Language and Ideology in Congress[*]

Daniel Diermeier[†]
Jean-François Godbout[‡]
Bei Yu[§]
Stefan Kaufmann[**]

Abstract:

The paper analyzes legislative speech records in the U.S. Senate from the 101st-108th Congresses. We apply a widely-used text classification algorithm - Support Vector Machines (SVM) - to extract the terms that are most indicative of conservative and liberal positions from legislative speeches and to predict Senators' ideological positions in the 108th Congress. When predicting the conservative/liberal positions of Senators based on previous legislative speech, the SVM classifier achieved a 94% level of accuracy. Our results indicate that the same ideological differences that are associated with roll-call based ideological measures in Congress are also reflected in political speech. This suggests that previous findings of low dimensional issue space in Congress are not simply the results of institutional factors such as party leadership or agenda control. Rather, it implies that voting and debating are different but correlated expressions of underlying ideological belief systems.

Note: Corresponding author, d-diermeier@kellogg.northwestern.edu

[*] We wish to thank Marcella Wagner for her superb research assistance. Generous support by the Canadian Institute for Advanced Research is gratefully acknowledged. Phil Burns of Northwestern Academic Technologies supplied us with an excellent Part-of-Speech tagger.
[†]Department of Managerial Economics and Decision Sciences (MEDS) and Ford Motor Company Center for Global Citizenship, Kellogg School of Management and Northwestern Institute on Complex Systems (NICO), Northwestern University.
[‡] Department of Political Science and Ford Motor Company Center for Global Citizenship, Kellogg School of Management, Northwestern University.
[§] Ford Motor Company Center for Global Citizenship, Kellogg School of Management and Northwestern Institute on Complex Systems (NICO), Northwestern University.
[**] Department of Linguistics, Northwestern University.

Over at least 50 years of research, *ideology* has been used to explain the political behavior of voters, legislators and other elites. In the context of mass political behavior, Converse (1964) has conceptualized ideology as a "belief system." Belief systems give structure to an individual's view on various issues. Intuitively, a political belief system expresses a view of which issue positions go together, the "knowledge of what-goes-with-what" (Poole 2003).

Empirically, belief systems allow us to predict an individual's position on an issue if we know his or her position on another issue. Speaking about belief "systems," however, does not necessarily imply a logically consistent political, economic, or social world-view. Indeed, as Converse (1964) argued, the association between issues may just have been contingent and reflect a particular, perhaps cultural or historical experience. Nevertheless, belief systems do constrain. It is quite unlikely (though not impossible) that a randomly selected U.S. voter who opposes universal health insurance, gun control, affirmative action, environmental regulation, abortion, and higher taxes also supports gay marriage. Converse expresses this idea as follows:

> "Constraint may be taken to mean the success we would have in predicting, given an initial knowledge that an individual holds a special attitude, that he holds certain further ideas and attitudes (Converse 1964, p.207)."

Converse (1964) and later Poole (2003) have argued that elite behavior, such as voting patterns in a legislature, can also be understood as the expression of underlying belief systems. The main difference is that belief systems at the elite level will always be more clearly defined – the idea being that elites are "true believers" (Poole 2003).

A parallel but closely connected literature is related to the development of the spatial model of voting (Downs 1957, Black 1948, Davis and Hinich 1966, Davis, Hinich, and Ordeshook 1970) which represents policy preferences as points in a multi-dimensional issue space. The spatial model has subsequently become *the* main building block in formal models of political behavior.

Measuring ideological orientations and belief systems, however, has always been a very difficult task. Unlike party affiliation, for example, ideology is not directly observable. Consequently, scholars have employed different strategies, ranging from survey responses to statistical estimates based on voting records. In legislative politics, and especially the U.S. Congress, the most widely used measure of ideology remains the vote-based score developed and refined by Poole and Rosenthal (1991; 1997; 2007) and McCarty, Poole, and Rosenthal (1997; 2006). The authors estimate ideology in Congress by applying a spatial voting model to Congressional roll call data. Legislators' ideal points are then estimated in choice spaces of various dimensions.[6] Perhaps the most important finding of the Poole-Rosenthal approach is that much of the voting behavior in Congress can be explained by a stable, low-dimensional issue space. Indeed, Poole and Rosenthal (1997) find that between 1789 and 1985, a two-dimensional spatial model (estimated with D-NOMINATE scores) can correctly classify about 85 percent of the individual voting decisions of each member of Congress. Moreover, for most periods of American history, a single dimension is sufficient. For example, in recent years (104th-106th Congress) a single dimension can account for about 90 percent of all roll call choices by members of Congress (Poole 2005).

[6] An alternative, but closely related statistical approach was developed by Heckman and Snyder (1997).

In addition to being successful in classifying complex voting behavior, the two spatial dimensions can also be given a straightforward ideological interpretation. Poole and Rosenthal (1997, 2007) have argued that the first dimension represents a traditional left-right dimension associated with the government's role in the economy and economic re-distribution, while the second dimension represents issues of state rights, slavery, and later racial and civil rights issues. What is most striking, according to Poole at al., is that after the civil rights reform of the 1960s, the second dimension has gradually lost some of its predictive power in explaining roll call outcomes. Poole (2003) argues that race-related issues are now largely correlated with questions of economic redistribution.

By now the finding of low dimensionality has been recognized as an important characteristic of Congressional decision-making.[7] Strictly speaking, low dimensionality in Congress implies that a representative's voting record on one issue will be a fairly good predictor of his/her vote choice over any other unrelated issue. For example, we know that liberal lawmakers favor fewer regulations of personal behavior and higher levels of income redistribution. We also know that conservatives typically favor more regulations of private personal behavior and fewer economic restrictions.

What is missing is an explanation for *why* we observe this apparent coherence. We can, for example, imagine a libertarian position which favors lower restrictions in both the economic and the personal domains -- e.g., one which opposes labor regulations and restrictions on marijuana use. These positions, however, are not represented in Congress to a significant degree, at least not in any member's voting record. The coherence is particularly striking if we restrict attention to issues of morality, culture, and the like. A legislator who is voting to oppose gun control is also likely to limit abortion

[7] Initially, this finding met with widespread disbelief. See Poole and Rosenthal (1997; p.8).

rights and vice versa. In other words, the simple fact that we can explain most issue voting by a single "redistributive" or left-right dimension does not explain why this bundling occurs and why we observe consistent voting in other (apparently) non-economic domains as well. A convincing explanation for the low dimensionality of Congressional voting behavior is still lacking.

In recent work, Poole (2003, 2005) has argued that low dimensionality of voting emerges as the *constrained* mapping of a high-dimensional space onto a lower-dimensional *hyperplane*. There has been a considerable amount of work trying to understand the complex strategic process by which high-dimensional policy options and preferences are mapped onto a low-dimensional space. One prominent idea is that the sample of policy alternatives that come to a vote is not random, but the result of strategic selection by party leaders and other institutional factors designed to fall along ideological or partisan lines. The constrained choice space is therefore the result of agenda control by strategic agents, especially party leaders (Aldrich 1995, Aldrich, Rohde, and Tofias 2004, Cox and McCubbins 1993, 2005, Kiewiet and McCubbins 1991, Rhode 1991, Rhode and Aldrich 2001, 2000). The cohesion between parties and the conflict between them is often said to reflect more directly an artifact of party leadership and their use of institutional prerogatives.

Conversely, this means that in the absence of such constraints a richer, higher-dimensional belief space should emerge. Thus if we still find a low-dimensional ideological space when institutional constraints are relaxed, the idea of constrained mappings shaped by institutional rules and strategic calculations becomes less plausible.

In other words, if there is nothing to constrain, the explanatory power of constraints vanishes.

This is the hypothesis that we explore in this paper by utilizing a different form of political behavior: Congressional speech. The idea is that voting and speech are both expressions of a common underlying belief system, but that speech during a Congressional debate is less constrained by institutional rules. This will be particularly true of speeches in the U.S. Senate, where germaneness requirements are less binding than in the House of Representatives. Note that our focus in this paper is on constraints in the legislature, especially agenda control. To be sure, Congressional speech may still be shaped by non-legislative constraints such as constituency concerns, but whether this is the case or not, such *electoral* constraints would be unrelated to the *legislative* agenda control proposed by institutional approaches. In other words, a finding of highly coherent Senatorial speech casts doubt on the idea that representatives' (unconstrained) belief spaces are highly complex and multi-dimensional and become projected onto a lower-dimensional space only through agenda control during the legislative process. The sources of ideological coherence then must lie elsewhere, perhaps in the electoral role played by parties. We investigate such alternative explanations below.

Political text has been an underutilized source of data in political science, in part due to the lack of rigorous methods to extract and process relevant information in a systematic fashion. Recent advances in computational linguistics, however, have opened up new venues for analyzing political language in various domains and contexts (for examples see Laver and Benoit 2002, Laver, Benoit, and Garry 2003, Monroe and Maeda 2004, Quinn et al. 2006, Simon and Xenos 2004). In this paper, we propose to analyze the ideological content of speeches made in the United States Senate between 1989 and 2004, fully capturing the 101st -108th Congress using a common text classification algorithm: Support Vector Machines (SVM).[8] We proceed as follows: First, we rank our sample of 177 unique Senators according to their DW-NOMINATE scores[9] In our first analysis we use the speeches made by the 25 most liberal and 25 most conservative Senators in the 101st to 107th Congress as our reference set. This reference set will be used to train the classification algorithm. We then employ the trained algorithm to classify the 25 most liberal and 25 most conservative Senators in the 108th Congress.[10] Note that this approach goes beyond testing whether the ideologies of a given Senator are consistent over time,[11] as the 25 most conservative and 25 most liberal Senators may or may not be identical to the list of Senators used for the training set.[12] Rather, if successful, our method will identify characteristics of political speech that are consistent over time and will be shared by members with similar ideologies. We then use the same methodology to investigate the moderate Senators -- i.e., those whose DW-NOMINATE scores did not rank among the 25 most liberal or conservative ones.

Previous work (e.g. Quinn et al. 2006) has shown that the issues discussed in Congress vary substantially from year to year. Quinn et al. (2006) used a multinomial mixture model to automatically cluster the 106th-108th Senatorial speeches into 42 topics.

---

[8] To the best of our knowledge, this is the first application of SVM algorithms in the study of legislators' ideologies. Other approaches are briefly discussed below. A more detailed analysis of these methods can be found in Yu et al. (2007).
[9] We use the Poole and Rosenthal DW-NOMINATE SCORES available at http://voteview.com/dwnomin.htm.
[10] 47 of these 50 "extreme" Senators had already served in the 107th Congress.
[11] This issue was investigated by Poole (2003) in the context of voting behavior. Poole found strong support for individual ideological consistency for members of Congress over time.
[12] 91 Senators in the 108th congress served in previous congresses. 44 of 50 extreme Senators in the 108th were rated as extreme in previous congresses.

By adding a time parameter, this model can also compute the changes of the topic distributions across Congresses, and thus describe the dynamics of the political agenda. While new topics emerge frequently, Quinn et al. do find stability among a set of identified topics. The relative attention paid to various topics, however, varies substantially, a fact Quinn et al. attribute to agenda control. In other words, in a specific year, the focus may be on the war in Iraq. In another, it may be on accounting reform, or on an appointment to the Supreme Court. For our classification tool to be successful, we need to show that despite the varying issues discussed in Congress, Conservatives and Liberals always talk about any of these topics in *distinct* and *stable* ways.

In the following, we focus first on the more extreme members of the Senate because we expect that their ideological belief systems are more sharply defined. Later we investigate the case of moderate Senators to see whether they exhibit different belief systems or whether their ideologies are just less well defined, constituting a "blend" between the more sharply defined belief systems found at the extremes.

Computational Analysis of Political Texts

Political scientists have traditionally relied on labor-intensive coding methodologies to study political texts, party platforms, and campaign speeches. (e.g. Budge et al. 2001, Baumgartner and Jones 1993, 2002, Baumgartner and Jones 2005). In the U.S. Congress, however, official records of speeches and debates are printed on a daily basis in large volume, which makes such an approach highly impractical. Recent developments in computational linguistics, however, have provided researchers with a series of tools that can be used to automatically process texts.[13]

Computational approaches have been mainly used in two areas: first, in the study of issues, topics, and legislative agendas, and second, in the estimation of *ideal points*. The main goal of the first application is to understand how policy agendas change over time. The most comprehensive work in this area is by Baumgartner and Jones (e.g., Baumgartner and Jones 1993, 2002, 2005), who relied on manual coding of transcripts of hearings and other documents. Other examples of manual coding in the context of Congressional speech and hearings include Bertelli and Grose (2007), the *Congressional Bill Project* by Adler and Wilkerson, and the *Policy Agenda Project* by Baumgartner, Jones and Wilkerson. Recent work has also applied fully or partially computerized approaches. These include dictionary-based approaches (e.g. Gerner et al. 1994, Laver and Garry 2000) which classify topics automatically using a manually coded dictionary, as well as fully automated approaches (Purpura and Hillard 2006, Quinn et al. 2006, and Schonhardt-Bailey 2006).

The second application of automated text analysis is in the area of ideal point estimation. Laver, Benoit, and Garry (2003) took a semi-automated approach in their analysis of European party manifestos, first using human coding to classify a few parties as reference points on a left-right ideological scale and then extracting word frequencies from their respective party manifestos. Other parties' ideological positions were then estimated based on how closely their party manifestos' word frequency distributions

[13] For a recent review see Cousins and McIntosh (2005).

matched the reference distributions.[14] An alternative approach was developed by Monroe and Maeda (2004). These authors use statistical techniques similar to Poole and Rosenthal (1997) to estimate "rhetorical ideal points." Their choice space, however, is a matrix of word counts rather than vote counts for each legislator. Unfortunately, as they discuss in their paper, their approach faces various statistical and computational difficulties.

In our own work, we use supervised learning techniques, a widely used approach in computational linguistics and computer science. Supervised learning approaches use annotated text to train a classifier, which is then tested on new text not contained in the training corpus. Success of the classifier is measured by its performance on this unseen text.

Our methodological approach is most closely related to Laver et al. (2003), though there are important differences. The main distinction is that we use supervised learning algorithms rather than direct comparisons of word frequencies, and that our methodology can be applied to any categorization, whether based on human coding (as in the party manifestos project) or not (as in the case of DW-Nominate scores). In addition, our approach allows us to identify the most relevant speech features that drive the classification. This provides additional insight into the underlying belief systems.

There are only a few applications of this approach to political domains. Examples include studies by Purpura and Hillard (2006), who used supervised learning techniques to code speech topics in the U.S. Congress, and Thomas, Pang and Lee (2006), who investigated (with modest success) whether speech classification of floor debates in the

[14] See also Benoit and Laver (2003, 2005), Laver and Benoit (2002), Benoit, Laver, Arnold, Pennings, and Ho (2005), Bertelli and Grose (2006), Imbeau (2005), and for a critical view Budge and Pennings (2007).

House of Representatives on a specific bill can be used as a predictor of subsequent agreement. Thomas et al. predicted speakers' opinions about a specific bill (support or opposition) based on their speeches. Their classifier was trained on 2740 speech segments in 38 bill debates and achieved an accuracy of 66% in predicting the opinions expressed in 860 speech segments from 10 different legislative debates.

Our own classification takes a different approach. We focus on the 25 most conservative and 25 most liberal Senators in the 101st and 107th Congress, using DW-NOMINATE scores as a proxy measure for ideology. These Senatorial speeches serve as the training set for our algorithm, which is then used to classify Senators from the 108th Congress. For each Senator, the classification by our algorithms is then compared to their respective DW-NOMINATE scores in the 108th Congress.

We chose to analyze Senatorial speech because, in the absence of cloture votes, individual Senators enjoy the right to unrestricted debate. In contrast to House members, Senators are also allowed to seek permission to speak out of turn on virtually any subject, or to interrupt proceedings to discuss any unrelated issue. Moreover, since germaneness of amendments is not required in the Senate, and lawmakers have the right to interrupt legislative proceedings with unrelated matters, we are presented with an ideal setting to assess whether party leadership and institutional rules can account for the low dimensionality of voting found in the U.S. Congress. If agenda control and other institutional constraints are essential to explain why a multidimensional issue space which deals with unrelated topics, like abortion and foreign policy, can be bundled into a low dimensionality voting space, then a classification based on the use of ideological language will perform poorly when it comes to predicting policy positions. However, if the observed low dimensionality in voting is explained better by a pre-existing political ideology (as expressed in speech) than by institutional factors, we should find that our classification scheme will perform effectively in predicting the ideological positions of Senators.

Methodology - Text Classification Algorithms and Support Vector Machines

Many supervised learning algorithms have been used for text document classification. Based on performance in previous classification tasks,[15] *Support Vector Machines* (SVM) have been identified as one of the most efficient classification methods (Dumais et al., 1998; Joachims, 1998; Yang and Liu, 1999; Sebastiani, 2002).[16] For binary ("yes" or "no") classification problems, the two most frequently used evaluation criteria are accuracy and the average of precision and recall.[17] The latter is often used for skewed data sets in which one category has much fewer examples than the other. The training set in this study, however, is quite balanced, with similar numbers of examples for both categories. Hence we chose accuracy as our evaluation criterion. SVM classification accuracies may vary widely between different classification tasks. They may be as high as over 95% for some topic categories in news articles, or as modest as around 75% for opinion classification of movie reviews (Pang et al., 2002). A classifier is usually

[15] The performance of classification algorithms is tested using common benchmark data sets. The Reuters-21578 news collection, the OHSUMED Medline abstract collection and the 20 Usenet newsgroups collection are the most widely used benchmark data sets. The Reuters-21578 collection is available at http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.
The OHSUMED collection is available at http://trec.nist.gov/data/t9_filtering.html. The 20 newsgroups collection is available at http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.

[16] We also compared our SVM algorithm to *naïve Bayes*, another popular classification method. Our experiment results show that SVM is slightly superior to naïve Bayes for ideological position classification (Yu et al., 2007).

[17] Accuracy is defined as the proportion of correct predictions among all predictions. Precision and recall are the proportions of documents correctly assigned to a category among all the documents assigned to that category, and among all the documents that truly belong to that category, respectively.

considered effective when the accuracy is higher than a baseline method, which can be something as simple as a random guess for balanced data sets or a majority vote for skewed data sets.

SVM is a supervised learning method based on the Structural Risk Minimization principle from statistical learning theory (Vapnik, 1982; Cortes and Vapnik, 1995; Vapnik, 1999). In our first analysis we have a binary classification problem with the two categories "(extreme) conservative" and "(extreme) liberal." The data -- documents in our application -- are represented as vectors in an *n*-dimensional space, each dimension corresponding to a feature deemed relevant to the classification task.[18] In the training phase, the category membership of each data point is known. We arbitrarily label one category as "negative" and the other as "positive," consistent with the terminology used in classification tasks. Thus a training set of l examples is represented as a set of pairs

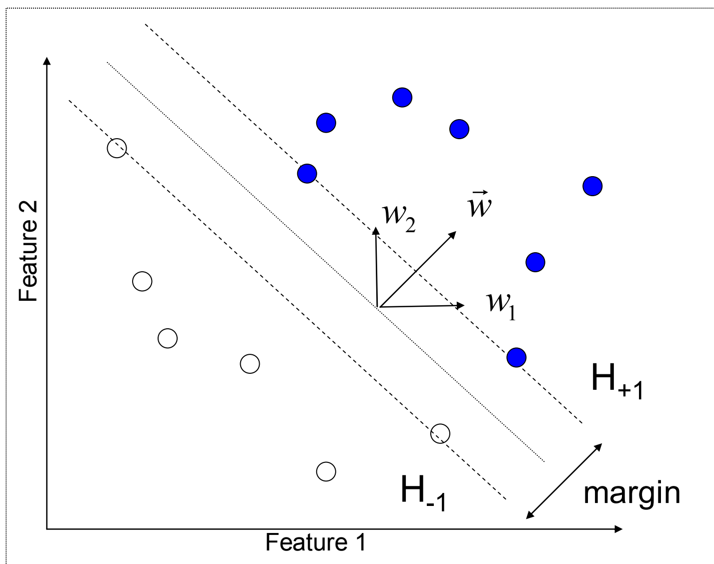
$$\{(x_1, y_1), \ldots, (x_l, y_l)\},\ x_i \in R^n,\ y_i \in \{-1, +1\} \tag{1}$$

The Support Vector Machine model is based on the following idea. If the data points in each of the categories are separable by a hyperplane, then there is a hyperplane that is *maximally separating*, in that the distance between it and the nearest data point is maximized. This "ideal" hyperplane lies at equal distances between two parallel hyperplanes, each of which is determined by one or more of the data points in one of the two categories. These data points on the parallel hyperplanes are called the *Support Vectors* (SV). The distance between the two parallel hyperplanes is called the *margin*. The task of a Support Vector Machine in the training phase is to find the two separating hyperplanes such that the margin is maximal.

[18] Details on the way in which these vectors were derived from the documents are discussed in the next section.

For illustration, consider the special case that each data point is represented by a pair of coordinates in a two-dimensional space, as shown in Figure 1. The general notion of a hyperplane corresponds to a line in two-dimensional space. In the figure, the "ideal" separating hyperplane and the two parallel hyperplanes running through the support vectors are shown as the dotted line and the two dashed lines H_{-1} and H_{+1}.

Figure 1 – Linear separating hyper-planes with maximized margin



The points x on the best separating hyperplane (the dotted line in the figure) satisfy the equation

$$w \bullet x + b = 0 \tag{2}$$

where w is a vector perpendicular to it and $|b|/\|w\|$ is its perpendicular distance from the origin ($\|w\|$ is the Euclidean norm of w). The SVM training algorithm yields values for b

and w such that for all (x_i, y_i) in the training set, $w \bullet x_i + b \leq -1$ if $y_i = -1$ and $w \bullet x_i + b \geq +1$ if $y_i = +1$; the corresponding equalities hold for the points on H_{-1} and H_{+1}. In the test phase, the classification of a new data point x – represented in the same vector space but of unknown category membership – is based on its location relative to the best separating hyperplane, given by the sign of $w \bullet x + b$.

Aside from its role in the classification of new data, the vector w is of interest in its own right, as it furnishes valuable information about the informativity of each feature in determining category membership. To see this, consider again the two-dimensional case illustrated in Figure 1, and suppose the training data were distributed in such a way that the best separating line runs parallel to the abscissa. In that case, the decision in determining the category membership of a new data point would be based entirely on its y-coordinate. In other words, knowing the value of Feature 2 would be sufficient for determining category membership, whereas the value of Feature 1 would be irrelevant. Notice, too, that in that case the vector w would be perpendicular to the abscissa, its x-component being zero.

In general, since the dimensions of the vector space correspond to the features used in the classification, the components of w can be used as feature ranking coefficients. If the classifier performs well, those components of w whose absolute values are highest correspond to the most informative features. This is a common property of all linear classifiers. Before the invention of SVM, Fisher's linear discriminate analysis and logistic regression were used for the same purpose. Linear SVM is a particular kind of linear discriminate classifier which maximizes margins. Linear SVM's feature ranking property has been exploited in some applications, for instance by Guyon et al. (2002) to identify a small subset of genes which are biologically relevant to a certain type of cancer. As a feature selection method, SVMs are superior to correlation coefficients since they are optimized during training to handle multiple features simultaneously, whereas the latter measure the correlation between individual features and the target concept. While there are differences between Guyon et al.'s biomedical application and our text classification task, we, too, can use *w* as a "weight vector" whose components measure the relative informativity of the linguistic features of documents as indicators of conservative or liberal ideology.

More formally, the SVM algorithm maximizes the margin between the two separating hyperplanes by finding the maximum of the functional

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{3}$$

subject to the constraints

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1,2,\ldots,l \tag{4}$$

Only the support vectors have non-zero α_i values. For the other data points, $\alpha_i = 0$. In (3) above, $K(x_i, x_j)$ is the *kernel function*. We use the linear kernel $K(x_i, x_j) = x_i \bullet x_j$ because it suits the text classification problem well. The linear kernel function can be replaced by other functions to handle non-linear boundaries. Studies show, however, that they do not improve text classification performance significantly (Leopold and Kindermann, 2002).

After training, the parameters *w* and *b* are given by (5) and (6). Given a text example *x*, the linear decision function is $\mathrm{sgn}(w \bullet x + b)$.

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{5}$$

$$b = y_{sv} - w \bullet x_{sv} \text{ [19]} \tag{6}$$

SVMs originally only dealt with linearly separable data, for which at least one hyperplane can be found so that all positive examples are on one side of the hyperplane and all negative examples are on the other side (as shown in figure 1). Cortes and Vapnik (1995) extended the method to non-separable data, which is often the case in real world classification problems.

There are several efficient implementations of the SVM algorithm, such as LIBSVM (Chang and Lin, 2001) and SVM*light* (Joachims, 1999). We used the SVM*light* package with its default setting in this study.

Document representation

As noted above, in our text classification model, documents are represented in a vector space whose dimensions correspond to the features that are relevant in the classification. In our implementation, the relevant features are words (more precisely, word *types*), and the vector representing each document is determined by the number of occurrences (or *tokens*) of each of the words in that document. The simplest and most widely used method for obtaining vectors from documents views the latter as "bags of words" (BOW).

For each word in the feature space, the value for a given document may be Boolean (recording the word's presence or absence in the document), the word's frequency (the number of occurrences in the document), its frequency normalized by

[19] *sv* can be any support vector. In SVM*light* the first support vector (according to its order in the input data) was used to compute *b*.

document length, or its frequency weighted by some weighting scheme. The most common word frequency weighting scheme is a family of measures subsumed under the label *tf***idf*. Here *tf* and *idf* stand for "term frequency" (the number of occurrences of the word in the document) and "inverse document frequency" (the inverse number of documents in which the word occurs). The idea is to offset the impact of high-frequency words in the given document by the extent to which they also occur in other documents, based on the assumption that words whose occurrences are dispersed over many documents are less useful in the classification task. Specifically, in our implementation, given a training set of n documents, a word with term frequency t_f in a document of length l and document frequency d_f, that word's *tf***idf* value is given by the formula

$$tf * idf = (t_f / l) \log(n / df) \tag{7}$$

The *tf***idf* weighting applies to the training data only. In the prediction step, the test examples are completely independent from each other, and the total number of test documents should be unknown to the classifier. Furthermore, normalization does not change the frequency ratios between words, and consequently does not change the classification result. Therefore we used the raw word frequencies to represent test document vectors in our *tf***idf* experiments.

The BOW model is insensitive to many potentially informative properties of documents, such as the location of words in the document and relative to each other, grammatical relations, the internal structure of the document at various levels (e.g., paragraph and sentence boundaries), and multi-word phrases such as "war on drugs." Some research studies tested more sophisticated document representations which incorporate such information (e.g. word position relation predicates and noun or verb phrases) in a variety of classifiers. However, experimental results showed that these complex features did not improve the classification performance significantly (Lewis, 1992; Cohen and Singer, 1999; Scott and Matwin, 1999; Moschitti and Basili, 2004). Therefore we applied the BOW model in this study.

The use of computational linguistics to study political behavior is a fairly new endeavor. It may therefore be useful to discuss performance criteria in more detail. As our paper is the first to use SVM to classify political ideology, we cannot compare our findings with pre-existing work in the same domain. However, there are some reference points in other domains, such as customer opinion classification with movie or restaurant reviews. In this application, the goal is to correctly classify reviews as "positive" or "negative." Opinion classifiers have achieved accuracy levels as high as 88% for product reviews (Dave et al., 2003) and 82% for movie reviews (Pang et al., 2002). However, Finn and Kushmerick (2006) found that an opinion classifier trained on movie reviews was not effective in predicting the polarity of restaurant reviews, and vice versa (see Table 1).[20]

Table 1: Domain transfer problem in customer review classification (Finn and Kushmerick, 2006)

	Movie	restaurant
Movie	76.8%	40.1%
Restaurant	55.5%	88.5%

This "domain-dependence phenomenon" is evidence that existing opinion classifiers can be used in detecting domain differences. In the political context, we are not dealing with different product domains, but different issue domains. A highly coherent

[20] For example, some typical adjectives in movie reviews (like *hilarious* and *boring*) are unlikely to occur in restaurant reviews, although some opinion descriptors (like *terrific* and *bad*) are universal.

ideology will view different issues in the same way. In the words of Converse (1964), knowing the attitude on a set of issues will be predictive of the views on another, unrelated issue. The analog of this assumption in the domain of customer reviews would lead to the prediction that customers will only use universal opinion descriptors, such as *terrific* or *disappointing*, rather than domain specific ones (*hilarious* and *funny* in the case of movies, *overpriced* and *overcooked* in the case of restaurants). In other words, we expect that low classification success across Congresses (given the variation of issues over time) will be evidence of highly issue-specific and orthogonal attitudes. On the other hand, high classification success will be consistent with a constrained belief system that sees different issues in the same way.

Data preparation

We downloaded all the Senatorial speeches of the 101st-108th Congresses from the website *thomas.gov*. We then converted the original HTML files to raw text by removing the HTML tags, headers, tables, lists, and unicode characters, and segmented the files into individual speeches. An individual speech is a Senator's speech given in a continuous time period until he or she stops. The beginning of a speech is always "Mr/Ms/Mrs. XXX," but the end of a speech can be the beginning of another Senator's speech, an officer's action, or a document inserted into the printed record. We created a set of heuristic rules to segment the speeches.

We used the DW-NOMINATE score as the measure to select the "extreme" Senators – the 25 most conservative and the 25 most liberal Senators in each Senate. Hence a training document is a Senator's complete set of speeches in each Senate over the 101st-107th Senates, and a test document is a Senator's complete set of speeches in the 108th Senate. Thus there are 350 training documents and 50 test documents, 400 in total.

Shallow natural language processing

The documents were subjected to a variety of pre-processing procedures in order to generate different vocabularies and document vectors for classification. These procedures comprised different combinations of tokenization, stemming, and part-of-speech tagging.[21] We used a simple tokenizer to split the speeches into individual words (for an example see Table 2). The tokenizer recognizes consecutive strings of alphabetical characters as valid words (see the example in Table 3). Part-of-speech tagging is of particular importance, as part of our goal was to investigate the role of each content word class – nouns, verbs, adjectives and adverbs – in isolation (see the example in Table 4).

To reduce the vocabulary size, we arbitrarily set a minimum term frequency of 50 and document frequency of 10 for a word to be selected as a feature, assuming that words with frequencies below that threshold are not useful for classification. We also removed the top 50 most frequent words as "stopwords" (mostly function words, such as *the*, *a*, and *of*). Stopwords are considered useless for classification because they occur frequently in every document. However, in those experiments in which the vocabularies were limited to content words – nouns, verbs, adjectives and adverbs – stopwords were not removed, since function words were already excluded in the part-of-speech selection. Since every Senator's ideological label is expected to be consistent across Congresses, and because 44 out of 50 Senators in the 108th Senate (the test set) are also members of

[21] Tokenization splits a piece of text into individual words. In stemming, words are stripped of certain inflectional and derivational morphemes; for example *stemming* may be converted to *stem*. Part-of-speech tagging is the assignment of the part-of-speech (e.g. nouns, verbs, etc.) to each word token in the text.

previous Senates (the training set), we conjecture that Senators' names are correlated with their ideological labels. To prevent the prediction model from degrading to a name classifier, we removed all names of Senators from the vocabularies. Similarly, there is a correlation between state names and ideological labels. The Senators in the training set represented 45 states; among them, the Senators from 17 states were all conservative and those from 18 other states were all liberal during the 101^{st}-107^{th} Senates. To prevent the classifier from being dominated by this correlation, we removed all state names and their abbreviations from the vocabularies. For parts-of-speech results, we can conveniently remove all the proper nouns.[22]

The Porter Stemmer (Porter, 1980) was used for suffix trimming. Stemming reduces the vocabulary size significantly by mapping different forms into the same stem, but it can also be harmful for information retrieval and classification if different forms of the same word contribute differently to the classification. For example, stemming verbs discards the tense information, which would harm the classification if verb tense were an important predictor. We used the MorphAdorner tagger[23] to tag the parts-of-speech. Because the tagger has its own tokenizer, the generated word forms in this case are slightly different from the results of the simple tokenizer.

[22] Apparently a named-entity recognizer would be necessary to completely remove the social network impact. We leave this process to future work.

[23] We wish to thank Pib Burns of Northwestern University's Academic Technologies group for providing us with the tagger.

Table 2: A speech example

```
<DOC>
<DOCNO>107-akaka-hi-1</DOCNO>
<TEXT>
   Mr. AKAKA. Mr. President, I express my sincerest sympathies to the families of those who have lost
loved ones in two unrelated incidents the U.S. military in Hawaii during the past week.
   On Friday afternoon, the U.S.S. Greeneville collided with the Ehime Maru, a Japanese fishing vessel.
I join President Bush in expressing my regret to the people of Japan for this tragedy. My heart goes out
to the families of the nine people who are still missing following this incident.
   On Monday evening, two UH-60 Blackhawk helicopters crashed during a training exercise at the
Kahuku Military Training Area, resulting in six deaths. My thoughts and prayers are with the families
and units who are mourning the loss of their loved ones. I also wish a speedy recovery to those soldiers
who are recovering from injuries sustained in this accident.
   I am certain that the investigations into these incidents will be thorough and comprehensive. But my
purpose today is not to question why these incidents occurred, but to express the genuine sadness and
concern that I share with the people of Hawaii and the rest of the nation over these two unfortunate
episodes.
</TEXT>
</DOC>
```

Table 3: An example of a tokenized sentence

```
mr/president/i/express/my/sincerest/sympathies/to/the/families/of/those/who/have/lost/
loved/…
```

Table 4: An example of part-of-speech tagged sentence fragment with selected nouns

```
president-NN
sympathies-NNS
families-NNS
…
```

Experimental setup and results

The combinations of classification algorithms, feature sets and feature weighting schemes lead to various classification methods. We tested SVMs using four feature weighting schemes – word presence/absence, frequency, normalized frequency, and *tf*idf* - and six feature sets – words, stemmed words, nouns, verbs, adjectives and adverbs (Table 5 lists the vocabulary size for each feature set). Hence we tested 4*6=24 different SVM methods in total.

Table 5: Feature set sizes

	Feature set					
	Words	Word stems	Nouns	Verbs	Adjectives	Adverbs
Size	19503	12245	8840	6903	3644	888

In evaluating a classifier on a given fixed data set, it is necessary to divide the data into a test set and a training set so as to approximate the real-world situation, in which the classifier would be applied to new data after the training phase. There are generally two ways to approximate this situation. One is to set aside a sizeable portion of the data as a “held-out” set which is ignored during training and only used for testing. This approach is sound for data sets with large numbers of labeled examples. However, for small data sets such as ours, it is problematic, since the arbitrary training/test split may accidentally lead to two data sets that are unlikely to have been produced by the same source. A common way to address this problem is “*n*-fold cross-validation.” Here the total data set is split into *n* subsets of equal size, each of which is held out and used to test a classifier trained on a corpus comprising the remaining *n*-1 sets. The overall accuracy of the method is then measured by the average of the accuracies obtained in the *n* tests.

The usual choice of *n* is 5 or 10. We performed 5-fold cross validation on the training set to estimate the effectiveness of our classification methods. The cross-validation evaluation results are shown in Table 6. The results show that *tf*idf*-SVM is the best classification method, reaching an accuracy level as high as 93.3% averaged over the six feature sets. The technical details of the relative evaluation experiment can be found in (Yu et al., 2007). This paper focuses on the analysis of the *tf*idf*-SVM classification results.

Table 6: 5-fold cross validation on the training set

Representations	Feature sets						Average acc
	Word	Stem	Noun	Verb	ADJ	ADV	
Boolean	87.4	86.3	87.4	81.1	88.3	74.6	84.2
Frequency	84.0	84.0	78.3	61.4	76.0	64.6	74.7
Normalized frequency	88.6	88.9	81.4	63.4	82.6	61.7	77.8
*tf*idf*	**94.3**	**94.6**	**96.6**	**93.7**	**93.4**	**87.1**	**93.3**

Table 7a: *tf*idf*-SVM extreme prediction results (with extreme training set)

*tf*idf*-SVM	Feature sets					
	Word	Stem	Noun	Verb	ADJ	ADV
extreme	**94.0**	86.0	84.0	88.0	**94.0**	52.0

The *tf*idf*-SVM method achieved an accuracy of 94% with the Word or Adjective feature set (table 7a). This result implies that out of the 50 most conservative and liberal Senators in the 108th Congress, 47 were correctly classified, solely in virtue of the words they used on the floor. Recall that 44 of the 50 "extreme" Senators were also extreme members of previous Senates. The remaining six non-extreme Senators are Allen, Reid, Chambliss, Cornyn, Graham, and Sununu. Sen. Allen was a new member of the 107th Congress, so he did not have speeches recorded in the Thomas database. Sen. Reid had served in the Senate for a long time. His DW-NOMINATE score, however, had changed gradually from moderate (-0.235 in the 101st Congress) to extreme (-0.381 in the 108th Congress), so he is not represented in the training data, but only in the test data. The other four Senators were new members of the 108th Congress. The three misclassified Senators are Dayton, Reid, and Schumer. Sen. Dayton and Sen. Schumer are represented in the training data, so the "in-sample" accuracy is 42/44 = 0.95 and the "out-of-sample" accuracy is 5/6=0.83 (Note, however, that the out-of-sample set is extremely small).

This very high classification success suggests that Senatorial speech (at least for extreme Senators) is highly consistent over time, just as consistent as their voting records. However, in contrast to voting, which is the consequence of highly structured agenda setting and proposal process, speech (at least in the Senate) faces no such constraints. To the extent that both speech and voting are expressions of the same underlying ideology, this suggests that Senators' ideologies are highly structured *before* any legislative institutions can further constrain it: There is nothing to constrain for institutional constraints.

Classifying Moderate Senators

We next turn to an analysis of moderate Senators. Our first goal in this analysis is to demonstrate the validity of our approach. If indeed our classification methodologies capture ideological positions, then the algorithms trained on extreme Senators should perform worse on the test set of moderate Senators. We again used the DW-NOMINATE score as the measure to select "moderate" Senators. That is, we divided the remaining 50 Senators into 25 (moderate) conservative and 25 (moderate) liberal Senators in each Senate. Note that the classification task is now to correctly classify moderate conservatives versus moderate liberals, not moderates versus extremes.[24] The results are shown in Table 7b.

Table 7b: *tf*idf*-SVM moderate prediction results (with extreme training set)

***tf*idf*-SVM**	Feature sets					
	Word	Stem	Noun	Verb	ADJ	ADV
Moderate	52.0	52.0	54.0	**62.0**	58.0	48.0

The higher error rate in predicting moderate Senators using classifiers trained on extreme Senators is consistent with the interpretation that the speech classification is

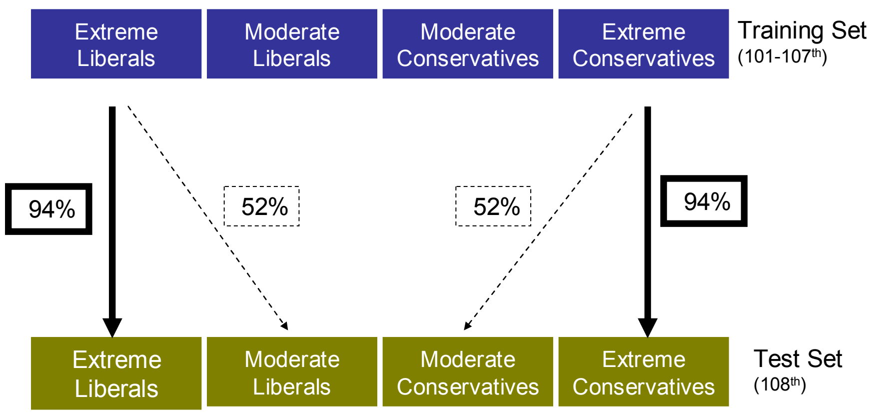
[24] We will investigate some of these other variants below.

indeed sensitive to ideological differences. However, it also suggests a new way to think about moderate Senators which uncovers a hidden assumption in the spatial representation of ideologies. According to the Poole-Rosenthal approach, the ideological position of members of Congress is identical to their ideal point in a multi-dimensional space. But once we fix an ideal point, moderate ideological positions should be just as well-defined and precise as “extreme” positions. There is, however, a different way to conceptualize moderates. We are not aware of a formal representation of this idea, but it does appear from time to time in popular political discourse. The idea here is that moderate liberals and conservatives do not constitute a separate position, but are simply more “blurry” versions of their extreme counterparts.

Our classification methodology allows us to investigate these competing accounts for moderates. According to the first hypothesis (moderate as a position), moderates would exhibit their own unique linguistic characteristics which distinguish them from the extreme Senators. The second hypothesis is that the moderates blend the vocabularies from the two extreme sides (moderates as blurry extremes). To investigate these hypotheses, we trained our classifiers on the two moderate segments and tested them on the extreme segments. This is the exact inverse of our previous experiment. If the “moderates as a position” is correct, then moderate categories should fail in predicting extreme Senators, but should succeed in classifying moderate Senators. If the “moderates as blurry extremes” hypothesis is true, however, we would expect the classifier trained on moderate examples not only to fail to predict extreme Senators, but also fail to classify moderate Senators. The following figures summarize these hypotheses.

First, as a reference point, recall the result of the first classification experiment, in which we used extreme Senators as our training set. We achieved high accuracy for extreme Senators, but low accuracy for moderate Senators (see Figure 2).

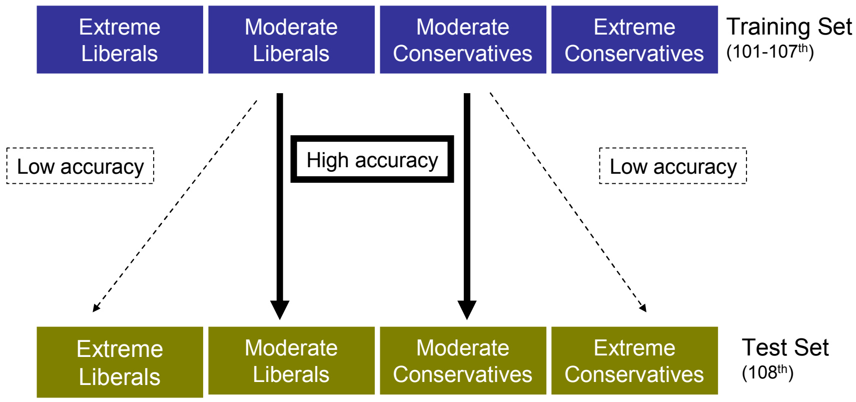
Figure 2 – Extremists to Moderates - Results



*tf*idf*-SVM classification using all words

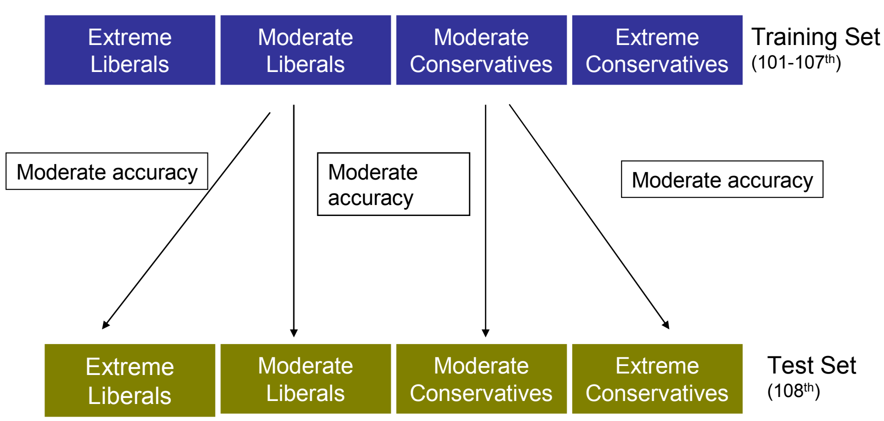
If the "moderates as a position" hypothesis is correct, we should see high accuracy for classifying moderate Senators, but low accuracy for classifying the corresponding extreme Senators (see Figure 3).

Figure 3 – "Moderates as a Position": Predictions



Finally, if the "moderates as blurry extremes" hypothesis holds, we should see moderate to low accuracy for extreme Senators, but also moderate to low accuracy for moderate Senators (see Figure 4).

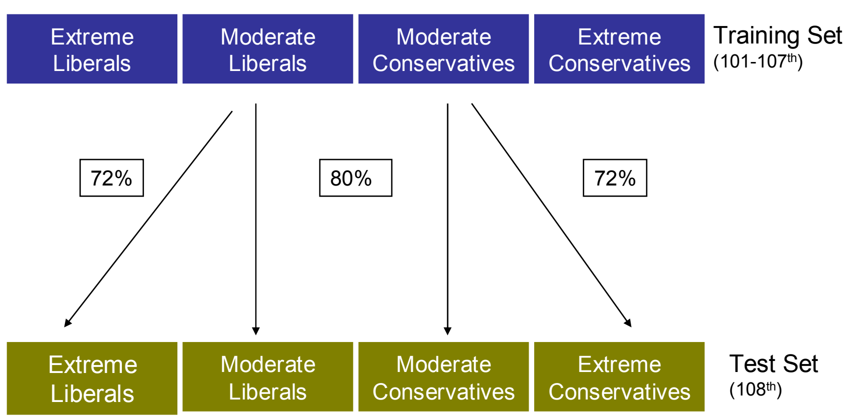
Figure 4 – "Moderates as Blurry Extremists": Predictions



To test the above hypotheses, we conducted the following experiments: First, we grouped the Senators in each Congress into four quartiles – extreme conservative, moderate conservative, moderate liberal and extreme liberal – based on their DW-NOMINATE scores. In contrast to our first experiment, which used extreme conservative and extreme liberal as the training set, this time we trained the classifier on moderate conservatives and moderate liberals. As before, we used the 101st-107th Congress speeches as training data and the 108th Congress for prediction. All other aspects of the experiments, including data preparation, part-of-speech tagging, etc., were identical.

Our results find clear support for the "moderate as blurry extremes" hypothesis and against the alternative hypothesis of "moderates as a position" (see Figure 5).

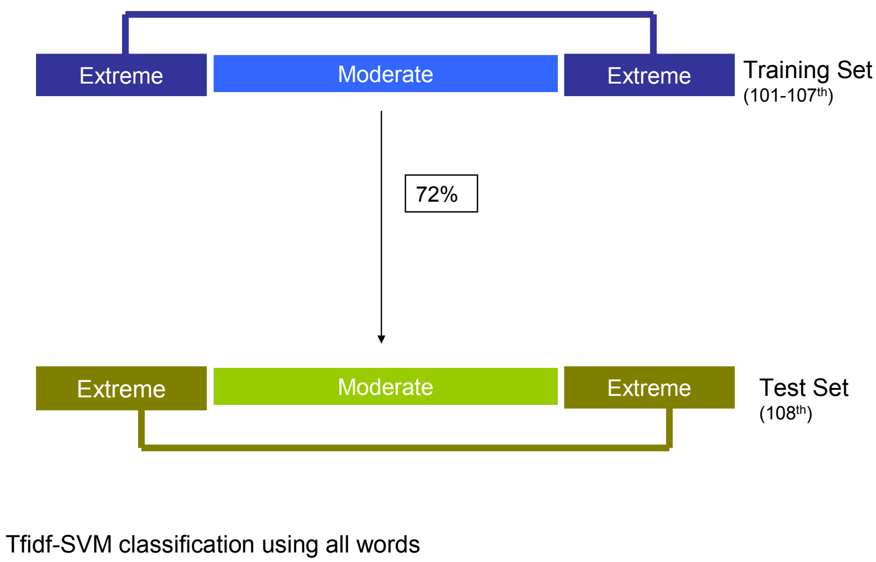
Figure 5 – Moderates to Extremists - Results



*tf*idf*-SVM classification using all words

One may argue that the classification task in this new experiment is inherently harder than the first one. That is, we may expect that a moderate liberal is *a priori* more likely to be misclassified as a moderate conservative than as an extreme liberal or an extreme conservative. (In both experiments a moderate liberal can be misclassified as a moderate conservative and vice versa). To address this concern, we conducted another experiment as a robustness check. This time, we used the first and last quartiles to form the "extreme class" and the intermediate two quartiles as the "moderate class." Now the goal is to classify extreme vs. moderate Senators, regardless of whether the extreme ones are conservative or liberal. If the "moderate as a position" hypothesis is true, we should see the moderate and extreme classes well separated. However, the low classification accuracy (see Figure 6) does not support the "moderate as a position" hypothesis, adding additional support to our conclusions from our initial results.

Figure 6 – Moderates vs. Extremists - Results



Our analysis suggests support for the view that extreme Senators have more clearly defined ideologies than moderate Senators. We can investigate this view further by analyzing the word rankings most responsible for the classification results.

Feature analysis

The word features were sorted by their coefficients in the generated SVM linear prediction function. “Positive” features indicate the linguistic characteristics of conservative speeches and “negative” features indicate the linguistic characteristics of liberal speeches. We report the corresponding list of vocabulary words obtained with the *tf*idf*-SVM feature set analysis. Table 8 shows the words with the highest feature weights in the classifier which achieved the 94% classification rate (see Table 7a).

Table 8: *tf*idf*-SVM Feature Set Analysis for all vocabulary

Words			
Liberal		***Conservative***	
FAS	lakes	habeas	homosexual
Ethanol	SBA	CFTC	PRC
Wealthiest	afterschool	surtax	tripartisan
Collider	nursing	marriage	scouts
WIC	arctic	cloning	ballistic
ILO	replacements	tritium	sandinistas
Handgun	orange	ranchers	salting
Lobbyists	libraries	BTU	M.I.A.
Enron	plutonium	grazing	NTSB
Fishery	veterans	catfish	FSX
Therein	disabilities	unfunded	gambling
Hydrogen	prescription	IRS	PAC
PTSD	NIH	unborn	abortion
Gun	lobbying	PLO	repurchase
Firestone	NRA	exclusionary	taxing
Trident		EMS	

Note: Words related to geography and individual names were removed. Words are decreasing in weight (from most Conservative or most Liberal). All words were converted into lower case during classification. In the above table the acronyms were recovered to upper case for the ease of reading.
Glossary: FAS: Federation of American Scientists ; WIC: Women, Infants, and Children Program; ILO: International Labor Organization; PTSD: Post-traumatic stress disorder; SBA: Small Business Arrangement; NIH: National Institute of Health; NRA: National Rifle Association; CFTC: Commodity Futures Trading Commission; BTU: British Thermal Unit; EMS: Emergency Medical Service; PRC: People’s Republic of China; M.I.A.: Missing in Action; NTSB: National Transportation Safety Board FSX: name of fighter aircraft ; PAC: Political Action Committee

The feature analysis yields some interesting insights. First, note that we find comparatively few words related to redistribution and taxation issues (*wealthiest*, *surtax*, *unfunded*). Notice also that Democrats are the ones using company names, unsurprisingly in the context of scandals (*Enron* and *Firestone* as in "Ford-Firestone"). Second, many of the top issues for Conservatives are from the domain of culture and values. Examples include *marriage*, *cloning*, *unborn*, *homosexual*, *gambling*, and *abortion*. Only *handgun* and *gun* (in the context of gun control) play that role for Democrats. We also find various terms related to local environmental and economic interests, such as *ethanol*, *hydrogen*, *arctic* (as in "Arctic National Wildlife Refuge") for Democrats, and *ranchers*, *catfish*, *grazing* for Republicans. Note also that ideologies express themselves not necessarily by talking differently about the same issues, but by talking about different issues - with Democrats, for example, using words related to corporate special interests and to the environment, while Republicans talk about abortion and other words related to values and morality.

We also conducted a classification based on nouns, verbs, adverbs, and adjectives.[25] Recall from Table 7a that the classification based on adjectives was particularly successful, followed by nouns and verbs, with adverbs providing by far the poorest classification.[26] A look at Table 9 explains why. Note again the predominance of culture and value related words, especially among the adjectives.

---

[25] We reproduce in Table 8-9 the most liberal and conservative words as they appear in our ranking, from first to the twentieth in rank order. However, for the purposes of this discussion, we selected words ranked in the top 50 to illustrate commonality.

[26] We note in passing that the SVM prediction based on adverbs obtains the lowest accuracy level in our extreme classification model. Only 52% of the Senators were correctly classified using this method, which is just about as efficient as a coin toss. Note also that the feature analysis yields are list of words without obvious political connotations. We believe that this poor classification rate is method related, specifically document-representation related, considering the fact that Adverbs still performed with a high level of accuracy in the cross validation test (87%). In Yu et al. (2007), we compare different classification methods

Table 9: *tf*idf*-SVM Feature Set Analysis for Nouns, Adjectives, Verbs, and Adverbs

Nouns		Adjectives		Verbs		Adverbs	
Lib	***Cons***	***Lib***	***Cons***	***Lib***	***Cons***	***Lib***	***Cons***
Disabilities	habeas	wealthiest	partial-birth	nursing	taxing	enormously	basically
Ethanol	surtax	reproductive	prepaid	policing	grazing	mentally	maybe
Hydrogen	ranchers	african-american	exclusionary	preexisting	tax	incidentally	morally
Libraries	cloning	managed	embryonic	battered	taxed	disproportionately	supposedly
Veterans	entitlements	toxic	wireless	crumbling	cloning	hugely	seldom
Handgun	catfish	homeless	chic	hate	married	critically	additionally
replacements	tritium	republican	ballistic	breaks	ranching	Chronically	abundantly
mammography	marriage	armor-piercing	unfunded	contaminated	adjourned	backward	objectively
Genocide	missile	generic	unborn	lifesaving	taxes	immensely	pretty
Collider	tripartisan	mail-order	ninth	lobbying	sued	ecologically	theoretically
Gun	marijuana	after-school	homosexual	laundering	importing	promptly	interestingly
sweepstakes	stem	mail-in	lawabiding	recessed	fracturing	woefully	purely
Handguns	interdiction	gay	fugitive	reformulated	sentencing	unacceptably	strictly
Fishery	ratepayers	community-based	nondefense	displaced	induced	comprehensively	irrespective
Plutonium	bureaucrats	dislocated	gaseous	gun	nationalize	densely	amok
firefighters	IOUS	superconducting	lineitem	bargain	deploying	crucially	assuredly
breast	soybean	coastal	out-of-state	revered	saluted	indiscriminately	downstream
guns	checkoff	pregnant	dual-use	promise	checking	candidly	likewise
Fisheries	circuit	cerebral	one-size-fits-all	smoking	tax	summarily	constitutionally
afterschool	adjournment	unanimous-consent	one-call	worry	court-martial	alike-	selflessly

Note: Words related to geography and proper names are removed. Words are decreasing in weight (from most Conservative or most Liberal).

Overall, we see that key issues discussed by liberals are energy and the environment (or alternative energy), corporate interests and lobbying, health care, inequality and education. For conservatives, the key issues discussed are taxation,

---

in terms of their extreme prediction results. For the adverb feature set, we find 74% prediction accuracy when using the presence/absence representation. So we can say tfidf causes overfitting for the Adverb feature set because it worked well on the training data (87.1% 5-fold CV accuracy) but poorly on the test data.

abortion, stem cell research, family values, defense, and (to a lesser extent) government administration. We also find some common topics of discussion, mainly procedural terms (such as *adjournment* or *unanimous-consent*) and idiosyncratic state-specific economic interests (e.g. *ranching*, *fishing*, *catfish*, and *grazing*).

How (or why) are moderate Senators misclassified?

Our previous analysis suggested that the vocabulary of moderates is a more blurry version of the ideological lexicon used by extreme Senators. In order to shed some more light on this relationship, we analyzed the prediction errors on moderate Senators made by the classifier trained on extreme Senators. To simplify the analysis, we use Boolean representation with a vocabulary limited to nouns. Recall from Table 6 that Boolean-SVM achieved 87.4% prediction accuracy in cross-validation, which is lower than the 96.6% accuracy obtained by *tf*idf*-SVM. However, the Boolean representation uses simple word presence/absence information, which makes it easier to understand which features cause the prediction errors than in the case of *tf*idf*. Thus in each document vector x in the n-dimensional space, x_i is either 1 or 0 for all i.

For each test example x, the SVM classifier computed a *decision value* Y_x as defined in (8), where w and b were computed during the training process.

$$Y_x = w \bullet x + b = \sum_{i=1}^{n} \left(w_i x_i + \frac{b}{n} \right) \tag{8}$$

To compare between examples with different decision values, we normalized the output to either +1 or -1:

$$\frac{Y_x}{|Y_x|} = \sum_{i=1}^{n} \frac{w_i x_i + \frac{b}{n}}{|Y_x|} \tag{9}$$

The value of $\frac{w_i x_i + \frac{b}{n}}{|Y_x|}$ can be understood as the contribution of the feature i to the final decision for x. Thus we can sort the features by their contributions to the decision. For the misclassified examples, we look at the noun features with the largest contributions to understand which features are responsible for the misclassification. Table 10 lists all misclassified moderate examples along with their decision values.

Table 10: Prediction errors in the moderate test case with Boolean noun features

False conservative (10)			False liberal (7)		
108-bayh-in	-1	0.714	108-alexander-tn	+1	-0.011
108-breaux-la	-1	0.445	108-collins-me	+1	-0.177
108-carper-de	-1	0.032	108-dewine-oh	+1	-0.524
108-chafee-ri	-1	0.455	108-mccain-az	+1	-0.317
108-conrad-nd	-1	0.210	108-smith-or	+1	-0.622
108-edwards-nc	-1	0.228	108-snowe-me	+1	-0.874
108-hollings-sc	-1	0.700	108-specter-pa	+1	-0.251
108-lincoln-ar	-1	0.034			
108-nelson-ne	-1	0.552			
108-pryor-ar	-1	0.405			

The table in Appendix A lists the 20 features with the largest contributions to each side for all misclassified examples listed in Table 10. Liberal contributions are listed in the first column. The highlighted features help us to understand why these examples are misclassified. This table shows that some moderate liberal Senators were misclassified as conservative because they talked about some traditionally “conservative issues,” such as bureaucracy, taxation/taxing, marriage, etc., although mentioning these issues, of course, does not necessarily imply support for the conservative position.[27] On the other hand, some moderate conservative Senators were misclassified as liberal because they mentioned "liberal issues" such as children, ecosystem, literacy, shelter for the homeless, etc.

Features may also be ranked with regard to the sum total of the contributions they make to the decisions for a set of *m* documents. For a feature *i*, this value is given by (10):

$$\sum_{k=1}^{m} \frac{w_i x_{ki} + \frac{b}{n}}{|Y_k|} \tag{10}$$

We calculated this value for all nouns in both the conservative and the liberal subsets of the moderate test set. High absolute contributions of "liberal" features (i.e., ones with negative weights in *w*) on the conservative subset indicate the major "misleading" words responsible for misclassification as liberal, and vice versa. Table 12 lists the top 20 features for both classes.

Table 12: Top "misleading" features in 108th moderate documents

Misclassification as conservative	Misclassification as liberal
taxation	gaps
bureaucracies	backgrounds
tactics	hazards
adjournment	outreach
businessmen	optimism
folks	disabilities
commonsense	ecosystem
strings	diseases
appetite	shores
designee	nurses
disincentive	gap

[27] For example, Senator Colman in the 106th Senate had to mention "grievous injury" before he expressed his objection to this amendment to the partial birth ban act.

quote prayers taxing tape hometown cattle insight manpower entity	colleges deaths villages bridges electricity breast illnesses fires mortality

Finally, Table 13 is an example of features with significant difference between the occurrence ratios (conservative:liberal) in the three data sets.

Table 13: An example of word occurrence differences in the three data sets

Word	Training set	Test-extreme	Test-moderate
"taxation"	148:103	21:15	17:16
"taxing"	99:61	17:5	8:6

It is interesting that *tax* and *taxpayers* are not discriminative words since both sides used them equally frequently. In contrast, *taxation* and *taxing*, which emphasize the action of taxing, are highly discriminative – the usage ratio between the extreme conservative and liberal Senators is around 3:2.

The Role of Parties

Our analysis has shown strong support for the hypothesis that Senators express a coherent ideological position in their speeches, in the sense that word patterns used in the past predict future word patterns well, at least in the case of extreme Senators. This suggests that legislative institutions are not responsible for the low dimensionality of voting scores exhibited in the Senate. Extreme Senators already come to those decisions with their ideologies well-formed. Moderate Senators' ideologies, on the other hand, are less well defined and seem to be a blurry version of the positions of extreme Senators.

This leads to the question of what can accounts for the low dimensionality in Senators' ideologies. One possible explanation is that legislators use speeches to communicate with their core electoral constituencies and signal that they are "true believers." If their general electoral constituency is less extreme, this feature can take on the nature of "dog whistle" politics. The core constituency understands the message, the general public may not. Some of our results suggest that these mechanisms are indeed at work. For example, among the most separating adjectives for Democrats we find the word *gay*, for the Republicans we find the word *homosexual*. We consider this a promising route for further exploration.

A key role in the electoral connection is played by parties. Note, however, that here we look at parties as electoral institutions (parties as "brands"). Our analysis casts doubt on the role of parties as legislative institutions (Parties as "legislative Leviathans" - Cox and McCubbins, 1993). To investigate the issue, we conducted a classification experiment on party membership with two categories, Republican and Democrat. We used all Senators in the 101st-107th Senate to train the party membership classifier. The classifier was then used to predict the party membership of the Senators in the 108th Senate. The prediction accuracy is 77.78%, good classification success for party membership compared to vote-based measures. To directly compare classification based on party membership with classification based on ideology, we divided our sample of Senators by their DW-scores at 0, creating a liberal and a conservative category. We then repeated our classification test on that sample. The ideology prediction accuracy is 80.81%. Thus the two classification results are highly similar.

To further investigate the relationship between ideology and party membership, we computed the kappa[28] agreement as a measure of consistency between the Senators' ideology labels and party membership. For our data set (800 Senatorial speech documents for the 101st-108th Congresses), kappa=0.932, which means almost perfect agreement. In other words, the ideology labels and the party membership are not independent.

These results imply that party membership and ideological classification are highly correlated. In return, this suggests that parties play an important in shaping ideological positions during the electoral process. To investigate this issue further, it would be instructive to study the relationship between party and ideology over a longer period, a project we intend to pursue in future research.

Conclusion

The main goal of this study was to examine whether and to what extent Senators' ideological positions, as measured by their DW-NOMINATE scores, can be determined solely based on underlying belief systems (as expressed in their speeches given on the floor) rather than institutional constraints such as agenda control. To this end, we divided the Senators in each of the 101st through 108th Congress into four groups – extreme conservative, moderate conservative, moderate liberal, and extreme liberal – and conducted a series of experiments in which we trained SVM classifiers on speeches from the 101st through 107th Congress and tested their performance on classifying speeches

[28] The kappa coefficient is often used to measure inter-rater agreement in annotation. We followed the kappa computation procedure described at http://www.musc.edu/dc/icrebm/kappa.html.

from the 108^{th} Congress. We reached a high accuracy of 94% in classifying extreme Senators in the 108^{th} Congress based on speeches by extreme Senators in the previous Congresses. In other words, 47 of the 50 extreme Senators in that Congress were classified correctly. The accuracy of the same classifier dropped to 52%, or 26 out of 50, on the "moderate" Senators in the 108^{th} Congress. Through a series of additional experiments, we then showed that moderates are best understood as blurry versions of extreme Senators rather than a distinct category. Together, these findings suggest that agenda control is the wrong explanation for coherent voting behavior in the U.S. Congress. Rather, both voting and debating appear to be different expressions or representations of an underlying belief system. However, we believe that the preceding analysis suggests that parties may in fact be the key factor in shaping these respective belief systems. But contrary to Cox and McCubbins (1993) and Aldrich, Rohde, and Toffias (2004), we have argued that belief systems are determined prior to legislative rules and agenda control in Congress. This suggests that parties would primarily serve as selection mechanisms to aggregate like-minded individuals or exercise any possible influence during the campaign stage.

Given the high success of our classification approach and our feature analysis, we believe that there are indeed distinctive lexica for conservative and liberal lawmakers. Our results demonstrate that the most extreme Senators from both the left and the right discuss different topics, with Democrats focusing on corporate special interests and environmental issues and Republicans focusing on moral and value concerns and (to a lesser extent) on taxes. In addition to providing an alternative explanation for the low dimensionality of voting in the Senate, the word frequency analysis offers at least some preliminary support for some of the arguments recently put forward by Lakoff (2002). Lakoff argues that the Republican and Democratic ideologies reflect different value systems deeply associated with personal views of morality. Our results do indicate a clear tendency of extreme Republicans to distinguish themselves on value issues. Whether our findings can be extended to indeed support Lakoff's theory remains the subject of future research.

Appendix A

The following table lists the 20 features with the most contributions to each side for all misclassified examples listed in Table 10. Conservative contributions are featured first. The highlighted features help us to understand why these examples are misclassified.

Table 11: prediction errors in the moderate test case with Boolean noun features

False conservative errors	False liberal errors
3 108-**bayh**-in -1 0.71392102 **lady beltway strings personality moneys turf payroll marriage marines quote morass prayers guy hometown motives adage exclusion infantry back notification** backgrounds childhood prejudice gaps cynicism counseling racism minorities executives mayors wounds bridges eloquence shores gender specialist segregation diseases ignorance sisters	1 108-**alexander**-tn +1 -0.01083375 bureaucracies taxation appetite taxing adjournment businessmen designee congressmen strings lifestyle sanctity tactics superintendent tape hill theft infringement cattle mandates folks **homelessness sewage backgrounds childhood heating self-sufficiency counseling disabilities infants racism starvation models hunger graduates analyses minorities executives artists optimism nurses**
8 108-**breaux**-la -1 0.44534852 **bureaucracies taxation businessmen standpoint lady personality taking micromanage moneys folks excuses nongermane mercy quote prayers politicians hometown viewpoint expressions contention** hazards backgrounds prejudice breaks shelters descent planet downturn epidemic depression mortgage infants outreach minorities optimism suburbs caregivers neglect giants treatments	15 108-**collins**-me +1 -0.17698836 bureaucracies taxation disincentive adjournment prioritize standpoint designee beltway lifestyle tyranny bombing tactics airmen moneys tape theft titles infringement trucking mandates **homelessness hazards backgrounds gaps breaks breast ecosystem heating shoreline ecosystems medication cops epidemic disabilities depression counterparts outreach models analyses minorities**
11 108-**carper**-de -1 0.032494613	17 108-**dewine**-oh +1 -0.52430643

taxation taxing standpoint lady congressmen beltway strings personality taking airmen moneys theft mandates folks sailors commonsense fitness beef usage payroll hazards backgrounds childhood gaps breaks literacy counseling epidemic infants outreach cutting-edge models hunger executives nurses mayors disparities bridges giants stress	taxing unborn adjournment businessmen prioritize accusation designee congressmen sanctity hike reasoning arrogance tactics irresponsibility theft etc folks habit excuses sailors **homelessness mortality sewage childhood prejudice breast ecosystem shoreline planet medication cops counseling epidemic disabilities depression infants outreach cutting-edge starvation sediment**
12 108-**chafee**-ri -1 0.45455036 **taxation hike tactics titles sailors prayers egg criminal allotments entity relation snow concession malpractice dilemma mom monument recipient endeavors pitch** prejudice breast ecosystem shoreline outreach analyses optimism nurses trauma neglect treatments villages jurisdictions gallons recession illnesses sediments photograph survivors treasures	37 108-**mccain**-az +1 -0.31654823 taxation appetite disincentive taxing bureaucrat adjournment businessmen prioritize standpoint designee congressmen lifestyle spenders tyranny micromanage bombing tactics airmen moneys grab **mortality hazards rent sewage backgrounds prejudice gaps breaks degradation breast ecosystem heating shelters insecurity self-sufficiency ecosystems planet medication cynicism epidemic**
16 108-**conrad**-nd -1 0.20967948 **taxation appetite disincentive taxing adjournment rancher businessmen prioritize standpoint forefathers spenders taking bombing tactics airmen moneys tape irresponsibility cattle folks** rent revitalization recessions gaps heating descent planet medication downturn cops epidemic depression mortgage racism counterparts sediment models graduates analyses executives	43 108-smith-or +1 -0.62176927 taxation taxing bureaucrat unborn bureaucrats accusation lady forefathers lifestyle sanctity taking tyranny micromanage bombing superintendent grass tape hill titles mandates **mortality revitalization childhood prejudice breast insecurity descent literacy self-sufficiency planet medication downturn orientation counseling epidemic relocation disabilities depression mortgage racism**
21 108-**edwards**-nc -1 0.22835342 **taxation tyranny hike bombing tape**	44 108-snowe-me +1 -0.87429835 bureaucracies disincentive unborn rancher

folks excuses commonsense bureaucracy nest prayers politicians commissioners hometown egg motives vacuum contacts back adversaries hazards rent sewage childhood breast insecurity explosions medication downturn cops mortgage cutting-edge graduates minorities executives nurses analysts bridges treatments lung	standpoint designee personality lifestyle tyranny bombing airmen tape titles mandates tendency turf habit sailors ranchers commonsense **mortality hazards rent revitalization backgrounds childhood gaps degradation breast ecosystem heating shoreline shelters literacy self-sufficiency ecosystems planet medication downturn counseling**
28 108-**hollings**-sc -1 0.69966938 **taxing lady forefathers personality superintendent titles trucking mandates folks turf beef usage payroll marriage launch liberals politicians commissioners wheat hometown** hazards gaps shelters insecurity descent planet relocation models analyses trauma analysts riders giants locks collaboration jurisdictions gallons recession researchers waterways	45 108-specter-pa +1 -0.25121877 taxation adjournment businessmen designee beltway personality lifestyle sanctity tyranny bombing tactics airmen moneys tape theft turf sailors bureaucracy reversal usage **mortality hazards rent childhood degradation breast shelters literacy medication downturn counseling epidemic disabilities mortgage outreach cutting-edge graduates analyses minorities executives**
35 108-**lincoln**-ar -1 0.034349571 **taxation disincentive adjournment prioritize forefathers tyranny reasoning tactics superintendent titles mandates folks excuses commonsense fitness beef payroll desires marriage marines** homelessness mortality hazards backgrounds gaps breast ecosystem heating shelters insecurity ecosystems medication orientation counseling epidemic disabilities depression mortgage infants counterparts	
41 108-**nelson**-ne -1 0.55242801 **appetite disincentive standpoint lady strings lifestyle sanctity taking bombing reasoning arrogance tactics**	

superintendent airmen moneys theft cattle folks sailors commonsense mortality hazards revitalization backgrounds childhood degradation ecosystem shelters self-sufficiency planet medication cops counseling epidemic disabilities mortgage racism outreach starvation models	
42 108-**pryor**-ar -1 0.40516298 **standpoint personality tyranny airmen theft titles mandates folks sailors commonsense fitness marriage marines nest foresight mankind quote prayers wheat hometown** hazards rent prejudice breast medication cops counseling epidemic disabilities mortgage counterparts cutting-edge minorities artists optimism nurses analysts perspectives stress vaccines	

References:

Aldrich, John H. 1995. *Why Parties?: The Origin and Transformation of Party Politics in America*. Chicago: The University of Chicago Press.

Aldrich, John H., David W. Rohde and Michael W. Tofias. 2004. Examining Congress with a Two-Dimensional Political Space. Paper presented at the American Political Science Association Meeting, Chicago, IL.

Baumgartner, Frank R., and Bryan D. Jones, eds. 2002. Policy Dynamics. Chicago: University of Chicago Press.

Baumgartner, Frank R., and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. Chicago: University of Chicago Press.

Baumgartner Frank R. and Bryan D. Jones. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago: University of Chicago Press.

Anthony M. Bertelli and Christian R. Grose. 2006. "The Spatial Model and the Senate Trial of President Clinton." American Politics Research Vol. 34, No. 4, 535-559

Bertelli, Anthony M. and Christian R. Grose. 2007. "Agreeable Administrators? The Public Positions of Cabinet Secretaries and Presidents." *Presidential Studies Quarterly*.

Benoit, Kenneth, Michael Laver, Christine Arnold, Paul Pennings, andMadeleine O. Hosli. 2005 (forthcoming). Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Wordscoring." *European Union Politics*.

2003. Benoit, Kenneth and Michael Laver. ``Estimating Irish Party Positions Using Computer Wordscoring: The 2002 Elections." *Irish Political Studies* 18(1): 97-107.

2005. Benoit, Kenneth and Michael Laver. ``Mapping the Irish Policy Space: Voter and Party Spaces in Preferential Elections." *Economic and Social Review* 36(2, Summer/Autumn): 83-108.

Black, D. (1958). *The theory of committees and elections*. Cambridge [Eng.]: University Press.

Budge, I., Pennings, P., 2007. Do they work? Validating computerized word frequency estimates against policy series. Electoral Studies 26 (1), 121e129.

Budge, I., Klingemann, H. D., Volkens, A., Bara, J. and Tanenbaum, E. (2001) *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998* (Oxford : Oxford University Press).

Burns, P. (2006). MorphAdorner: Morphological Adorner for English Text. (c) 2006-2007 by Northwestern University.

Chang, C. C., & Lin, C. J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cohen, W., & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems*, 17(2),141–173.

Cortes, C., & Vapnik V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Davis, O. A., Hinich, M. J., & Ordeshook, P. C. (1970). An expository development of a mathematical model of the electoral process. *The American Political Science Review, 64*(2), 426-448.

Davis, O.A., and Hinich, M. (1966).A mathematical model of policy formation in a democratic society. In Bernd, J. (Ed.), *Mathematical Applications in Political Science II,* 175–208 Southern Methodist University Press: Dallas, TX.

Downs, A. (1957). *An economic theory of democracy*. New York,: Harper.

Finn, A., & Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of American Society of Information Science and* Technology, 57(11), 1506-1518.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, edited by D.E. Apter. New York: Free Press.

Cousins, Ken, and Wayne McIntosh. 2005. "More than Typewriters, More than Adding Machines: Integrating Information Technology into Political Research." *Quality and Quantity* 39: 591-614.

Cox, Gary W., and Mathew D. McCubbins. 1993. Legislative Leviathan: Party Government in the House. Berkeley: University of California Press.

Cox, Gary W., and Mathew D. McCubbins. 2005. Setting the Agenda: Responsible Party Government in the US House of Representatives. Cambridge University Press.

Dave, K, Lawrence, S, & Pennock, D. M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, 519-528. Retrieved May 28, 2007, from ACM Digital Library.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the 7th international conference on information and knowledge management*, 48-155. Retrieved May 28, 2007, from ACM Digital Library.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3), 389–422.

Joachims, T. (1998). Text categorization with Support Vector Machines: learning with many relevant features. *10th European Conference on Machine Learning, volume 1398 of Lecture Notes in Computer Science*, 137-142. Berlin: Springer Verlag.

Joachims, T. (2004). SVM^{light}: Support Vector Machine (Version 6.01). Software available at http://svmlight.joachims.org/.

Gerner, Deborah J., Philip A. Schrodt, Ronald A. Francisco, and Judith L. Weddle. 1994. "The Machine Coding of Events from Regional and International Sources." *International Studies Quarterly* 38:91-119.

Imbeau, L. M. 2005. “Policy Discourse, Fiscal Rules, and Budget Deficit: A Median Voter Model.” Durham, UK, Annual meeting of the European Public Choice Society.

Heckman, J. J., & Snyder, J. (1997). Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics, 28*, S142-S189.

Kiewit, Roderick D., and McCubbins, Matthew D. 1991. *The Logic of Delegation: Congressional Parties and the Appropriation Process.* Chicago: University of Chicago Press.

Lakoff, George. 2002. Moral Politics: How Liberals and Conservatives Think. 2nd Edition. Chicago, IL: The University of Chicago Press.

Laver, Michael and Kenneth Benoit. 2002. ``Locating TDs in Policy Spaces: Wordscoring Dáil Speeches." *Irish Political Studies* 17(1, Summer): 59-73.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. “Extracting Policy Positions from Political Texts Using Words as Data.” *American Political Science Review* 97, 2: 311-337.

Laver, Michael, and John Garry. 2000. “Estimating Policy Positions from Political Texts.” *American Journal of Political Science* 44: 619-34.

Leopold, E., & Kindermann, J. (2002). Text categorization with support vector machines. how to represent texts in input space? *Machine Learning*, 46(1-3), 423–444.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the 15th annual international conference on research and development of information retrieval,* 37–50. Retrieved May 28, 2007, from ACM Digital Library.

McCarty, Nolan, Keith T. Poole, and Howard Rosenthal. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. Boston: MIT Press.

McCarty, Nolan, Keith Poole and Howard Rosenthal. 1997. *Income Redistribution and the Realignment of American Politics.* Washington D.C.: American Enterprise Institute.

Monroe, Burt L., and Ko Maeda. 2004. "Rhetorical Ideal Point Estimation: Mapping Legislative Speech." Society for Political Methodology, Stanford University, Palo Alto.

Moschitti, A. & Basili, R. (2004). Complex linguistic features for text classification: a comprehensive study. *European conference on information retrieval, volume 2997 of Lecture Notes in Computer Science*, 181–196. Berlin: Springer Verlag

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on empirical methods in natural language processing*, 79–86. Retrieved May 28, 2007, from ACM Digital Library.

Poole, K. T., & Rosenthal, H. (1991). Patterns of Congressional Voting. *American Journal of Political Science, 35*(1), 228-278.

Poole, Keith T., and Howard Rosenthal. 1997 Congress: A Political-Economic History of Roll Call Voting. New York: Oxford

Poole, Keith T., and Howard Rosenthal. 2007. Ideology and Congress. Transaction Publisher.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting*. Cambridge: Cambridge University Press.

Poole, Keith T. 2003. Changing Minds? Not in Congress. Unpublished Manuscript.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

Purpura, S., & Hillard, D. (2006). Automated classification of congressional legislation. *Proceedings of the 2006 international conference on Digital government research*, 219-225. Retrieved May 28, 2007, from ACM Digital Library.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2006. "An Automated Method of Topic-Coding Legislative Speech Over Time with Application to the 105th-108th U.S. Senate." Unpublished Manuscript.

Rohde, David W. (1991) Parties and Leaders in the Postreform House (Chicago, Ill.: University of Chicago Press).

D. Rohde and John Aldrich, *The Consequences of Party Organization in the House: The Role of the Majority and Minority Parties in Conditional Party Government*, in Polarized Politics: Congress and the President in a Partisan Era, edited by Jon Bond and Richard Fleischer (2000), pp. 3172, CQ

D. Rohde and John H. Aldrich, *The Logic of Conditional Party Government: Revisiting the Electoral Connection*, in Congress Reconsidered, 7th Edition, edited by Lawrence Dodd and Bruce Oppenheimer (2001), pp. 26292, CQ Press.

Schonhardt-Bailey, Cheryl. 2006. *From the Corn Laws to Free Trade: Interests, Ideas, and Institutions in Historical Perspective.* Cambridge: MIT Press.

Scott, S. and Matwin, S. (1999). Feature engineering for text classification. *Proceeding s of the 16th international conference on machine learning* (pp. 379-388). San Francisco: Morgan Kaufmann.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Simon, A. F. and M. Xenos. 2004. “Dimensional reduction of word-frequency data as a substitute for intersubjective content analysis.” *Political Analysis* 12:63-75.

Thomas, M., Pang, B. and Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of the 2006 conference on empirical methods in natural language processing,* 327–335. Retrieved from ACL Digital Archive.

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vapnik, V. (1982). *Estimating of Dependencies Based on Empirical Data.* New York: Springer-Verlag.

Yang, Y., & Liu, X.. A re-evaluation of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, 42–49. Retrieved May 28, 2007, from ACM Digital Library.

Yu, B., Kaufmann, S., Godbout, J-F, & Daniel Diermeier. Ideological vocabulary extraction and position prediction. Unpublished manuscript, Northwestern University 2007