure low-frequency words.

The ultimate use for WordScores, however, is for the spatial placement of documents. When the W_{kw} are taken forward to the next step, the impact of any word is proportional to its relative frequency. That is, the implicit evaluation measure, $W_{kw}^{*(D-R)}$, is

$$W_{kw}^{*(D-R)} = \frac{y_{kw}^{(D)}/n_k^{(D)} - y_{kw}^{(R)}/n_k^{(R)}}{y_{kw}^{(D)}/n_k^{(D)} + y_{kw}^{(R)}/n_k^{(R)}} n_{kw}. \tag{4}$$

In the case of two ''documents,'' as is the case here, this is nearly identical to the difference of proportions measure. In this example, they correlate at over $+0.998$. So, WordScores demonstrates the same failure to account for sampling variation and the same overweighting of high-frequency words. Lowe (2008) (in this volume) gives much greater detail on the workings of the WordScores procedure and how it might be given a probabilistic footing.

3.3 *Model-Based Approaches*

In our preferred model-based approaches, we model the choice of word as a function of party, $P(w|p)$. We begin by discussing a model for the full collection of documents and then show how this can be used as a starting point and baseline for subgroup-specific models. In general, our strategy is to first model word usage in the full collection of documents and to then investigate how subgroup-specific word usage diverges from that in the full collection of documents.

3.3.1 The likelihood

Consider the following model. We will start without subscripts and consider the counts in the entire corpus, $\mathbf{y}$:

$$\mathbf{y} \sim \text{Multinomial}(n, \boldsymbol{\pi}), \tag{5}$$

where $n = \sum_{w=1}^{W} y_w$ and $\boldsymbol{\pi}$ is a W-vector of multinomial probabilities. Since $\boldsymbol{\pi}$ is a vector of multinomial probabilities, it is constrained to be in the $(W-1)$–dimensional simplex. In some variants below, we reparameterize and use the (unbounded) log odds transformation

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

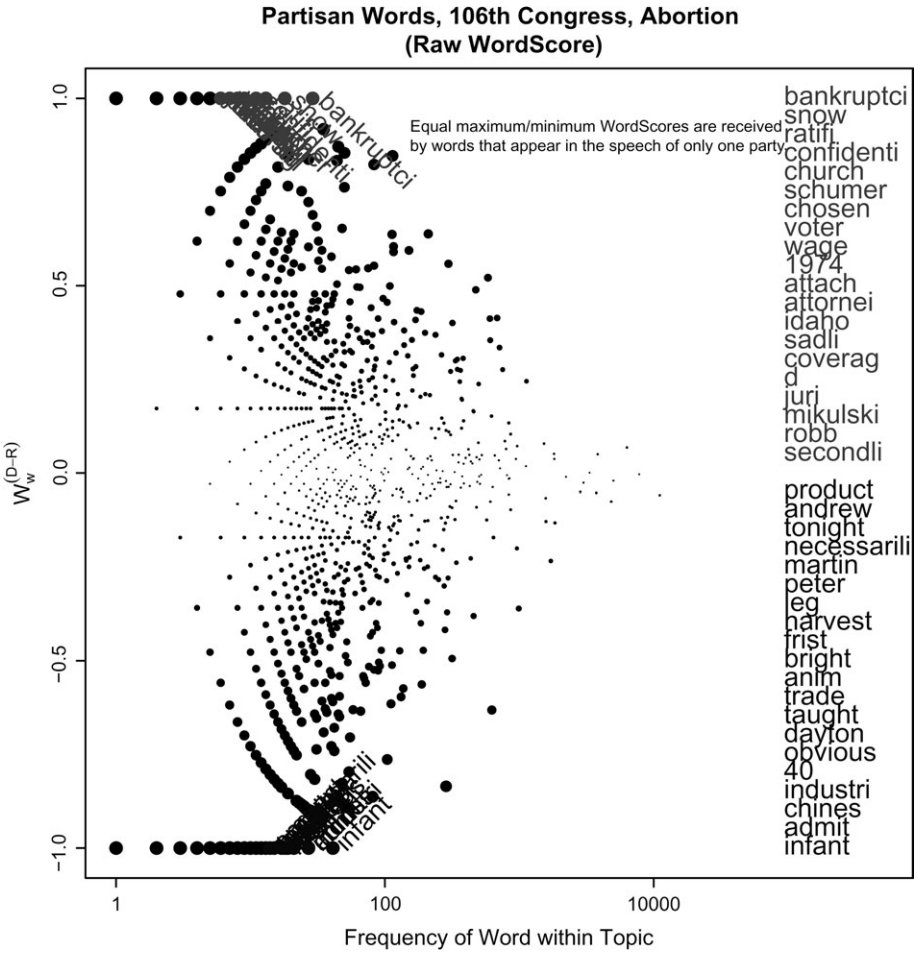

Fig. 3 Feature evaluation using W_{kw}. Plot size is proportional to evaluation weight, $|W_{kw}|$. The top 20 Democratic and Republican words are labeled and listed in rank order to the right.

$$\beta_w = \log(\pi_w) - \log(\pi_1) w = 1, \ldots, W. \tag{6}$$

and work with $\boldsymbol{\beta}$ instead of $\boldsymbol{\pi}$. The inverse transformation

$$\pi_w = \frac{\exp(\beta_w)}{\sum_{j=1}^{W} \exp(\beta_j)}. \tag{7}$$

allows us to transform $\boldsymbol{\beta}$ estimates back to estimates for $\boldsymbol{\pi}$. The likelihood and log-likelihood functions are:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{w=1}^{W} \left(\frac{\exp(\boldsymbol{\beta}_w)}{\sum_{j=1}^{W} \exp\left(\boldsymbol{\beta}_j\right)} \right)^{y_w}, \tag{8}$$

and

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

$$\ell(\boldsymbol{\beta}|\mathbf{y}) = \sum_{w=1}^{W} y_w \log\left(\frac{\exp(\boldsymbol{\beta}_w)}{\sum_{j=1}^{W}\exp\left(\boldsymbol{\beta}_j\right)}\right). \tag{9}$$

Within any topic, k, the model to this point goes through with addition of subscripts:

$$\mathbf{y}_k \sim \text{Multinomial}(n_k, \boldsymbol{\pi}_k), \tag{10}$$

with parameters of interest, β_{kw}, and log-likelihood, $\ell(\boldsymbol{\beta}_k|\mathbf{y}_k)$, defined analogously.

Further, within any group-topic partition, indexed by i and k, we superscript for group to model:

$$\mathbf{y}_k^{(i)} \sim \text{Multinomial}\left(n_k^{(i)}, \boldsymbol{\pi}_k^{(i)}\right), \tag{11}$$

with parameters of interest, $\beta_{kw}^{(i)}$, and log-likelihood, $\ell(\boldsymbol{\beta}_k^{(i)}|\mathbf{y}_k^{(i)})$, defined analogously.

If we wish to proceed directly to ML estimation, the lack of covariates results in an immediately available analytical solution for the MLE of $\beta_{kw}^{(i)}$. We calculate

$$\hat{\boldsymbol{\pi}}^{\text{MLE}} = \mathbf{f} = \mathbf{y} \cdot (1/n), \tag{12}$$

and $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ follows after transforming.

3.3.2 Prior

The simplest Bayesian model proceeds by specifying the prior using the conjugate for the multinomial distribution, the Dirichlet:

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{13}$$

where $\boldsymbol{\alpha}$ is a W-vector, $\alpha_w > 0$, with a very clean interpretation in terms of "prior sample size." That is, use of any particular Dirichlet prior defined by $\boldsymbol{\alpha}$ affects the posterior exactly as if we had observed in the data an additional $\alpha_w - 1$ instances of word w. This can be arbitrarily uninformative, for example, $\alpha_w = 0.01$ for all w. Again, we can carry this model through to topics and topic-group partitions with appropriate sub- and superscripting.

3.3.3 Estimation

Due to the conjugacy, the full Bayesian estimate using the Dirichlet prior is also analytically available in analogous form:

$$\hat{\boldsymbol{\pi}} = \left(\mathbf{y} + \boldsymbol{\alpha}\right) \cdot 1/\left(n + \alpha_0\right). \tag{14}$$

where $\alpha_0 = \sum_{w=1}^{W} \alpha_w$.

Again, all this goes directly through to partitions with appropriate subscripts if desired.

3.3.4 Feature evaluation

What we have to this point is sufficient to suggest the first approach to feature evaluation. Denote the odds (now with probabilistic meaning) of word w, relative to all others, as

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

$\Omega_w = \pi_w/(1-\pi_w)$, again with additional sub- and superscripts for specific partitions. Since the Ω_w are functions of the π_w, estimates of these follow directly from the $\hat{\pi}_w$.

Within any one topic, k, we are interested in how the usage of a word by group i differs from usage of the word in the topic by all groups, which we can capture with the log-odds-ratio, which we will now define as $\delta_w^{(i)} = \log(\Omega_w^{(i)}/\Omega_w)$. The point estimate for this is

$$\hat{\delta}_{kw}^{(i)} = \log\left[\frac{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)}{\left(n_k^{(i)} + \alpha_{k0}^{(i)} - y_{kw}^{(i)} - \alpha_{kw}^{(i)}\right)}\right] - \log\left[\frac{(y_{kw} + \alpha_{kw})}{(n_k + \alpha_{k0} - y_{kw} - \alpha_{kw})}\right]. \quad (15)$$

In certain cases, we may be more interested in the comparison of two specific groups. This is the case in our running example, where we will have exactly two groups, Democrats and Republicans. The usage difference is then captured by the log-odds-ratio between the two groups, $\delta_w^{(i-j)}$, which is estimated by

$$\hat{\delta}_{kw}^{(i-j)} = \log\left[\frac{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)}{\left(n_k^{(i)} + \alpha_{k0}^{(i)} - y_{kw}^{(i)} - \alpha_{kw}^{(i)}\right)}\right] - \log\left[\frac{\left(y_{kw}^{(j)} + \alpha_{kw}^{(j)}\right)}{\left(n_k^{(j)} + \alpha_{k0}^{(j)} - y_{kw}^{(j)} - \alpha_{kw}^{(j)}\right)}\right]. \quad (16)$$

Without the prior, this is of course simply the observed log-odds-ratio. This would emerge from viewing word counts as conventional categorical data in a contingency table or a logit. For each word, imagine a 2×2 contingency table, with the cells including the counts, for each of the two groups, of word w and of all other words. Or, we can specify a logit of the binary choice, word w versus any other word, with our party group indicator the only regressor. With more than two groups, the same information, with slightly more manipulation (via risk ratios), can be recovered from a multinomial logit or an appropriately constrained Poisson regression (Agresti 2002). With the prior, this is a relabeling of the smoothed log-odds-ratio discussed before.

So, if we apply the measure, with equivalent prior, we get results identical to those shown in Figure 2. This has the same problems, with the dominant words still the same list of obscurities. The problem is clearly that the estimates for infrequently spoken words have higher variance than frequently spoken ones.

We can now exploit the first advantage of having specified a model. Under the given model, the variance of these estimates is approximately:

$$\sigma^2\left(\hat{\delta}_{kw}^{(i)}\right) \approx \frac{1}{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)} + \frac{1}{\left(n_k^{(i)} + \alpha_{k0}^{(i)} - y_{kw}^{(i)} - \alpha_{kw}^{(i)}\right)} + \frac{1}{(y_{kw} + \alpha_{kw})} + \frac{1}{(n_k + \alpha_{k0} - y_{kw} - \alpha_{kw})}, \quad (17)$$

$$\approx \frac{1}{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)} + \frac{1}{(y_{kw} + \alpha_{kw})}, \quad (18)$$

and

$$\sigma^2\left(\hat{\delta}_{kw}^{(i-j)}\right) \approx \frac{1}{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)} + \frac{1}{\left(n_k^{(i)} + \alpha_{k0}^{(i)} - y_{kw}^{(i)} - \alpha_{kw}^{(i)}\right)} + \frac{1}{\left(y_{kw}^{(j)} + \alpha_{kw}^{(j)}\right)} + \frac{1}{\left(n_k^{(j)} + \alpha_{k0}^{(j)} - y_{kw}^{(j)} - \alpha_{kw}^{(j)}\right)}, \quad (19)$$

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

$$\approx \frac{1}{\left(y_{kw}^{(i)} + \alpha_{kw}^{(i)}\right)} + \frac{1}{\left(y_{kw}^{(j)} + \alpha_{kw}^{(j)}\right)}. \tag{20}$$

Where the approximations in Equations 17 and 19 assume $y_{kw}^{(i)} \gg \alpha_{kw}^{(i)}$, $y_{kw} \gg \alpha_{kw}$ and ignore covariance terms that will typically be close to 0 while Equations 18 and 20 additionally assume that $n_k^{(i)} \gg y_{kw}^{(i)}$ and $n_k \gg y_{kw}$. The approximations are unnecessary but reasonable for documents of moderate size (at 1000 words only the fourth decimal place is affected) and help clarify the variance equation. Variance is based on the absolute frequency of a word in all, or both, documents of interest, and its implied absolute frequency in the associated priors.

3.4 *Accounting for Variance*

Now we can evaluate features not just by their point estimates but also by our certainty about those estimates. Specifically, we will use as the evaluation measure the z-scores of the log-odds-ratios, which we denote with ζ:

$$\hat{\zeta}_{kw}^{(i)} = \hat{\delta}_{kw}^{(i)} / \sqrt{\sigma^2\left(\hat{\delta}_{kw}^{(i)}\right)}, \tag{21}$$

and

$$\hat{\zeta}_{kw}^{(i-j)} = \hat{\delta}_{kw}^{(i-j)} / \sqrt{\sigma^2\left(\hat{\delta}_{kw}^{(i-j)}\right)}. \tag{22}$$

Figure 4 shows the feature weightings based on the $\hat{\zeta}_{kw}^{(D-R)}$ for our running example. The prominent features now much more clearly capture the core partisan differences. Republican word choice reflects framing the debate from the point of view of the *baby/child* and emphasizing the details of the *partial birth abortion procedure*. In contrast, Democrats framed their speech from the point of view of the *woman/women* and *her/their right* to *choose*.

There are still problems here that can be observed in Figure 4. Although function words no longer dominate the lists as with several previous techniques, some still appear to be too prominent in their implied partisanship, among them *the*, *of*, *be*, *not*, and *my*. A related problem is that, although there is variation, every word in the vocabulary receives nonzero weight, which could lead to overfitting if these estimates are used in some subsequent model. We could perhaps use some threshold value of $\hat{\zeta}_{kw}^{(D-R)}$ as a selection mechanism (Figure 4 demonstrates with the familiar cutoff at 1.96.)

A nearly identical selection mechanism, also following from binomial sampling foundations, can be developed by subjecting each word to a χ^2 test and selecting those above some threshold. Coverage is nearly identical to the z-test. Or words can be ranked by p-value of χ^2 test, again nearly identically to ranking by p-value of z-tests. This was the approach used by Gentzkow and Shapiro (2006). There is no implied directionality (e.g., Democratic vs. Republican), however, and the χ^2 value itself does not have any obvious interpretation as an evaluation weight.

3.5 *Shrinkage and Regularization*

The problems we have seen to this point are typical of many contemporary data-rich problems, emerging not just from computational linguistics but from other similarly scaled or ''ill-posed'' machine and statistical learning problems in areas like image processing and gene expression. In these fields, attempts to avoid overfitting are referred to as *regularization*. A related concept, and more familiar in political science, is Bayesian *shrinkage*. There are several approaches, but the notion in common is to put a strong conservative prior on

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

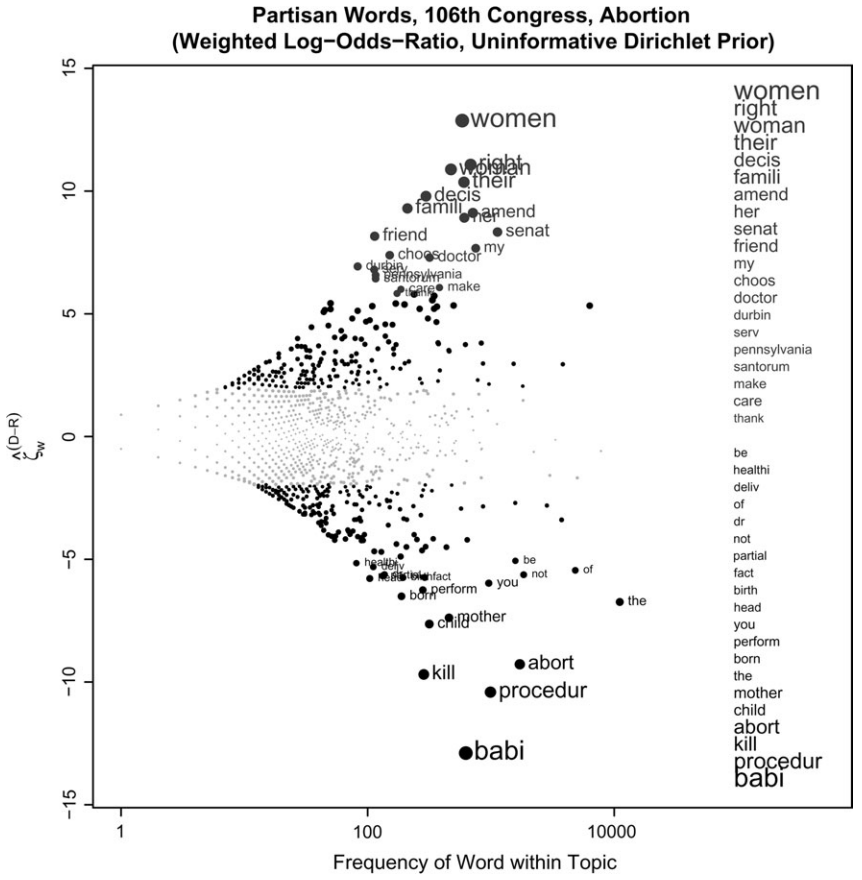

Fig. 4 Feature evaluation and selection using $\hat{\zeta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $\left|\hat{\zeta}_{kw}^{(D-R)}\right|$; those with $\left|\hat{\zeta}_{kw}^{(D-R)}\right|<1.96$ are gray. The top 20 Democratic and Republican words are labeled and listed in rank order to the right.

the model. We bias the model toward the conclusion of no partisan differences, requiring the data to speak very loudly if such a difference is to be declared.

In this section, we discuss two approaches. In the first, we use the same model as above, but put considerably more information into the Dirichlet prior. In the second, we use a different (Laplace) functional form for the prior distribution.

3.5.1 Informative Dirichlet prior

One approach is to use more of what we know about the expected distribution of words. We can do this by specifying a prior proportional to the expected distribution of features in a random text. That is, we know *the* is used much more often than *nuclear*, and our prior can reflect that information. In our running example, we can use the observed proportion of words in the vocabulary in the context of Senate speech, but across multiple Senate topics.[18] That is,

[18] Although this is technically not a legitimate subjective prior because the data are being used twice, nearly all the prior information is coming from data that are not used in the analysis. Qualitatively similar empirical Bayes results could be obtained by basing the prior on speeches on all topics other than the topic in question or, for that matter, on general word frequency information from other sources altogether.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

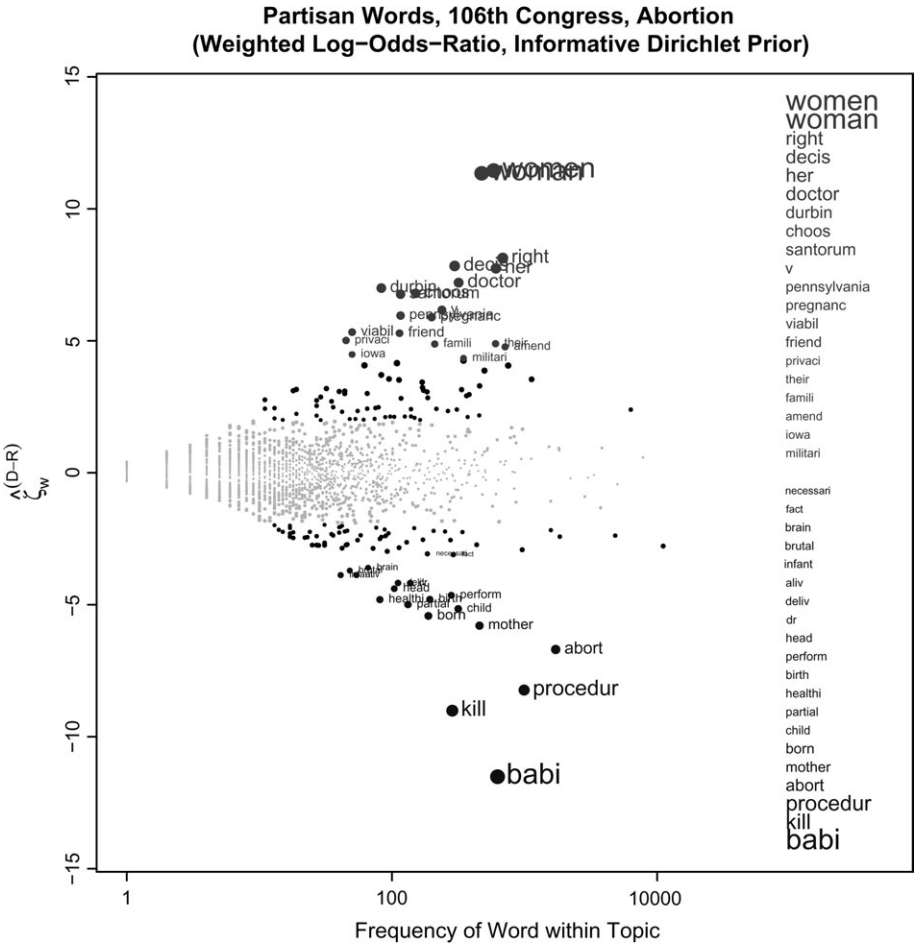

Fig. 5 Feature evaluation and selection based on $\hat{\zeta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $\hat{\zeta}_{kw}^{(D-R)}$; those with $\left|\hat{\zeta}_{kw}^{(D-R)}\right|<1.96$ are gray. The top 20 Democratic and Republican words are labeled and listed in rank order to the right.

$$\alpha_{kw}^{(i)} = \alpha_{k0}^{(i)} \hat{\boldsymbol{\pi}}^{\text{MLE}} = \mathbf{y} \cdot \frac{\alpha_0}{n} \tag{23}$$

where $\alpha_{k0}^{(i)}$ determines the implied amount of information in the prior. This prior shrinks the $\pi_{kw}^{(i)}$ and $\Omega_{kw}^{(i)}$ to the global values, and shrinks the feature evaluation measures, the $\zeta_{kw}^{(i)}$ and the $\zeta_{kw}^{(i-j)}$, toward zero. The greater $\alpha_{k0}^{(i)}$ is, the more shrinkage we will have.

Figure 5 illustrates results from the running example using this approach. We set α_0 to imply a ''prior sample'' of 500 words per party every day, roughly the average number of words per day used per party on each topic in the data set.

As is shown in Figure 5, this has a desirable effect on the function-word problem noted above. For example, the Republican top 20 list has shuffled, with *the*, *you*, *not*, *of*, and *be* being replaced by the considerably more evocative *aliv*, *infant*, *brutal*, *brain*, and *necessari*.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

We also see an apparent improvement in selection, with fewer words now exceeding the same ζ threshold.

The amount of shrinkage is determined by the scale parameter in the Dirichlet prior relative to the absolute frequency of speech in the topic. The prior used in this example has almost no effect on estimates in topics with much more speech and strong partisan differences (like defense, post-Iraq), and overwhelms estimates—as it should—in topics with very little speech or no partisan differences (like symbolic constituent tributes).

Relative to the technique in the next section, this simple approach has several advantages. First, because of the analytical solution, it requires very little computation. Second, this low computational overhead makes it relatively straightforward to develop more flexible variants, such as a dynamic version we discuss below. Third, the fact that these estimates could be recovered from appropriate standard Bayesian generalized linear models means the approach is easily extended beyond the setting of a strict partitioning of the document space into discrete groups. Multiple observable characteristics (party *and* gender *and* geographic district) can be modeled simultaneously, subject to the identification constraints typical of a regression problem.

This approach still suffers as a technique for feature selection: it is unclear where one draws the line between words that qualitatively matter and words that qualitatively do not. So, there is still the problem for the qualitative analyst of interpreting a long (weighted/ordered) list and the problem for the quantitative analyst of possibly overfitting if these are used in subsequent modeling.

3.5.2 Laplace prior

We address this here by using an alternative functional form for the prior, one which shrinks most contrasts between estimates not toward zero, but *to* zero. In this approach, we specify a prior for $\boldsymbol{\beta}_k^{(i)}$. Specifically, we assume

$$\beta_{kw}^{(i)} \overset{ind.}{\sim} \text{Laplace}(\hat{\beta}_{kw}^{\text{MLE}}, \gamma), \quad w = 2, \ldots, W. \tag{24}$$

Or, even more precisely,

$$p(\boldsymbol{\beta}_k^{(i)}|\gamma) = \prod_{w=2}^{W} \frac{\gamma}{2} \exp\Big(-\gamma \Big| \beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\text{MLE}} \Big| \Big). \tag{25}$$

Here β_k^{MLE} is the maximum likelihood estimate from an analysis of the pooled $\mathbf{y}$ which serves as the ''prior'' mode for $\beta_k^{(i)}$ and $\gamma > 0$ serves as an inverse scale parameter.

It is fairly well known that such a prior is equivalent to an L1 regularization penalty (Williams 1995). In practice, such a prior (or regularization penalty) causes many, if not most, of the elements of the posterior mode of $\boldsymbol{\beta}_k^{(i)}$ to be *exactly* equal to the prior mode $\hat{\boldsymbol{\beta}}_k^{\text{MLE}}$.

To finish the prior specification, we consider two types of hyper-priors for γ. The first is that γ follows an exponential distribution.

$$p_E(\gamma) = a \, \exp(-a\gamma). \tag{26}$$

The second is the improper Jeffreys prior:

$$p_J(\gamma) \propto 1/\gamma. \tag{27}$$

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

It is well known that this corresponds to a uniform prior for $\log(\gamma)$.

Under the exponential hyper-prior for γ, the full posterior becomes:

$$p(\boldsymbol{\beta}_k^{(i)}, \gamma | y_k^{(i)}) \propto p(\mathbf{y}_k^{(i)} | \boldsymbol{\beta}_k^{(i)}) p(\boldsymbol{\beta}_k^{(i)} | \gamma) p_E(\gamma)$$

$$\propto \left\{ \prod_{w=1}^{W} \left(\frac{\exp(\beta_{kw}^{(i)})}{\sum_{i=1}^{W} \exp(\beta_{ki}^{(i)})} \right)^{y_{kw}} \right\} \left\{ \prod_{w=2}^{W} \frac{\gamma}{2} \exp\left(-\gamma \left| \beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\text{MLE}} \right| \right) \right\} a \exp(-a\gamma) \qquad (28)$$

and under the Jeffreys hyper-prior for γ, the full posterior becomes:

$$p(\boldsymbol{\beta}_k^{(i)}, \gamma | y_k^{(i)}) \propto p(\mathbf{y}_k^{(i)} | \boldsymbol{\beta}_k^{(i)}) p(\boldsymbol{\beta}_k^{(i)} | \gamma) p_J(\gamma)$$

$$\propto \left\{ \prod_{w=1}^{W} \left(\frac{\exp(\beta_{kw}^{(i)})}{\sum_{i=1}^{W} \exp\left(\beta_{ki}^{(i)} \right)} \right)^{y_{kw}} \right\} \left\{ \prod_{i=2}^{W} \frac{\gamma}{2} \exp\left(-\gamma \left| \beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\text{MLE}} \right| \right) \right\} 1/\gamma. \qquad (29)$$

The sharp mode of the Laplace prior means that words whose partisanship is not clear will receive partisan contrasts that are exactly zero. This is useful for feature selection. The nondifferentiability at the mode does considerably increase the difficulty and computational overhead associated with estimation. This procedure is discussed in the Appendix.

With the Laplace model, we do not downweight (further) by variance, but look directly at the $\hat{\delta}_{kw}^{(D-R)}$. Figure 6 shows the results for our running example. Note first that Figure 6 is largely blank, with the vast majority of words having estimates shrunk to zero. The estimates are dominated by a few words: *kill*, *babi*, *procedur*, *child*, *mother*, and *abort* for Republicans; *women*, *woman*, *decis*, *famili*, *friend*, and *right* for Democrats. A higher smoothing parameter would result in even fewer nonzero features; a smaller one would result in more.[19] As with the above, the effect of any particular smoothing parameter will be affected by the relative volume of speech and the polarization of word use. This is ideal for an application where we value parsimony in the output.

This has a few drawbacks. Most notably, the nondifferentiability of the Laplace prior dramatically increases computational complexity even for point estimates. Extending the model—dynamically, for example—is also difficult. These properties make this approach less than ideal for feature evaluation and applications where the estimates feed forward to another analysis. With a sufficiently high smoothing parameter, however, this is an excellent choice for feature selection.

4 Applications of Feature Selection and Evaluation

In this section we explore the broader usefulness of our proposed models for analyzing substantive questions relating to issue framing, representation, polarization, and dimensionality. Although the merging of the text-as-data approach with shrinkage methods could fill several articles for each of these topics, our goal is to illustrate how future research might utilize these methods to answer substantive questions in political science. In turn, we show how our

[19] For these calculations, we define the prior with $a = 100$.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

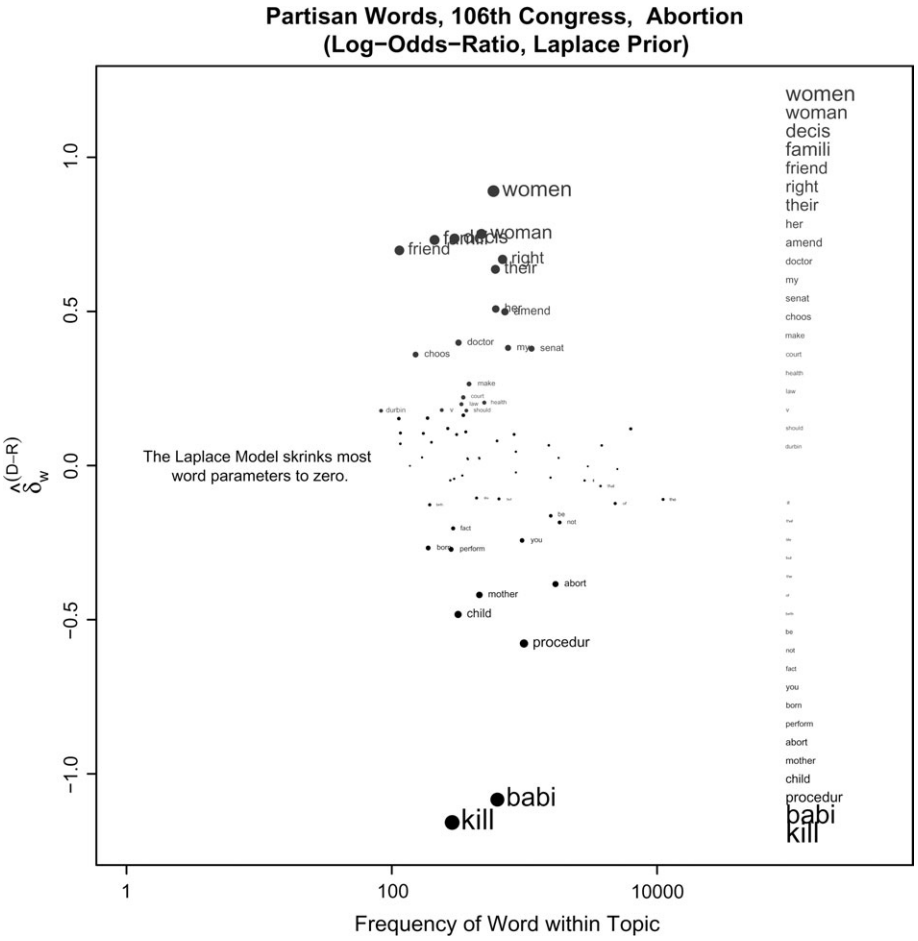

Fig. 6 Feature evaluation and selection using $\hat{\delta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $\hat{\delta}_{kw}^{(D-R)}$. The top 20 Democratic and Republican words are labeled and listed in rank order to the right.

methods can be easily expanded to analyze different issues, extra-partisan grouping, change over time, and political divisions across more than two categories.

We begin by expanding the issues within which we search for partisan frames. Specifically, we are interested in whether our methods find frames where they are likely to be (as in the abortion topic) and do not find partisan frames where there are likely to few. Next, we show that distinctions in word use can be identified not only across party but across gender (within party) also, as has been argued in the representation literature (Williams 1998; Diaz 2005). This is followed by examples that illustrate the changes in partisan word usage over time. Finally, we explore how feature selection and weighting across geographic units can be used to understand the underlying dimensionality of politics.

4.1 *Partisan framing*

Chong and Druckman (2007, 106) note that the analysis of ''frames in communication'' is an endeavor that ''has become a virtual cottage industry.'' In such analyses, ''an initial set of frames for an issue is identified inductively to create a coding scheme'' (Chong and

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

Druckman 2007, 107). The result of this initial stage is something like the 65-argument coding scheme developed by Baumgartner et al. (2008) to exhaustively capture the pro- and antiframes used in debate over the death penality (Baumgartner et al. 2008, 103–12, 243–51). These then serve as the basis for a content analysis of relevant texts, like newspaper articles, to evaluate variations in frame across source and across time. The feature selection techniques we describe here can be useful for such a purpose.

Consider the topic of taxes. Figure 7 shows the results of applying the Laplace model, again in the 106th Congress. Note the more compressed *y*-axis, relative to the abortion analysis of Figure 6, indicating that during this period taxation was a considerably more partisan topic. The prominent features identify the most prominent partisan frames. Republicans were pushing for elimination of the *death tax*, lowering of *capital gains taxes*, elimination of the *marriage penalty*, and *flatter brackets*; Democrats advocated for a tax credit for *prescription drug coverage*, for *college*, and against lowering the *estate tax* or giving *breaks* to the *wealthiest*. Indeed, *death tax* and *marriage penalty* are two of the more famous and successful examples of party-coordinated language choice.

Such features could then be used to evaluate something like media use of taxation frames. A full analysis is well beyond our scope here, but a brief example is suggestive of the potential here. We conducted Lexis-Nexis searches of two newspapers, *The New York Times* and *The Washington Times*, and two broadcast news sources, *National Public Radio* and *Fox News*, widely thought to be, relatively speaking, at different partisan extremes. We counted articles or transcripts that used variants of the word *tax* along with variants of Democratic features (*estate*, *prescription*, *wage*, *wealthiest*) and four prominent Republican frames (*death*, *bracket*, *flat*, *penalty*), for 1999–2000. Sticking to the familiar metric of odds ratios, the odds of *The New York Times* using these Democratic over Republican frames are roughly 20% higher than those of the *Washington Times*. Similarly, the odds of NPR using Democratic over Republican frames are roughly 63% higher than Fox News.[20] This is a toy example, of course, but suggests possibilities for a variety of more interesting applications across source, time, or political topic.

More central to our motivations for the Laplace model are the advantages of regularization. In this context, one notable advantage is that the model will not suggest differences where there are none. For example, our Senate analysis finds a large number of ''sports'' speeches congratulating a home state team. We would not expect these to have partisan structure. If we apply exactly the same Laplace model as above to these speeches, the model finds only a tiny handful of words with very small partisan content.[21] We would correctly conclude that there is virtually no partisan content to such speech.

4.2 *Women's Representation*

Not only is it useful to estimate our model on different topics but we can also analyze extra-partisan distinctions. For example, a long-standing question in the representation literature concerns the relative roles of descriptive representation, symbolic representation, and substantive representation instantiated by diversity in the demographics of

[20]The counts of *estate*, *prescription*, *wage*, *wealthiest*, *death*, *bracket*, *flat*, *penalty*, respectively, are (2078, 620, 1159, 208, 1280, 164, 431, 743) for *The New York Times*, (1032, 376, 692, 84, 731, 189, 399, 720) for *The Washington Times*, (172, 214, 224, 68, 209, 47, 68, 174) for National Public Radio, and (167, 362, 238, 54, 450, 55, 197, 281) for Fox News.

[21]Specifically, *football* is very slightly Republican, whereas *great* and *women* are slightly Democratic. The graphic for this is omitted because it is largely blank, as expected, but is available in the web appendix.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

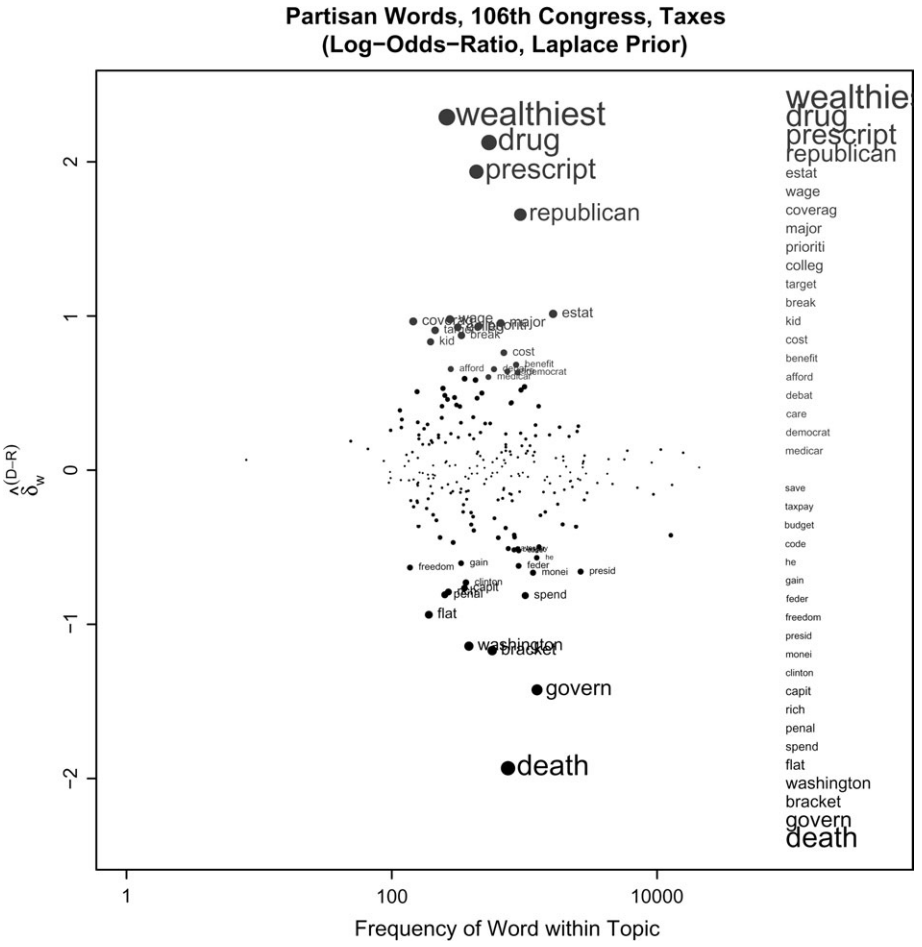

Fig. 7 Feature evaluation and selection using $\hat{\delta}_{kw}^{(D-R)}$. Plot size is proportional to evaluation weight, $\hat{\delta}_{kw}^{(D-R)}$. The top 20 Democratic and Republican words are labeled and listed in rank order to the right.

a legislature. One of the more prominent is the concern over whether women legislators behave differently than male legislators and whether they offer a fundamentally different form of representation by speaking ''in a different voice'' (Gilligan 1993; Williams 1998; Diaz 2005).

These models are natural for such questions. Not only can we look at the relative differences in language use between men and women, as we did with party, but we can also do so while controlling for party. This is important because of a general gender gap across the two parties both in core values (Rokeach 1973) and in framing (Lakoff 2004). Democrats are both more likely than Republicans to address ''women's issues,'' and also have more women members serving in the Senate than do the Republicans.

Figure 8 shows a simple example of gender differences, within Democrats, in language use for our running example of abortion in the 106th Senate, using the informative Dirichlet model. The most prominent and obvious difference between female and male Senators is the distinction between discussing the adult involved in an abortion as a *woman*, or as a *mother*. We keep it simple here, but more elaborate types of cross-nesting of categories,

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

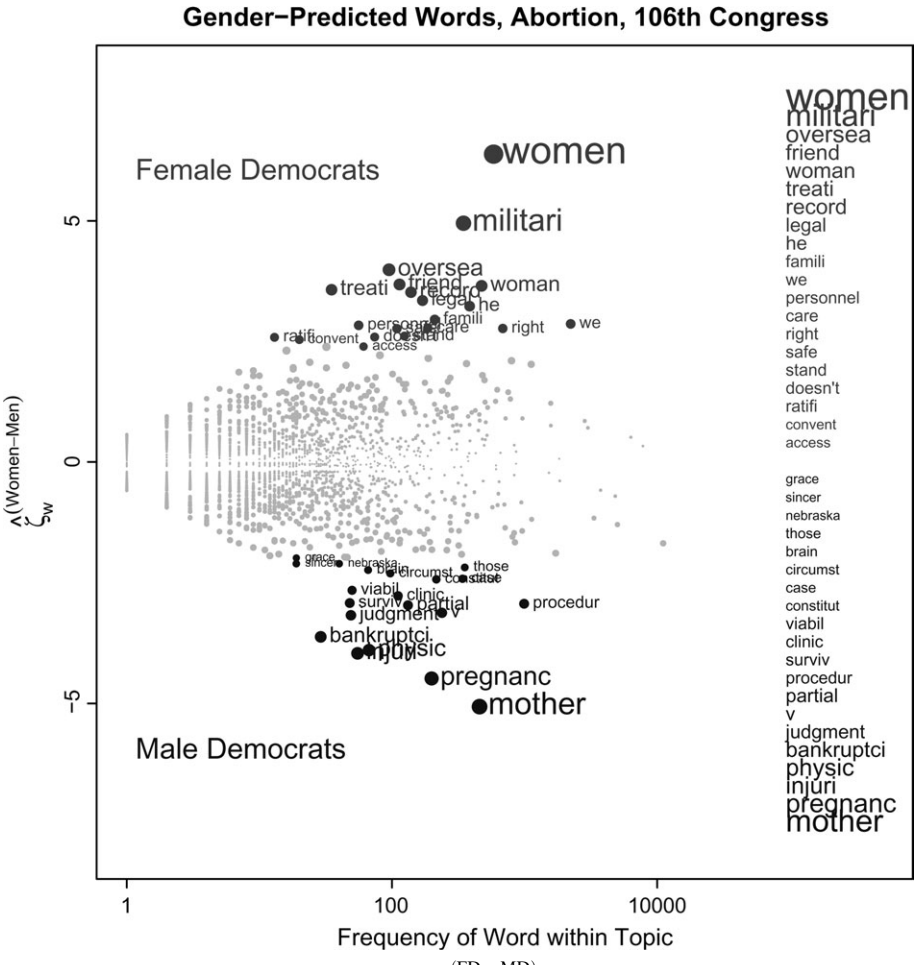

Fig. 8 Feature evaluation and selection using $\hat{\zeta}_{kw}^{(\mathrm{FD}-\mathrm{MD})}$. Plot size is proportional to evaluation weight, $\hat{\zeta}_{kw}^{(\mathrm{FD}-\mathrm{MD})}$. Top 20 Female (Democrat) and Male (Democrat) words are labeled and listed in rank order.

as well as continuous covariates (like urbanity of state), are simple extensions of the Dirichlet model.

4.3 *Dynamics: Issue Evolutions and Semantic Realignment*

We can also extend our analysis across time to analyze dynamics across partisan (or extra-partisan) word use. One of the fundamental questions in American politics has concerned the existence, content, timing, and mechanisms of parties shifting against the ideological background (Carmines and Stimson 1990; Poole and Rosenthal 1997). A large-scale example clearly matching the criteria for a ''realignment'' or an ''issue evolution'' is the reversal of the parties on race and civil rights issues over the course of the 20th century; a small-scale example, which may or may not meet the criteria, is the polarization over time of the parties on the issue of abortion.

Empirical investigation of such historical phenomena is difficult. One approach, exemplified by Mucciaroni and Quirk's *Deliberative Choices* (2006) is to examine in great detail

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

the traceable behaviors—in this case, Congressional speeches, testimony, etc.—for a particular policy debate held over a relatively short time scale. A second approach, exemplified by Carmine and Stimson's *Issue Evolution* (1990) and similar work that followed its lead (Adams 1997), looks for changes across more abstract traceable behaviors—for example, rollcall votes in Congress, survey responses—in a single issue area over a long time scale: race for Carmines and Stimson, abortion and others for later work. A third approach, exemplified by Poole and Rosenthal's *Congress* (1997), looks at a single behavior—Congressional rollcall votes—across all issues and long time scales, looking for abstract changes in the partisan relationship, such as the dimensionality of a model required to explain the behaviors.

Speech offers interesting potential leverage on the problem. We have seen that static snapshots can illuminate the content of a particular partisan conflict. If these techniques can be extended dynamically, we can look directly at how the content of partisan conflict changes.

The analytic solution of the Dirichlet model makes dynamic extensions straightforward. First, the y are extremely noisy on a day-to-day basisso we apply a smoother to the data.[22] Then we calculate *zeta* over a moving time window of the data.

A genuine partisan realignment or issue evolution would be evidenced by massive shifts in the partisanship of language. That is, a necessary condition for an issue evolution would be for some words to switch sides, to flip sign in our measure. This generally occurred (or been agreed to have occurred) only over fairly long time frames. So, over the relatively short time frame of these data, 8 years, we would not expect to find much that would qualify as such. However, we do find many microscale movements that suggest such analyses can prove fruitful.

For example, we find several instances where Republicans ''own'' an issue or frame while in opposition to Clinton are successful in obtaining a policy change during the Bush honeymoon period of early 2001 or in the rally period after 9/11, only to find Democrats ''owning'' the issue when the policy becomes unpopular. Dramatic examples include *debt* when discussing the budget, *oil* when discussing energy policy, and *Iraq* when discussing defense policy.

Figure 9 shows an estimate of the dynamic partisanship[23] of *iraq* on the topic of defense, an example which provides a unique window into the dynamics of partisan conflict. In order to observe shifts in the tectonic plates of partisan conflict, we need both salience (the parties and presumably the public must care) and some exogenous shock or change in the status quo (to motivate change). Iraq supplies both of these, and in turn, we see relative nonpartisanship pre-9/11 shift to a Republican talking point post-9/11 only to see events on the ground and the cost of the war increase Democratic leverage on the issue. This technique may be of particular importance in the future for any analyst interested in tracing whether the Iraq issue settles into a nonpartisan or renewed partisan equilibrium.

Similarly, dynamics can reveal how institutions affect the ''ideological'' justifications utilized by parties. Figure 10 illustrates one example, the dynamic partisanship of the word

[22]Specifically, we smooth the data by first calculating a b-day moving average, m, of word use, and apply an exponential smoother, with smoothing factor A: $m^{(i)}_{kwt} = \sum_{\tau=t-b}^{t} y^{(i)}_{kw\tau}$; $s^{(i)}_{kw(b+1)} = m^{(i)}_{kw(b+1)}$; $s^{(i)}_{kwt} = Am^{(i)}_{kwt} + \left(1-A\right)s^{(i)}_{kw(t-1)}$. From this we calculate the ζ_{kwt} for each day, using the $s^{(i)}_{kwt}$ where we would have used the $y^{(i)}_{kwt}$ previously.

[23]These particular graphs reflect daily estimates based on a $b = 180$ day (6 months) moving average component and an $A = .01$ smoothing factor.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

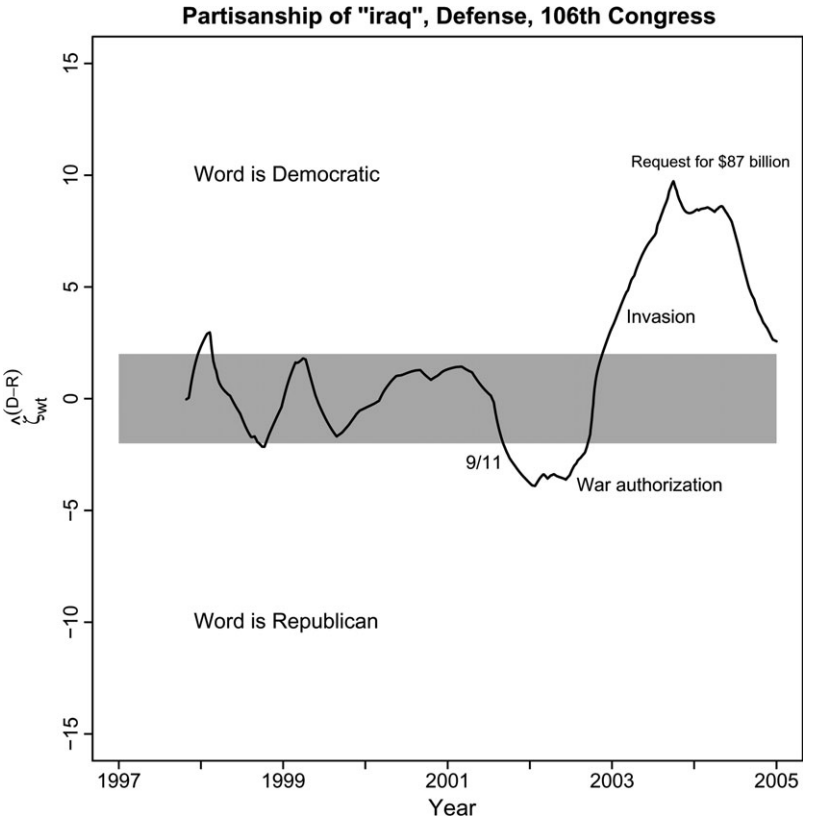

Fig. 9 Dynamic partisanship of ''iraq'' in the context of defense.

vote on the issue of judicial nominations. During the Clinton era, Republicans consistently blocked his judicial nominees. Democrats demanded that votes be held, with increasing urgency as the Clinton presidency neared its end. During the Bush era, Democrats consistently filibustered his judicial nominees, and Republicans discovered a deeply held belief in presidential deference and up-or-down votes. On the theoretical issue of majority rule, where you stand depends on where you sit.[24]

4.4 *Polarization*

Partisan polarization has been a central focus for much recent scholarship American politics.(McCarty et al. 2006; Sinclair 2006; Theriault 2008) The primary measure of polarization at the elite level has come from rollcall analysis, following Poole and Rosenthal (1997). Variation in language use offers an alternative measure, one that can be differentiated across political topics and at a finer dynamic grain.

[24]We can also look at the entire vocabulary over time in an animation. An example for the topic of defense is given here: http://qssi.psu.edu/PartisanDefenseWords.html.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

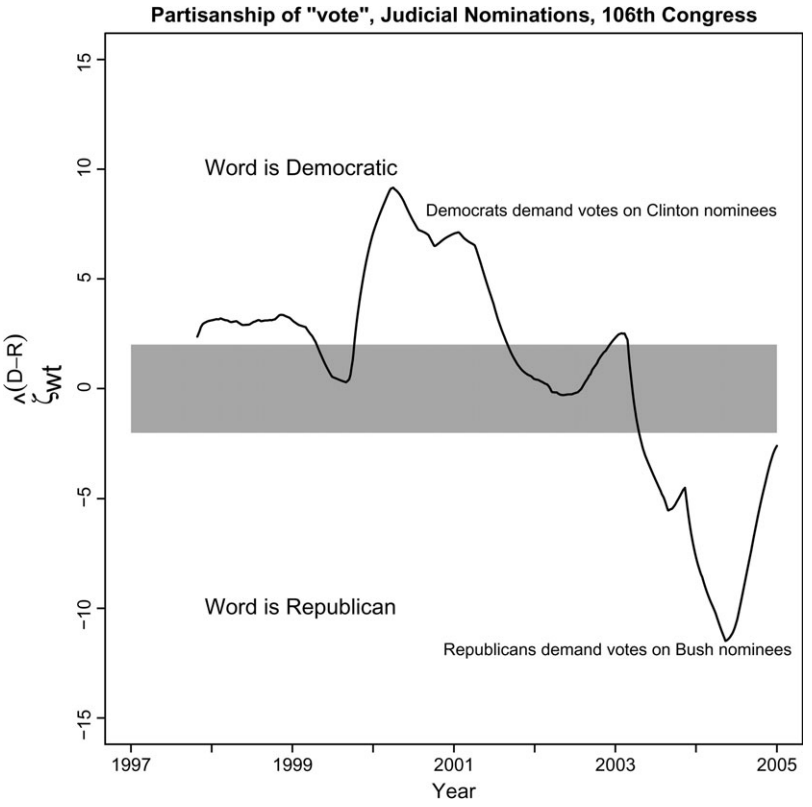

Fig. 10 Dynamic partisanship of ''vote'' in the context of judicial nominations.

A simple measure of polarization is the variance of $\zeta_{kwt}^{(D-R)}$ across all words at any given time, *t*. Figure 11 illustrates this for four topics over the period: budget, defense, education, and sports. Defense is particularly interesting, reaching near zero polarization in the aftermath of 9/11 and dramatically increasing through and after the invasion of Iraq. Note also the buildup and then collapse of polarization around education peaking after the passage of No Child Left Behind, the apparent annual (out of phase) cycles in polarization around the salient topics, and the consistently low partisan polarization of sports.

4.5 *Dimensionality*

Another central concern of the Congressional literature is the dimensionality of the political space in which parties and politicians position themselves. Spatial theorists have argued that underlying preferences must be distributive and multidimensional (Aronson and Ordeshook 1981) and that parties and institutions are reshaped to manage the inherent instability (Black 1958; Riker 1986). Rollcall-based work suggests that a single ideological dimension accounts for virtually all Congressional divisions (Poole 2005). But if bills are structured by a partisan agenda-setter, in the manner of Cox and McCubbins (2005) for example, there will never be enough cross-cutting rollcall votes to detect underlying

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

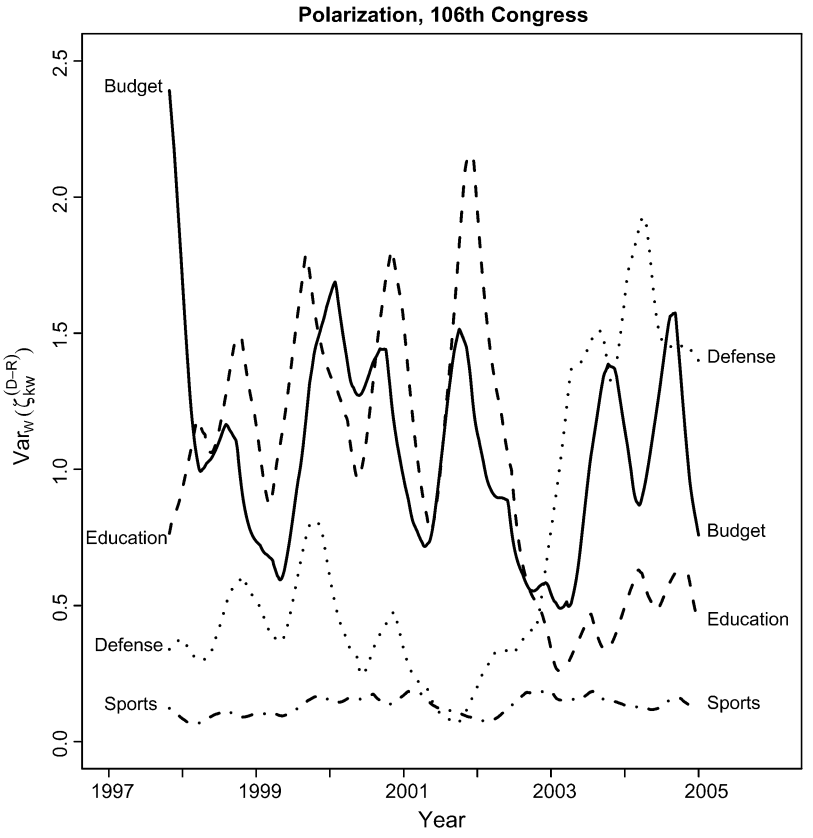

Fig. 11 The dynamics of language polarization.

49-dimensional state versus state preferences. Speech, on the other hand, can reveal such subtleties.

To give one example, Figure 12 shows the state-predicted words on the issue of public lands and water, for four states with a range of likely interests in the subject. The delegations from each state each have some distinctive speech on the issue, with prominent words including the state names, major features (Great Lakes, Finger Lakes), and primary concerns (e.g., invasive species in Michigan, emissions in New York, or hunting in Alabama). There is also clear variation in the intensity of preferences, with great intensity on multiple issues from the Alaska delegation, but limited intensity from the Alabama delegation. This is suggestive that there are measurable distributive (and therefore multidimensional) preference differences across Senators (which are unlikely to be present in rollcall votes once partisan coalitions have been built around particular bills).

5 Discussion and Conclusion

Contemporary representative democratic systems are shaped, perhaps defined, by the relationships among citizens, elected representatives, political parties, and the substance of politics. Creating tools to analyze these relationships, and how they interconnect dynamically over the life of a democracy, is a fundamental challenge to political methodology. One bottleneck has been our measurement and modeling of the relationship between

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

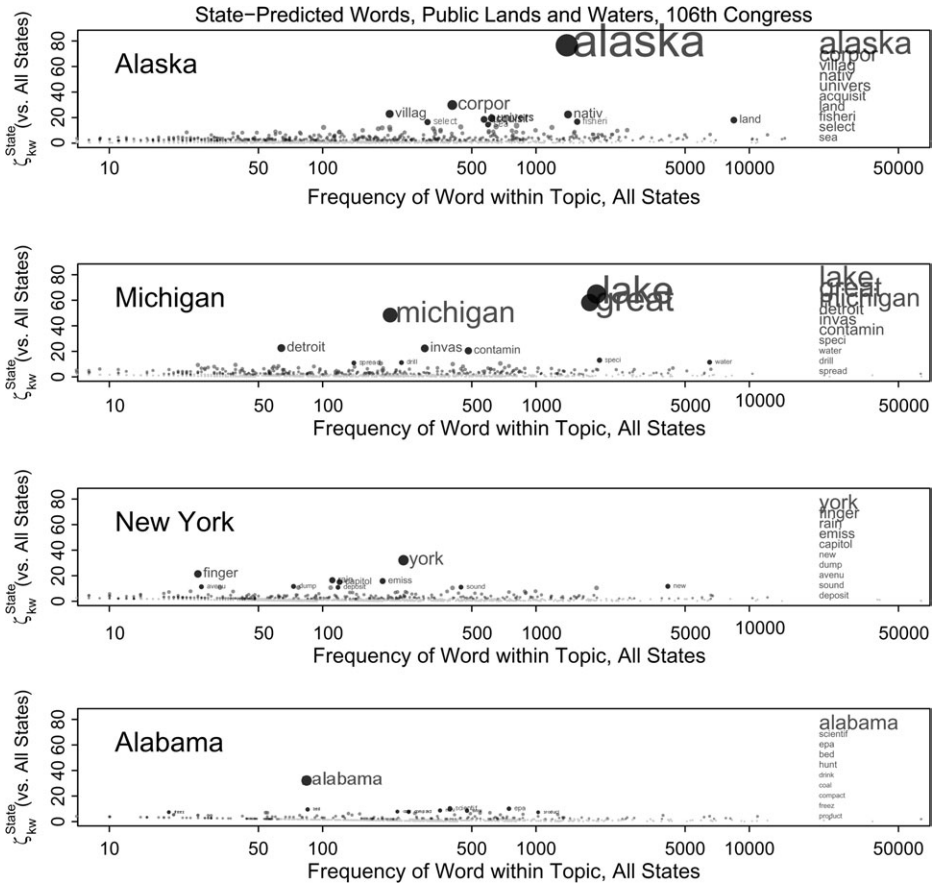

Fig. 12 Feature evaluation and selection using $\hat{\zeta}_{kw}^{(\text{State})}$. Plot size is proportional to evaluation weight, $\hat{\zeta}_{kw}^{(\text{State})}$. Top 10 state-predicted words for each state are labeled and listed in rank order.

political substance and partisan dynamics. Efforts to systematically analyze political conflict, especially in the American context, have previously focused on behaviors that are easily quantified but obscure substance (e.g., votes). More recent attempts to investigating political dynamics that summarize and organize text have fallen into other traps, including inefficiency (e.g., classification techniques), and overfitting idiosyncratic group differences in low- or high-frequency words (e.g., *tf.idf* or difference of proportions). These problems lead researchers to select and weight feature/word lists that have low semantic validity or vary wildly depending on arbitrary decisions.

Here we have highlighted the problems with these existing techniques for feature selection and evaluation as well as presented two potential solutions. By using a model-based approach with Bayesian shrinkage and regularization, we argue, analysts can at least partially side-step previous pitfalls. Most importantly, the word lists and weights produced by our preferred techniques reduce the emphasis placed on words used very frequently across all groups as well as those words that are very infrequently used across both groups. Our methods also do not necessitate that the researcher create an ad hoc filter that erases words from the political vocabulary. Instead, Bayesian shrinkage supplies rules that are consistently applied across the whole of the vocabulary, allowing consistent distinctions across groups to become apparent.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

We are, of course, ultimately concerned with whether these tools are useful for political analysis. Our examples have shown that questions relating to issue framing, representation, polarization, and dimensionality can be explicitly explored with text data through our techniques. We expect that others working within these substantive fields will be able to carry out more detailed and extensive applications with these tools.

The next step is to further refine and craft methods that will be useful for unlocking the exciting potential of political texts. Our model-based regularization approach is one such attempt. Although off-the-shelf algorithms and techniques have proven valuable, we must look under the hood to know if we are validly measuring the concepts we care about. There is much that is theoretically and conceptually unique about the production of language in politics, and there is a need for new methods to be developed and applied accordingly.

Funding

National Science Foundation (grant BCS 05-27513 and BCS 07-14688).

Appendix—Estimation of the Laplace Model

Maximization of the posterior densities in equations 28 and 29 is complicated by the discontinuities in the partial derivatives that are introduced by the Laplace prior. We make use of a version of the algorithm of Shevade and Keerthi (2003) that has been modified to work with a multinomial likelihood. Calculation of the relevant derivatives is tedious but not difficult. In addition to moving from a Bernoulli to a multinomial likelihood, we also extend the work of Shevade and Keerthi (2003) and Cawley and Talbot (2006) to allow an exponential prior for γ.

Marginalizing over γ with a Jeffreys Prior

Using results from Cawley and Talbot (2006), it is easy to show how the value of $\boldsymbol{\beta}_k^{(i)}$ that maximizes

$$p(\boldsymbol{\beta}_k^{(i)}|y_k^{(i)}) \propto \int p(\mathbf{y}_k^{(i)}|\boldsymbol{\beta}_k^{(i)}) p(\boldsymbol{\beta}_k^{(i)}|\gamma) p_J(\gamma) \mathrm{d}\gamma \tag{A1}$$

can be calculated via a modified version of the iterative approach of Shevade and Keerthi (2003) in which, at each iteration, a working version, $\tilde{\gamma}$, of γ is defined to be:

$$\tilde{\gamma} \equiv W / \sum_{w=1}^{W} \left| \beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\mathrm{MLE}} \right|, \tag{A2}$$

and then the maximization over $\beta_{kw}^{(i)}$ is carried out conditional on $\tilde{\gamma}$.

Marginalizing over γ with an Exponential Prior

Using ideas similar to those described in the previous subsection to show how the value of $\boldsymbol{\beta}_k^{(i)}$ that maximizes

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

$$p(\boldsymbol{\beta}_k^{(i)}|y_k^{(i)}) \propto \int p(\mathbf{y}_k^{(i)}|\boldsymbol{\beta}_k^{(i)})p(\boldsymbol{\beta}_k^{(i)}|\gamma)p_E(\gamma)\mathrm{d}\gamma \tag{A3}$$

can be calculated via a modified version of the iterative approach of Shevade and Keerthi (2003) in which, at each iteration, a working version, $\tilde{\gamma}$, of γ is defined to be:

$$\tilde{\gamma} \equiv \frac{(W+1)}{\left(a + \sum_{w=1}^{W} \left|\beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\mathrm{MLE}}\right|\right)} \tag{A4}$$

and then the maximization over $\beta_{kw}^{(i)}$ is carried out conditional on γ.

Maximizing over γ with an Exponential Prior

It is also possible to show that values of $\beta_k^{(i)}$ and γ that maximizes

$$p(\boldsymbol{\beta}_k^{(i)}, \gamma|\mathbf{y}_k^{(i)}) \propto p(\mathbf{y}_k^{(i)}|\boldsymbol{\beta}_k^{(i)})p(\boldsymbol{\beta}_k^{(i)}|\gamma)p_E(\gamma) \tag{A5}$$

can be calculated via a modified version of the iterative approach of Shevade and Keerthi (2003) in which, at each iteration, a local conditional maximizer for γ is:

$$\hat{\gamma} \equiv \frac{W}{\left(a + \sum_{w=1}^{W} \left|\beta_{kw}^{(i)} - \hat{\beta}_{kw}^{\mathrm{MLE}}\right|\right)}, \tag{A6}$$

and then the maximization over $\beta_{kw}^{(i)}$ is carried out conditional on $\hat{\gamma}$.

Note that when $a = 0$, the results from this procedure are equivalent to those from the setup where γ is given a Jeffreys prior and then integrated out of the posterior.

References

Adams, Greg D. 1997. Abortion: Evidence of issue evolution. *American Journal of Political Science* 41(3):718–37.

Agresti, Alan. 2002. *Categorical data analysis*. 2nd ed. Hoboken, NJ: Wiley.

Aizawa, Akiko. 2003. An information-theoretic perspective of tf-idf measures. *Information Processing and Management* 39(1):45–65.

Aronson, Peter H., and Peter C. Ordeshook. 1981. Regulation, redistribution, and public choice. *Public Choice* 37(1):69–100.

Baumgartner, Frank R., Suzanna L. DeBoef, and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge: Cambridge University Press.

Black, Duncan. 1958. *The theory of committees and elections*. Cambridge: Cambridge University Press.

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

Carmines, Edward, and James Stimson. 1990. *Issue evolution*. New York: Princeton University Press.

Cawley, Gavin C., and Nicola L. C. Talbot. 2006. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* 22(19):2348–55.

Chong, Dennis, and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science* 10:103–26.

Cox, Gary W., and Mathew D. McCubbins. 2005. *Setting the agenda: Responsible party government in the US House of Representatives*. Cambridge: Cambridge University Press.

Diaz, Mercedes Mateo. 2005. *Representing women? Female legislators in West European parliaments*. Colchester, UK: ECPR Press Monographs.

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu, and Stefan Kaufmann. 2007. *Language and ideology in congress*. Chicago, IL: Midwestern Political Science Association.

Emily, Senay, Eli H. Newberger, and Rob Waters. 2004. *From boys to men*. New York: Simon and Schuster.

Ericson, Matthew. 2008. The words they used. *The New York Times,* September 4, http://www.nytimes.com/interactive/2008/09/04/us/politics/20080905_WORDS_GRAPHIC.html (accessed November 1, 2008).

Freund, Yoav, and Robert Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55(1):119–39.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

Gentzkow, Matthew, and Jesse M. Shapiro. 2006. *What drives media slant? Evidence from U.S. daily newspapers*. University of ChicagoTechnical report.
Gilligan, Carol. 1993. *In a different voice: Psychological theory and women's development*. Cambridge, MA: Harvard University Press.
Hand, David J. 2006. Classifier technology and the illusion of progress. *Statistical Science* 21(1):1–14.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning*. New York: Springer.
Hiemstra, Djoerd. 2000. A probabilistic justification for using tfxidf term weighting in information retrieval. *International Journal on Digital Libraries* 3:131–9.
Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2007. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology and Politics* 4(4):31–46.
Hopkins, Daniel, and Gary King. 2007. Extracting systematic social science meaning from text. Chicago, IL: Midwestern Political Science Association.
Jurafsky, Daniel, and James H. Martin. 2000. *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.
Kleinfeld, Judith, and Andrew Kleinfeld. 2004. Cowboy nation and American character. *Society* 41(3):43–50.
Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology.* 2nd ed. New York: Sage.
Lakoff, George. 2004. *Don't think of an elephant! Know your values and frame the debate*. White River Junction, VT: Chelsea Green Publishing.
Laver, Michael, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review* 97:311–31.
Lowe, Will. 2008. Understanding Wordscores. *Political Analysis* doi:10.1093/pan/mpn004 (accessed October 4, 2008).
Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *An introduction to information retrieval*. Cambridge: Cambridge University Press.
Many Eyes (IBM). *Tag Cloud Guide*. http://manyeyes.alphaworks.ibm.com/manyeyes/page/Tag_Cloud.html (accessed July 13, 2008).
McCaffrey, Dawn, and Jennifer Keys. 2008. Competitive framing processes in the abortion debate: Polarization-vilification, frame saving, and frame debunking. *Sociological Quarterly* 41(1):41–61.
McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The dance of ideology and unequal riches*. Cambridge, MA: MIT Press.
Monroe, Burt L., Cheryl L. Monroe, Kevin M. Quinn, Dragomir Radev, Michael H. Crespin, Michael P. Colaresi, Jacob Balazar, and Steven P. Abney. 2006. *United States congressional speech corpus*. State College, PA: Pennsylvania State University.
Monroe, Burt L., and Ko Maeda. 2004. Rhetorical ideal point estimation: Mapping legislative speech. *Society for Political Methodology.* Palo Alto: Stanford University.
Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. *An Animated History of Senate Debate on Defense, 1997–2004*. http://qssi.psu.edu/PartisanDefenseWords.html (accessed November 1, 2008).
Mucciaroni, Gary, and Paul J. Quirk. 2006. *Deliberative choice: Debating public policy in congress*. Chicago, IL: University of Chicago Press.
Nunberg, Geoffrey. 2006. *Talking right: How conservatives turned liberalism into a tax-raising, latte-drinking, sushi-eating, volvo-driving, New York Times-Reading, Body-Piercing, Hollywood-Loving, Left-Wing Freak Show.* New York: Public Affairs.
Poole, Keith T. 2005. *Spatial models of parliamentary voting*. New York, Cambridge: Cambridge University Press.
Poole, Keith T., and Howard Rosenthal. 1997. *Congress: A political-economic history of roll-call voting*. Oxford: Oxford University Press.
Porter, Martin F. 2001. The English (Porter2) stemming algorithm. http://snowball.tartarus.org/algorithms/english/stemmer.html (accessed July 3, 2008).
Quinn, Kevin, Burt L. Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev. 2006. *How to analyze political attention with minimal assumptions and costs*. Davis, CA: Society for Political Methodology.
Riker, William H. 1986. *The art of political manipulation*. New Haven, CT: Yale University Press.
Rokeach, Milton. 1973. *The nature of human values*. New York: The Free Press.
Sacerdote, Bruce, and Owen Zidar. 2008. Campaigning in poetry: Is there information conveyed in the candidates' choice of words. Technical report, University of Dartmouth.
Schonhardt-Bailey, Cheryl. 2008. The congressional debate on partial-birth abortion: Constitutional gravitas and moral passion. *British Journal of Political Science* 38(3):383–410.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013

Shevade, S. K., and S. S. Keerthi. 2003. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 19(17):2246–53.

Sinclair, Barbara. 2006. *Party wars: Polarization and the politics of national policy making*. Norman, OK: University of Oklahoma Press.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52(3):705–22.

Theriault, Sean M. 2008. *Party polarization in congress*. Cambridge: Cambridge University Press.

Vapnik, Vladimir. 2001. *The nature of statistical learning theory.* New York: Springer-Verlag.

Williams, Melissa S. 1998. *Voice, trust and memory: Marginalized groups and the failings of liberal representation*. Princeton, NJ: Princeton University Press.

Williams, Peter M. 1995. Bayesian regularization and pruning using a Laplace prior. *Neural Computation* 7(1):117–43.

Wordle. 2008. http://wordle.net (accessed July 13, 2008).

Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology and Politics* 5(1):33–48.

Downloaded from http://pan.oxfordjournals.org/ at University of Pennsylvania Library on June 3, 2013