B.J.Pol.S. **42**, 31–55 *Copyright © Cambridge University Press, 2011*
doi:10.1017/S0007123411000160
First published online 23 May 2011

Language and Ideology in Congress

DANIEL DIERMEIER, JEAN-FRANÇOIS GODBOUT, BEI YU AND STEFAN KAUFMANN*

Legislative speech records from the 101st to 108th Congresses of the US Senate are analysed to study political ideologies. A widely-used text classification algorithm – Support Vector Machines (SVM) – allows the extraction of terms that are most indicative of conservative and liberal positions in legislative speeches and the prediction of senators' ideological positions, with a 92 per cent level of accuracy. Feature analysis identifies the terms associated with conservative and liberal ideologies. The results demonstrate that cultural references appear more important than economic references in distinguishing conservative from liberal congressional speeches, calling into question the common economic interpretation of ideological differences in the US Congress.

Over at least fifty years of research, *ideology* has been used to explain the political behaviour of voters, legislators and other political agents. Ideologies give structure to an individual's view on various issues. Intuitively, a political ideology specifies which issue positions go together, the 'knowledge of what-goes-with-what'.[1] Ideologies do not necessarily imply a logically consistent political, economic or social world view. Indeed, as Converse argued, the association between issues may just be contingent and reflective of a particular, perhaps cultural or historical, experience.[2] Nevertheless, ideologies constrain. It is quite unlikely (though not impossible) that a randomly selected American 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.'[3] Measuring ideological orientations, however, has always been a difficult task. Unlike party affiliation, for example, ideology is not directly observable. Consequently, scholars have employed different strategies to measure ideological positions, ranging from survey responses to statistical estimates based on voting records. In legislative politics, and especially for the

* 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 (email: d-diermeier@kellogg.northwestern.edu); Department of Political Science, University of Montreal; School of Information Studies, Syracuse University; and Department of Linguistics, Northwestern University, respectively. The authors wish to thank seminar participants at the annual meetings of the American Political Science Association and the Midwest Political Science Association, as well as the members of the Institutions, Organizations and Growth research group at the Canadian Institute for Advanced Research (CIFAR) for their helpful comments. Financial support from the Ford Motor Company Center for Global Citizenship, Kellogg School of Management, Northwestern University, and CIFAR is gratefully acknowledged.

[1] Keith T. Poole, 'Changing Minds? Not in Congress', *Public Choice*, 131 (2007), 435–51.

[2] Philip E. Converse, 'The Nature of Belief Systems in Mass Publics', in David E. Apter, ed., *Ideology and Discontent* (New York: The Free Press, 1964), pp. 206–61.

[3] Converse, 'The Nature of Belief Systems and Mass Publics', p. 207.

US Congress, the most widely used measure of ideology remains the vote-based score developed and refined by Poole and Rosenthal and later by McCarty, Poole and Rosenthal.[4] 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. Perhaps the most important finding of the Poole–Rosenthal approach is that much of the voting behaviour in Congress can be explained by a stable, low-dimensional issue space. Indeed, Poole and Rosenthal find that between 1789 and 1985, a two-dimensional spatial model (estimated with NOMINATE scores) can correctly classify about 85 per cent of the individual voting decisions of each member of Congress.[5] Moreover, for most periods of American history, a single dimension is sufficient to explain nearly all of the variance in voting. For example, in recent years (104th–106th Congress) a single dimension can account for about 90 per cent of all roll-call choices by members of Congress.[6]

Though by no means without controversy, the finding of low dimensionality has been recognized as an important characteristic of congressional decision making.[7] Yet, many questions remain. First, there is an extensive and ongoing debate on whether the voting patterns reflect a legislator's ideological position or whether the low-dimensionality is an artefact of legislative agenda control, the need to preserve party brands, or constituency concerns and the like.[8] Indeed, party cohesion, pressure by party whips and strategic

[4] Keith T. Poole and Howard Rosenthal, 'Patterns of Congressional Voting', *American Journal of Political Science*, 35 (1991), 228–78; Keith T. Poole and Howard Rosenthal, *Congress: A Political-Economic History of Roll Call Voting* (New York: Oxford University Press, 1997); Nolan McCarty, Keith T. Poole and Howard Rosenthal, *Income Redistribution and the Realignment of American Politics* (Washington, D.C.: American Enterprise Institute, 1997); Nolan McCarty, Keith T. Poole and Howard Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches* (Boston, Mass.: MIT Press, 2006); Keith T. Poole and Howard Rosenthal, *Ideology and Congress* (New Brunswick, N.J.: Transaction Publishers, 2007).

[5] Poole and Rosenthal, *Congress*.

[6] Keith T. Poole, *Spatial Models of Parliamentary Voting* (New York: Cambridge University Press, 2005).

[7] Initially, this finding met with widespread disbelief. See Poole and Rosenthal, *Congress,* p. 8. However, the low-dimensionality of legislative voting has been confirmed by other scholars using different estimation methodologies, such as Bayesian procedures (Joshua Clinton, Simon Jackman and Doug Rivers, 'The Statistical Analysis of Roll Call Data', *American Political Science Review*, 98 (2004), 355–70) or factor analysis (James J. Heckman and James M. Snyder Jr, 'Linear Probability Models of the Demand for Attribution with an Empirical Application to Estimating the Preferences of Legislators', *RAND Journal of Economics*, 28 (1997), S142–89) for estimating ideal points.

[8] Institutional features such as gate-keeping powers of committees (Kenneth A. Shepsle and Barry R. Weingast, 'Structure-induced Equilibrium and Legislative Choice', *Public Choice*, 37 (1981), 503–19), pre-floor legislative activities (such as co-sponsorship), strategic voting (Jeffery C. Talbert and Matthew Potoski, 'Setting the Legislative Agenda: The Dimensional Structure of Bill Cosponsoring and Floor Voting', *Journal of Politics*, 64 (2002), 864–91) or institutional constraints such as the presidential veto (Jason M. Roberts, 'The Statistical Analysis of Roll Call Data: A Cautionary Tale', *Legislative Studies Quarterly*, 22 (2007), 341–60; Joshua D. Clinton, 'Lawmaking and Roll Calls', *Journal of Politics*, 69 (2007), 457–69) can all affect the measurement of ideal points and reduce the dimensionality of legislative voting in Congress. It is also possible that exogenous factors, such as electoral incentives, could help explain why parties aim to present a coherent legislative agenda, and avoid intra-party voting divisions. Indeed, Snyder and Ting (James M. Snyder and Michael M. Ting, 'An Informational Rationale for Political Parties', *American Journal of Political Science*, 46 (2002), 90–110; James M. Snyder and Michael M. Ting, 'Party Labels, Roll Calls, and Elections', *Political Analysis*, 11 (2003), 419–44) and Woon and Pope (Jonathan Woon and Jeremy C. Pope, 'Made in Congress? Testing the Electoral Implications of Party Ideological Brand Names', *Journal of Politics*, 70 (2008), 823–36) argue that parties can use their aggregate roll-call record to produce a coherent ideological brand name in order to communicate with the

voting have been shown significantly to alter the measurement of ideal points in Westminster-style parliamentary systems, and there is at least some evidence that this could also affect the scaling of legislative votes in the US Congress.[9]

Secondly, putting aside these controversies, we would want to know the content of the estimated ideological positions. This is of particular importance from a comparative point of view: for example, to understand the differences between 'conservatism' in the American context and in other political systems.

Poole and Rosenthal and others have hypothesized that the first dimension represents a traditional left–right conflict associated with the government's role in the economy and economic redistribution, while the second dimension represents issues that have the potential to divide both parties internally, like regional conflicts over slavery, and later racial and civil rights. Moreover, Poole has argued that the waning importance of the second dimension after the civil rights reform of the 1960s suggests that race-related issues are now largely correlated with questions of economic redistribution, and this finding has been confirmed by survey data and correlations with interest-group ratings.[10] Other detailed studies of specific debates have indicated the importance of additional issue dimensions in legislative conflicts that are not necessarily related to economic or redistributive questions.[11] In addition, the lack of comparable data makes both a comparative and an inter-temporal analysis of the content of ideologies infeasible.

Our goal in this article is to investigate systematically the claims regarding the *content* of congressional ideologies. We do not aim to add to the literature examining the validity of the Poole and Rosenthal measures, but rather to take these measures as given and then investigate what they signify. In other words, we are not taking a stance on whether these ideological orientations reflect the personal belief systems of elected politicians or whether they are the consequence of agenda control, constituency constraints, interest-group influence, career concerns or any of the other factors suggested in the literature. But we assume that a legislator's 'effective' ideology, as measured by NOMINATE scores, ultimately influences political behavior, such as his or her legislative voting record, even if it is shaped and constrained by other factors.

In contrast to Poole and Rosenthal and most of the existing literature on legislative ideologies, we are using congressional speech rather than voting behaviour, because it is a

(F'note continued)

electorate. In this context, the observed unidimensionality in legislative voting would be facilitated by electoral incentives, rather than by institutional rules or agenda control.

9 On the Westminster style parliamentary systems, see Spirling and McLean (Arthur Spirling and Iain McLean, 'UK OC OK? Interpreting Optimal Classification Scores for the U.K. House of Commons', *Political Analysis*, 15 (2006), 85–6). On the US Congress case, see Clinton, 'Lawmaking and Roll Calls', and Roberts, 'The Statistical Analysis of Roll Call Data'.

10 See, for example, the NPAT candidate survey of Stephen Ansolabehere, James M. Snyder Jr and Charles Stewart III, 'The Effects of Party and Preferences on Congressional Roll-Call Voting', *Legislative Studies Quarterly*, 26 (2001), 533–72, which looks at the correlation between first factor NOMINATE and first factor NPAT scores; or the Poole and Rosenthal study of NOMINATE scores and interest group ratings (Poole and Rosenthal, *Congress*; Poole and Rosenthal, *Ideology and Congress*).

11 One such example is Cheryl Schonhardt-Bailey, 'The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion', *British Journal of Political Science*, 38 (2008), 383–410. In her study of the US Senate debates on partial-birth abortion, Schonhardt-Bailey identifies two dimensions of conflict, where the first dimension represents an emotive conflict over the abortion procedure, while the second dimension is related to the constitutionality of the bill. Schonhardt-Bailey argues that legislative voting correlates with this second dimension.

richer source of data with which to study the content of political ideologies. While our focus here is on the US Congress, our methods can also be used to compare ideological content over time or across political systems. Political text has been a neglected 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. Traditionally, scholars had to rely on labour-intensive coding methodologies to study political texts, party platforms and campaign speeches.[12] Recent advances in computational linguistics, however, have opened up new avenues for analysing political language in various domains and contexts.[13]

Researchers have also begun to apply automated text analysis in order to measure ideological positions. For instance, Laver, Benoit and Garry took a semi-automated approach to analysing European party manifestos: first, by using domain experts to place a few parties as reference points on a left–right ideological scale; and then by extracting word frequencies from their respective party manifestos. Other parties' ideological positions were then estimated from how closely their party manifestos' word frequency distributions matched the reference distributions.[14] Monroe and Maeda developed alternative approaches, as did Slapin and Proksch.[15] These authors use statistical techniques similar to Poole and Rosenthal's to

[12] Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum, *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998* (Oxford: Oxford University Press, 2001); Frank R. Baumgartner and Bryan D. Jones, *Agendas and Instability in American Politics* (Chicago: University of Chicago Press, 1993); Frank R. Baumgartner and Bryan D. Jones, eds, *Policy Dynamics* (Chicago: University of Chicago Press, 2002); Frank R. Baumgartner and Bryan D. Jones, *The Politics of Attention: How Government Prioritizes Problems* (Chicago: University of Chicago Press, 2005).

[13] For examples, see Michael Laver and Kenneth Benoit, 'Locating TDs in Policy Spaces: Wordscoring Dáil Speeches', *Irish Political Studies*, 17 (2002), 59–73; Michael Laver, Kenneth Benoit and John Garry, 'Extracting Policy Positions from Political Texts Using Words as Data', *American Political Science Review*, 97 (2003), 311–37; Kenneth Benoit and Michael Laver, 'Estimating Irish Party Positions Using Computer Wordscoring: The 2002 Elections', *Irish Political Studies*, 18 (2003), 97–107; Kenneth Benoit and Michael Laver, 'Mapping the Irish Policy Space: Voter and Party Spaces in Preferential Elections', *Economic and Social Review*, 36 (2005), 83–108; Burt L. Monroe and Ko Maeda, 'Rhetorical Ideal Point Estimation: Mapping Legislative Speech' (presented at the Society for Political Methodology, Palo Alto: Stanford University, 2004); Adam F. Simon and Michael Xenos, 'Dimensional Reduction of Word-frequency Data as a Substitute for Intersubjective Content Analysis', *Political Analysis*, 12 (2004), 63–75; Jonathan B. Slapin and Sven O. Proksch, 'A. Scaling Model for Estimating Time-Series Party Positions from Texts', *American Journal of Political Science*, 52 (2008), 705–22; Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev, 'How to Analyze Political Attention with Minimal Assumptions and Costs', *American Journal of Political Science*, 54 (2010), 209–28. For a recent review, see Burt Monroe and Philipp A. Schrodt, 'Introduction to the Special Issue: The Analysis of Political Text', *Political Analysis*, 16 (2008), 351–5; and also Ken Cousins and Wayne McIntosh, 'More than Typewriters, More than Adding Machines: Integrating Information Technology into Political Research', *Quality and Quantity*, 39 (2005), 591–614; and Tae Yano, William W. Cohen and Noah A. Smith, 'Predicting Response to Political Blog Posts with Topic Models', *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics conference (NAACL)* (2009), 477–85.

[14] Examples include Laver, Benoit and Garry, 'Extracting Policy Positions from Political Texts Using Words as Data'. See also Benoit and Laver, 'Estimating Irish Party Positions Using Computer Wordscoring'; Benoit and Laver, 'Mapping the Irish Policy Space'; Laver and Benoit, 'Locating TDs in Policy Spaces; Kenneth Benoit, Michael Laver, Christine Arnold, Paul Pennings and Madeleine O. Hosli, 'Measuring National Delegate Positions at the Convention on the Future of Europe Using Computerized Wordscoring', *European Union Politics*, 6 (2005), 291–313. For a critical view, see Ian Budge and Paul Pennings, 'Do They work? Validating Computerized Word Frequency Estimates against Policy Series', *Electoral Studies*, 26 (2007), 121–9.

[15] Monroe and Maeda, 'Rhetorical Ideal Point Estimation'; Slapin and Proksch, 'A Scaling Model for Estimating Time-Series Party Positions from Texts'.

estimate 'rhetorical ideal points'. In contrast to Poole and Rosenthal, their choice space is a matrix of word counts rather than vote counts for each legislator. The use of text data creates various statistical and computational difficulties, but it does allow researchers to estimate ideological positioning even in the absence of voting data – as long as reliable text documents, such as party manifestos or speech records, are available.

This line of research, however, is in the tradition of estimating ideal points from roll-call data, with voting behaviour replaced by 'text as data'. That is, 'unstructured data' (i.e., text) are translated into 'structured data' (i.e., numbers) that can then be further analysed by various quantitative methods. Our emphasis here is less on measurement. Rather, our focus is on providing quantitative approaches that can be used to understand the content of these ideologies better.

More precisely, we will use supervised learning algorithms, Support Vector Machines (SVMs), to analyse the ideological content of legislative speech. The approach works as follows: a group of reference texts is selected 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. The approach then allows us to identify the most relevant speech features that drive the classification, which will provide us with additional insight into the content of political ideology.

There are only a few applications of this method in the political domain. Examples include Purpura and Hillard, who used supervised learning techniques to identify speech topics in congressional legislation, and Thomas, Pang and Lee, who investigated whether speech classification of floor debates in the House of Representatives on a specific bill can be used as a predictor of subsequent agreement.[16]

Our focus will be on the study of congressional ideologies using congressional legislative speech data. Ideologies are measured by DW-NOMINATE scores, the most common measure of legislative behaviour.[17] Notice that we are not interested in providing alternative approaches for locating legislators in an ideological space. Rather, we are using pre-existing ideological measures (which are based on voting behaviour) to shed light on the internal structure of political ideologies. Our approach is not an assessment of these measures, but an investigation into their content. Our first goal is to investigate the validity of our approach. That is, can we indeed use a classifier trained on past speech to classify future speech? Once the validity of our approach is established, we can then use it to examine the content of the identified ideological segments.

In this study, we analyse the ideological content of speeches made in the US Senate between 1989 and 2004, fully capturing the 101st to 108th Congresses. We use Senate rather

[16] Stephen Purpura and Dustin Hillard, 'Automated Classification of Congressional Legislation', *Proceedings of the 2006 International Conference on Digital Government Research* (2006), 219–25, retrieved 28 May 2007, from the ACM Digital Library; Bo Pang, Lillian Lee and Shivakumar Vaithyanathan, 'Thumbs up? Sentiment Classification Using Machine Learning Techniques', *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, (2002), 79–86, retrieved 28 May 2007, from the ACM Digital Library; Matt Thomas, Bo Pang and Lillian Lee, '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* (2006), 327–35, retrieved from the ACL Digital Archive, predicted speakers' opinions about a specific bill (support or opposition) based on their speeches. Their classifier was trained on 2,740 speech segments in 38 bill debates and achieved an accuracy of 66 per cent in predicting the opinions expressed in 860 speech segments from ten different legislative debates.

[17] We used the Poole and Rosenthal DW-NOMINATE SCORES available at http://voteview.com/dwnomin.htm.

than House speeches because Senate speeches are considerably longer and unconstrained by the germaneness rule, which permits a less restricted expression of political ideologies.[18]

We proceed as follows. First, we rank the senators in each Congress according to their DW-NOMINATE scores. Then, we use the speeches made by the twenty-five most liberal and twenty-five most conservative senators in the 101st to 107th Congresses as our reference set to train an SVM classifier.[19] Next, we employ this classifier to predict the ideological positions of the twenty-five most liberal and twenty-five most conservative senators in the 108th Congress.[20] Our choice of more extreme senators is guided by the hypothesis that their ideological positions are more clearly defined. We will later investigate whether this is indeed the case by studying moderate senators.

Notice that our approach goes beyond testing whether the ideology of a given senator is consistent over time,[21] as the twenty-five most conservative and twenty-five most liberal senators may or may not be identical to the list of senators used for the training set.[22] Rather, if successful, our method will identify characteristics of political speech that are consistent over time and shared by members with similar ideologies.

In the next section, we briefly discuss our methodology, followed by a section on data preparation. We then outline the details of the estimations and our results. Given the very high success rate in correctly classifying senators over the relevant time frame, we then identify the features (or words) that drove the classification. One of the most important findings of this analysis is the importance of value concepts related to cultural controversies such as abortion and gay rights. We then apply our approach to moderate senators. This provides not only an additional validity check, but also an examination of the difference between speeches by moderate and extreme senators, while at the same time offering deeper insights into the content of congressional ideologies. In the discussion section, we consider various methodological issues, extensions, and consequences for the study of political ideologies and parties. This is followed by a brief conclusion. For the interested reader, Appendix A offers additional details on our methodology, while Appendix B provides supplementary estimation results.

METHODOLOGY: TEXT CLASSIFICATION ALGORITHMS AND SUPPORT VECTOR MACHINES

Many supervised learning algorithms have been used for text document classification. On the basis of the performance in previous classification tasks,[23] Support Vector Machines (SVMs)

[18] We will discuss differences between the House and the Senate below. We will also suggest how the approach can be utilized when studying other legislatures.

[19] The DW-NOMINATE scores for the same senators can be different across congresses. As a result, when we prepare the senatorial speeches as training and testing documents (each document is called an 'example' in machine learning terms), a senator could be assigned to the extreme category in one congress but moved to the moderate category in another. Therefore, we treat the same senators in different congresses as different training/testing examples.

[20] Forty-five of these fifty 'extreme' senators had already served in the 107th Congress.

[21] This issue was investigated by Poole ('Changing Minds? Not in Congress') in the context of voting behaviour. Poole found strong support for individual ideological consistency in members of Congress over time.

[22] Ninety-one senators in the 108th Congress served in previous congresses. Forty-four of the fifty extreme senators in the 108th Congress were rated as extreme in previous congresses.

[23] The performance of classification algorithms is tested using common benchmark datasets. The Reuters-21578 news collection, the OHSUMED Medline abstract collection, and the 20 Usenet newsgroups collection are the most widely used benchmark datasets. The Reuters-21578 collection is

have been identified as one of the most effective text classification and feature selection methods.[24] We chose accuracy as our evaluation criterion, defined as the proportion of correct predictions among all predictions. SVM-based text classification accuracies can vary widely between different tasks. They may be as high as over 95 per cent for some topic categories in news articles, or as modest as around 75 per cent for opinion classification of movie reviews.[25] A classifier is usually considered effective when the accuracy is higher than a baseline method, which can be something as simple as a random guess for balanced datasets or a majority of cases for skewed data sets.

In our analysis, we have a binary classification problem with 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 linguistic feature deemed relevant to the classification task.[26] In the training phase, the category membership of each data point is given. In our case, it would be a 'conservative' or 'liberal' label based on the legislator's DW-NOMINATE score. To simplify the notation, we label one category as '+1' and the other as '−1'. Thus, a training set of l examples is represented as a set of 'vector and label' 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.[27] If the data points in the positive and negative 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

(F'note continued)
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.html.

[24] Susan Dumais, John Platt, David Heckerman and Mehran Sahami, 'Inductive Learning Algorithms and Representations for Text Categorization', *Proceedings of the 7th International Conference on Information and Knowledge Management* (1998), 48–155, retrieved 28 May 2007, from the ACM Digital Library; Isabelle Guyon, Jason Weston, Stephen Barnhilland, Vladimir Vapnik, 'Gene Selection for Cancer Classification Using Support Vector Machines', *Machine Learning*, 46 (2002), 389–422; George Forman, 'An Extensive Empirical Study of Feature Selection Metrics for Text Categorization', *Journal of Machine Learning Research*, 3 (2003), 1289–305; Thorsten Joachims, 'Text Categorization with Support Vector Machines: Learning with Many Relevant Features', *10th European Conference on Machine Learning, Vol. 1398 of Lecture Notes in Computer Science* (Berlin: Springer Verlag, 1998), pp. 137–42; Dunja Mladenic, Janez Brank, Marko Grobelnik and Natasa Milic-Frayling, 'Feature Selection Using Linear Classifier Weights: Interaction with Classification Models', *Proceedings of the 27nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '04), (Sheffield: 25–29 July 2004), pp. 234–41; Yiming Yang and Xin Liu, 'A Re-evaluation of Text Categorization Methods', *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1999), 42–9, retrieved 28 May 2007, from the ACM Digital Library; Fabrizio Sebastiani, 'Machine Learning in Automated Text Categorization', *ACM Computing Surveys*, 34 (2002), 1–47. 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.

[25] Pang, Lee and Vaithyanathan, 'Thumbs up?'

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

[27] Vladimir Vapnik, *Estimating of Dependences Based on Empirical Data* (New York: Springer-Verlag, 1982); Corinna Cortes and Vladimir Vapnik, 'Support-vector Networks', *Machine Learning*, 20 (1995), 273–97; Vladimir Vapnik, *The Nature of Statistical Learning Theory* (New York: Springer-Verlag, 1999).

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* (SVs). 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 maximized. This is achieved by computing the parameters w and b in the linear decision function $y = w \bullet x + b$. 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 the relative location of this point to the maximally separating hyperplane, which is 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 insights about the informativeness of each feature in determining category membership. 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. We will use this feature-ranking method to find the most informative word indicators of liberal and conservative ideologies. More mathematical details of SVM can be found in Appendix A.[28]

DOCUMENT REPRESENTATION

As noted above, in our text classification model, documents are represented in a vector space whose dimensions correspond to the linguistic 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).

The BOW model is insensitive to many potentially informative properties of documents, such as the location of words in the document, grammatical relations, the internal structure of the document at various levels (for example, paragraph and sentence boundaries) and multi-word phrases such as 'war on drugs'. Some researchers have tested more sophisticated document representations which incorporate such information (for example, relative word position and syntactic structure) in a variety of classifiers. However, experimental results showed that these complex features did not improve the classification performance significantly.[29] Therefore, we use the BOW approach in this study.

[28] There are several efficient implementations of the SVM algorithm, such as LIBSVM and SVM^{light} (Thorsten Joachims, 'SVM^{light}: Support Vector Machine (Version 6.01)', (2004)). We used the SVM^{light} package with its default setting in this study. See Chih C. Chang and Chih J. Lin, 'LIBSVM: A Library for Support Vector Machines' (2001), Software available at http://www.csie.ntu.edu.tw/ ~ cjlin/libsvm.

[29] David D. Lewis, '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* (1992), pp. 37–50, retrieved 28 May 2007, from the ACM Digital Library; William W. Cohen and Yoram Singer, 'Context-sensitive Learning Methods for Text Categorization', *ACM Transactions on Information Systems*, 17 (1999), 141–73; Sam Scott and Stan Matwin, 'Feature Engineering for Text Classification', *Proceedings of the 16th International Conference on Machine Learning* (San Francisco: Morgan Kaufmann, 1999), pp. 379–88; Alessandro Moschitti and Roberto Basili, 'Complex Linguistic Features for Text Classification: A Comprehensive Study', *European Conference on Information Retrieval, Vol. 2997 of Lecture Notes in Computer Science* (Berlin: Springer Verlag, 2004), pp. 181–96.

The value of each word type in the feature space can be determined in various ways: It may simply record the word's presence or absence in the document, leading to a Boolean representation, or it can list the word's frequency of occurrence in the document. In turn, this frequency can be normalized by document length or weighted by some other 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, the *tf*idf* value of a word with term frequency t_f in a document of length l and document frequency d_f is given by the formula:

$$tf^*\,idf = (t_f/l)\log(n/d_f). \tag{2}$$

The *tf*idf* weighting applies only to the training data. In the prediction step, the test examples are completely independent from each other, and the total number of test documents is assumed to 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 use of computational linguistics to study political behaviour is a fairly new endeavour. It may, therefore, be useful to discuss performance criteria in more detail. As our article is the first to use SVM in classifying 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 for movie or restaurant reviews. In this last application, the goal is to correctly classify reviews as 'positive' or 'negative'. Opinion classifiers have achieved accuracy levels as high as 88 per cent for product reviews and 82 per cent for movie reviews.[30] However, Finn and Kushmerick found that an opinion classifier trained on movie reviews was not effective in predicting the polarity of restaurant reviews, and vice versa.[31]

This domain dependence suggests that classification methods can be used to assess coherence across different types of documents. In the political context, we are not dealing with product domains, but different issue domains. A highly coherent ideology will view different issues in the same way. In other words, on the one hand, 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 categorizes different issues in a predictable manner.

[30] Kushal Dave, Steve Lawrence and David M. Pennock, 'Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews', *Proceedings of the 12th International Conference on World Wide Web* (2003), 519–2, retrieved 28 May 2007, from the ACM Digital Library; Pang, Lee and Vaithyanathan, 'Thumbs up?'

[31] Aidan Finn and Nicholas Kushmerick, 'Learning to Classify Documents according to Genre', *Journal of American Society of Information Science and Technology*, 57 (2006), 1506–18. 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.

DATA PREPARATION

We downloaded all 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',[32] and 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. Details of the speech segmentation method can be found in Yu *et al.*[33]

We used Poole and Rosenthal's DW-NOMINATE score as the measure to select the 'extreme' senators – the twenty-five most conservative and the twenty-five most liberal senators in each senate. In a typical hold-out test for text classification tasks, a certain percentage of documents are set aside for training purposes and the rest are used as test data. In our study, 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.

The documents were subjected to a variety of pre-processing procedures. These procedures comprised different combinations of tokenization, stemming and part-of-speech tagging. We used a simple tokenizer to split the speeches into individual words. The tokenizer recognizes consecutive strings of alphabetical characters as valid words. Part-of-speech tagging is of particular importance, since part of our goal was to investigate the role of each content word class – nouns, verbs, adjectives and adverbs – in isolation.

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 have low coverage and thus are not useful for classification. We also removed the top fifty most frequent words as 'stopwords' which do not bear solid meaning (mostly function words, such as *the*, *a*, and *of*). Since every senator's ideological label is expected to be consistent across congresses, and because forty-five out of fifty senators in the 108th Senate (the test set) are also members of previous senates (the training set), we conjecture that senators' names are correlated with their ideological labels. To prevent the model from predicting based on name matching, we removed all names of senators from the vocabularies. Similarly, there might be a correlation between state names and ideological labels. The senators in the training set represented forty-five states; among them, the senators from seventeen states were all conservative and those from eighteen other states were all liberal during the 101st–107th senates. To prevent the classifier from being dominated by this correlation, we removed all state names and the names for the residents (for example, 'Marylanders') from the vocabularies.

The Porter Stemmer was used for suffix trimming.[34] 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 discard the tense information, which would harm the classification if verb tense were an important predictor.[35]

[32] Miss was not included because all single female Senators (e.g. Susan Collins and Barbara Mikulski) were saluted as 'Ms'.

[33] Bei Yu, Daniel Diermeier and Stefan Kaufmann 'Classifying Party Affiliation from Political Speech', *Journal of Information Technology & Politics*, 5 (2008), 33–48.

[34] M. F. Porter, 'An Algorithm for Suffix Stripping', *Program*, 14 (1980), 130–7.

[35] We used the MorphAdorner tagger to tag the parts of speech. Since the tagger has its own tokenizer, the generated word forms in this case are slightly different from the results of the simple tokenizer.

CLASSIFICATION DESIGN AND RESULTS

The combinations of classification algorithms, feature sets and feature weighting schemes lead to various classification methods. We tested SVMs using three feature weighting schemes – Boolean (word presence/absence), normalized frequency and *tf*idf* – and six feature sets – 'word' (individual words), 'stem' (stemmed words), 'noun' (nouns), 'verb' (verbs), 'adj.' (adjectives) and 'adv.' (adverbs). Table 1 lists the vocabulary size for each feature set. Hence, we tested 3*6 = 18 different SVM methods in total.

TABLE 1 *Feature Set Sizes*

	Feature set					
	word	stem	noun	verb	adj.	adv.
Size	19,459	11,395	8,831	6918	3665	890

In evaluating a classifier on a given fixed dataset, it is necessary to divide the data into a test set and a training set. We used an '*n*-fold cross-validation' approach. Here, the total dataset 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. An extreme case of *n*-fold cross validation is 'leave-one-out'; here, n equals the size of the dataset, the test is run n times, and one example is held out each time.[36]

We performed 'leave-one-out' cross-validation on the training set to estimate the effectiveness of our classification methods. The cross-validation evaluation results are shown in Table 2. The results show that *tf*idf*–SVM is the best classification method, reaching an accuracy level as high as 93.3 per cent averaged over the six feature sets. This article focuses on the analysis generated by the *tf*idf*–SVM classification results.

The *tf*idf*–SVM method achieved an accuracy of 92 per cent with the Word feature set (Table 3).[37] This result implies that out of the fifty most conservative and liberal senators in the 108th Congress, forty-six were correctly classified based solely on the words they used in their floor speeches. Recall that forty-five of the fifty 'extreme' senators were also extreme members of previous senates. The remaining five extreme senators are Reid, Chambliss, Cornyn, Graham and Sununu. Senator Reid had served in Congress 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), and therefore 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 four misclassified senators are Dayton, Frist, Reid and

[36] This is a standard approach in classification tasks; see, e.g., Tom Mitchell, *Machine Learning* (Toronto: McGraw Hill, 1997). An alternative approach consists in setting 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 datasets with large numbers of labelled examples. However, for small datasets such as ours, it is problematic since the arbitrary training/test split may accidentally lead to two datasets that are unlikely to have been produced by the same source.

[37] The accuracy was even higher (94 per cent) when adjectives were used as feature sets. Since there are only fifty test examples, 2 per cent accuracy improvement corresponds to one more correctly predicted example. Therefore, we do not think the accuracy difference is significant.

TABLE 2 *'Leave-one-out' Cross Validation on the Training Set Representations*

	Feature sets						
	word	stem	noun	verb	adj.	adv.	Average acc.
Boolean	88.9	88.0	89.5	82.9	89.2	77.2	86.1
Normalized frequency	88.6	89.7	84.6	66.7	84.6	61.8	90.7
*tf*idf*	95.4	94.0	95.4	93.4	93.4	88.9	93.3

TABLE 3 *tf*idf*–SVM *Extreme Prediction Results (with Extreme Training Set)*

	Feature sets					
*tf*idf*–SVM	word	stem	noun	verb	adj.	adv.
Extreme	92.0	86.0	84.0	88.0	94.0	52.0

Schumer, all represented in the training data except Reid, and therefore the 'in-sample' accuracy is $42/45 = 0.93$ and the 'out-of-sample' accuracy is $4/5 = 0.80$.[38]

This high classification success speaks for the validity of our approach. First, it shows that senatorial speech (at least for extreme senators) is highly consistent over time. This is true even though senators may face different issues in different congresses. Secondly, speech-based classification on a given set of senators also applies to new senators of the same ideological orientation. In others words, ideology, as expressed in congressional speech, appears to be shared. This latter conclusion, however, needs to be treated with caution as it is based on a small sample. We now turn to feature analysis to examine the content of the political ideologies.

FEATURE ANALYSIS

As discussed above, one of the main goals of this study is to shed light on the content of ideologies. This task can be accomplished by analysing the most discriminative features generated by the SVM classifier. The word features were sorted by their coefficients in the generated SVM linear prediction function. In our implementation, 'positive' features indicate the linguistic characteristics of conservative speeches and 'negative' features indicate those of liberal speeches.[39]

We report the corresponding lists of vocabulary words obtained with the *tf*idf*–SVM feature set analysis. Table 4 shows the words with the highest feature weights in the classifier, which achieved a 92 per cent classification rate (as seen in Table 3). 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, especially those known from scandals (*Enron* and *Firestone* as in 'Ford–Firestone'). Secondly, many of the top issues for

[38] Note, however, that the out-of-sample set is small due to lack of turnover among members of the Senate.
[39] These polarities are arbitrary. See the methodology section for technical details.

TABLE 4 *tf*idf*–SVM *Feature Set Analysis for All Vocabulary*

Words			
Liberal		Conservative	
FAS: −199.49	SBA: −113.10	habeas: 193.55	homosexual: 103.07
Ethanol: −198.92	Nursing: −109.38	CFTC: 187.16	everglades: 102.87
Wealthiest: −159.74	Providence: −108.73	surtax: 151.81	tower: 101.67
Collider: −142.28	Arctic: −108.30	marriage: 145.79	tripartisan: 101.23
WIC: −140.14	Orange: −107.98	cloning: 141.71	PRC: 102.90
ILO: −139.89	Glaxo: −107.81	tritium: 133.49	scouts: 97.55
Handgun: −129.01	Libraries: −107.70	ranchers: 132.95	nashua: 99.32
Lobbyists: −128.95	Disabilities:−106.44	BTU: 121.92	ballistic: 97.22
Enron: −127.71	Prescription: −106.31	grazing: 121.59	salting: 94.28
Fishery: −127.30	NIH: −105.52	unfunded: 120.82	abortion: 91.94
Hydrogen: −122.59	Lobbying: −105.35	catfish: 120.82	NTSB: 93.81
Souter: −121.40	NRA: −105.20	IRS: 114.91	Haiti: 97.28
PTSD: −119.87	Trident: −104.15	unborn: 111.88	PAC: 92.85
Gun: −119.52	RNC: −103.46	Taiwan: 111.13	taxing: 90.39
Firestone: −117.90	Lobbyist: −99.38	PLO: 106.56	nonseverability: 89.26
Lakes: −114.84	Homelessness: −95.68	EMS: 103.99	embryonic: 88.83

Notes: Words are decreasing in weight (from most conservative or most liberal). All words were converted into lower case during classification. In the above table, acronyms and proper names were recovered to upper case for 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 Administration; NIH: National Institute of Health; NRA: National Rifle Association; CFTC: Commodity Futures Trading Commission; BTU: British Thermal Unit; IRS: Internal Revenue Service; PLO: Palestine Liberation Organization; EMS: Emergency Medical Service; PRC: People's Republic of China; NTSB: National Transportation Safety Board; PAC: Political Action Committee.
Source: A complete list of the features can be found at http://textmining.syr.edu/beiyu/BJPS/train-adj-tf50-df10-tfidf-numlabel.svm.fw.sorted.txt.

Conservatives are from the domain of culture and values. Examples include *marriage*, *cloning*, *unborn*, *homosexual* 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 issues related to values and morality.[40]

We also conducted a classification based on nouns, adjectives, verbs and adverbs.[41] Recall from Table 3 that the classification based on adjectives was particularly successful,

[40] This is related to the literature on framing. For a recent review, see Jamie Druckman and Dennis Chong , 'Framing Theory', *Annual Review of Political Science*, 10 (2007), 103–26.

[41] We reproduce in Table 4 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 fifty to illustrate commonality.

followed by nouns and verbs, with adverbs providing by far the poorest classification. A look at Table 5 explains why. Note again the predominance of culture and value-related words, especially among the adjectives.

Overall, we see that the 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, abortion, stem cell research, family values, defence and (to a lesser extent) government administration. The issue of taxation is quite subtle. For example, the word *tax* as a noun (as in *income tax*) does not discriminate, while the verb *tax* and in the form *taxing* is highly discriminating for conservatives. In other words, conservatives tend to highlight the action of taxing, rather than the issue itself. We also find some common topics of discussion, mainly procedural terms (such as *adjournment* or *unanimous consent*) and idiosyncratic state-specific economic interests (such as *ranching*, *fishing*, *catfish* and *grazing*).

In contrast to the findings by Poole and Rosenthal, our analysis does not show a prevalence of economic (redistributive) factors in separating liberals from conservatives. Rather, 'value' issues such as abortion and stem-cell research play a big role, especially for conservatives. 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 separating adjectives for Democrats we find the word *gay*, and for the Republicans we find the word *homosexual*. In other words, the correct use of terms signals one's political 'type' to constituencies that care a great deal about these issues.

Up to this point, our analysis has focused on 'extreme' senators. This approach was based on the hypothesis that their ideological belief systems were more sharply defined as compared to 'moderate' senators. To assess the validity of this approach, we investigated the case of moderate senators to see whether their ideologies constitute a 'blend' between the more sharply defined lexica found at the extremes. In addition to providing a robustness check for our results, this analysis also sheds some additional light on the internal structure of ideologies.

CLASSIFYING MODERATE SENATORS

As before, we used the DW-NOMINATE score as the measure to select 'moderate' senators. That is, we divided the fifty non-extreme senators into moderate conservative and moderate liberal senators in each senate. If indeed our classification methods captured ideological positions, then the algorithms trained on extreme senators should perform worse on the test set of moderate senators. The results are shown in Table 6. The higher error rate in predicting moderate senators using classifiers trained on the extreme senators is consistent with the interpretation that speech classification is indeed sensitive to ideological differences as expected. To understand how and why moderate senators are misclassified, we analysed the prediction errors made by the classifier trained on extreme senators. To simplify the analysis, we used the Boolean representation with a vocabulary limited to nouns. Recall from Table 2 that Boolean–SVM achieved an 86.1 per cent prediction accuracy in cross-validation, which is lower than the 93.3 per cent accuracy obtained by the *tf*idf*–SVM. However, since the Boolean representation uses simple word presence/absence information, it is easier to understand which features cause

TABLE 5 *tf*idf*–SVM *Feature Set Analysis for Nouns, Adjectives, Verbs and Adverbs*

Nouns		Adjectives		Verbs		Adverbs	
Lib	*Cons*	*Lib*	*Cons*	*Lib*	*Cons*	*Lib*	*Cons*
FAS −148.75	habeas 159.99	wealthiest −119.55	partial-birth 89.71	nursing −304.52	taxing 277.88	enormously −331.04	basically 368.12
B-2 −122.09	everglades 138.47	reproductive −95.06	prepaid 86.65	policing −292.22	grazing 270.52	mentally −296.98	maybe 330.88
Souter −120.03	surtax 111.98	African-American −85.18	exclusionary 85.08	pre-existing −259.31	tax 253.61	incidentally −283.60	morally 329.40
disabilities −118.50	ranchers 105.13	Armenian −79.85	embryonic 79.58	Battered −242.68	taxed 242.64	disproportionately −275.30	supposedly 265.18
ethanol −115.71	cloning 96.64	managed −72.33	wireless 74.17	crumbling −225.85	cloning 180.55	hugely −257.50	seldom 239.26
Hydrogen −113.67	entitlements 90.62	toxic −70.89	chic 73.66	hate −208.90	married 178.63	critically −251.69	additionally 234.82
libraries −109.70	catfish 89.32	Salvadoran −67.47	ballistic 71.16	Breaks −198.49	ranching 173.51	chronically −249.22	abundantly 220.35
veterans −108.75	tritium 87.88	Homeless −66.95	unfunded 63.93	contaminated −196.41	adjourned 169.80	backward −244.66	objectively 218.62
handgun −107.54	marriage 86.91	republican −66.61	unborn 59.78	lifesaving 196.41	taxes 166.76	immensely −230.15	pretty 215.74
replacements −102.24	missile 82.16	armor-piercing −63.50	ninth 59.66	lobbying −191.44	sued 158.72	ecologically −221.92	theoretically 215.68
mammography −99.10	missile 82.16	generic −61.55	homosexual 56.04	laundering −190.12	importing 156.03	promptly −212.72	interestingly 211.23
genocide −98.83	tri-partisan 81.88	mail-order −60.15	law-abiding 53.91	Recessed −180.47	fracturing 149.05	thereon −202.67	purely 204.73
collider −98.20	c-17 81.00	after-school −58.65	fugitive 51.53	reformulated −177.95	sentencing −147.77	woefully −202.67	strictly 195.35
gun −95.93	marijuana 79.73	mail-in −58.56	nondefense 50.88	Displaced −175.86	induced 142.44	Unacceptably −186.46	irrespective 192.12
sweepstakes −95.84	stem 79.19	gay −55.58	gaseous 50.71	gun 169.55	nationalize 139.03	comprehensively −186.37	amok 185.80

TABLE 5 *(Continued)*

Nouns		Adjectives		Verbs		Adverbs	
Lib	*Cons*	*Lib*	*Cons*	*Lib*	*Cons*	*Lib*	*Cons*
handguns −94.56	interdiction 77.98	community-based −55.34	line-item 48.89	bargain −166.08	deploying 135.50	densely −183.86	assuredly 184.52
sir −91.65	ratepayers 75.57	dislocated −54.21	500-per-child 47.94	revered −164.13	saluted 132.83	home −182.02	downstream 181.82
fishery −91.06	bureaucrats 75.09	superconducting −54.21	flag 47.69	promise −156.74	checking 132.83	crucially −180.39	likewise 179.97
plutonium −86.70	IOUS 74.70	coastal −51.51	space-based 47.66	smoking −154.81	tax 131.11	indiscriminately −175.47	constitutionally 178.58
firefighters −86.50	soybean 74.30	pregnant −48.94	Chinese 47.26	worry −154.39	court-material 129.27	candidly −172.47	selflessly 167.03

Note: Words related to geography and proper names are removed. Words are decreasing in weight (from most Conservative or most Liberal). The weights are comparable within each feature set (noun, adj., adv., or verb), but are not comparable between feature sets (e.g. comparing the weights of the verb 'tax' and the noun 'surtax' is not meaningful.) The complete list of features and weights can be accessed from *http://nico.tech.northwestern.edu/~byu422/BJPS-feature-weights/*. In the Verb category, the first 'tax' refers to the verb in the present tense (as in 'the government should not tax the rich'), while the second 'tax' refers to the verb itself (as in 'to tax the rich is just wrong'). Our algorithm treats different part-of-speech tags as different words. We removed the part-of-speech tags for better readability.

TABLE 6 *tf*idf*–SVM *Moderate Prediction Results (with Extreme Training Set)*

	Feature sets					
Representations	word	stem	noun	verb	adj.	adv.
moderate	48.0	52.0	54.0	62.0	58.0	48.0

the prediction errors than in the case of *tf*idf*. Each document vector consists entirely of 1s and 0s.

For each test example x, the SVM classifier computed a *decision value* Y_x as defined in Equation 3, where w and b were computed during the training process:

$$Y_x = w \bullet x + b = \sum_{i=1}^{n} \left(w_i x_i + \frac{b}{n} \right). \tag{3}$$

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

$$\frac{Y_x}{|Y_x|} = \sum_{i=1}^{n} \frac{w_i x_i + \frac{b}{n}}{|Y_x|}. \tag{4}$$

For each i, 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 of them are responsible for the misclassification. Table 7 lists all examples of misclassified moderate Senators along with their decision values.

TABLE 7 *Prediction Errors in the Moderate Test Case with Boolean Noun Features*

False conservative (10)			False liberal (7)		
Bayh	−1	0.716	Alexander	+1	−0.019
Breaux	−1	0.456	Collins	+1	−0.171
Carper	−1	0.038	Dewine	+1	−0.523
Chafee	−1	0.461	McCain	+1	−0.315
Conrad	−1	0.218	Smith	+1	−0.625
Edwards	−1	0.240	Snowe	+1	−0.876
Hollings	−1	0.702	Specter	+1	−0.247
Lincoln	−1	0.043			
Nelson	−1	0.560			
Pryor	−1	0.404			

Table B1 in Appendix B lists the twenty features with the largest contributions to each side for all misclassified examples listed in Table 7 (liberal contributions are listed in the first column). The highlighted features help us understand why these examples are misclassified.

TABLE 8 *Example of Words with Significantly Different Numbers of Occurrences in the Conservative and Liberal Speeches*

	Document frequency			Term frequency		
Word	training-extreme	test-extreme	test-moderate	training-extreme	test-extreme	test-moderate
'taxation'	148:103	21:15	17:16	953:423	280:62	158:145
'taxing'	99:62	17:5	8:6	333:141	59:5	35:31

Note: Format of table is: (occurrences in conservative speeches: occurrences in liberal speeches).

This table shows that, on the one hand, some moderate liberal Senators were misclassified as conservative because they talked about some traditionally 'conservative issues', such as bureaucracy, taxation, marriage, etc., although mentioning these issues, of course, does not necessarily imply support for the conservative position.[42] 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.

To investigate this in more detail, we took a closer look at the issue of taxation. Table 8 lists the term frequencies and document frequencies of the words 'taxing' and 'taxation' for three subsamples: *101st–107th extreme senators, 108th extreme senators, 108th moderate senators*. Recall from Table 5 that *tax* and *taxpayers* are not discriminating between the two ideologies. In contrast, *taxation* and *taxing*, which emphasize the action of taxing, are highly discriminative. Extreme conservative senators are much more likely to use these words than extreme liberal senators: For example, *taxation* was used 953 times by extreme conservative senators in the 101st–107th Congresses and 280 times in the 108th Congress; however, it was used only 423 times by extreme liberal senators in the 101st–107th Congresses and 62 times in the 108th Congress. In contrast, *taxation* was used by the 108th moderate liberal and conservative senators with nearly equal frequencies (158:145). The usage of *taxing* follows a very similar pattern. This means that extreme senators collectively use *taxing* and *taxation* as strong ideology indicators, while moderate senators, typically, do not. Indeed, moderate senators who do use *taxing* and *taxation* in similar ratios as their more extreme peers are more likely to be misclassified.

DISCUSSION

While the results of our analysis are promising, many questions remain. Of particular importance are robustness issues. For example, we may want to investigate whether our results for the Senate also hold for the US House of Representatives. Unfortunately, we cannot directly compare these findings because the DW-NOMINATE scores are not equivalent across chambers.[43] However, we can use a simpler but somewhat less informative

[42] For example, Senator Colman in the 106th Senate mentioned 'grievous injury' before he expressed his objection to this amendment to the partial-birth ban act.

[43] To compare the two chambers directly, it is necessary to use a common space score for both the House and the Senate. See, for example, Royce Carroll, Jeff Lewis, James Lo, Nolan McCarty, Keith Poole and Howard Rosenthal, '"Common Space" (Joint House and Senate) DW-NOMINATE Scores with Bootstrapped Standard Errors' (2009).

measure, such as party membership. 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 (Boolean-SVM and *tf*idf*-SVM). The two classifiers were then used to predict the party membership of the senators in the 108th Senate. The *tf*idf*-SVM classifier reached 77.8 per cent prediction accuracy and the Boolean-SVM one reached 86.9 per cent, both indicating a good classification success for party membership. While this is a lower accuracy compared to our previous classification results, this drop is to be expected as we are no longer restricting attention to senators with more extreme ideological positions. Notice that a drop in accuracy would also provide additional evidence for our hypothesis that extreme senators have a more clearly defined ideology. To directly compare classification based on party membership with classification based on ideology, we divided our entire 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 per cent for the *tf*idf*-SVM classifier and 83.84 per cent for the Boolean-SVM variant. Thus, the two classification results are highly similar. To investigate the relationship between ideology and party membership further, we computed the kappa[44] agreement score as a measure of consistency between the senators' ideology labels and party membership. For our dataset (800 senatorial speech documents for the 101st–108th Congresses), the kappa equals 0.932, which means almost perfect agreement. In other words, the previous result implies that party membership and ideological classification are highly correlated.

With this reassurance, we can now assume that it is possible to compare the classification success rates between the House and the Senate. Some preliminary results along these lines can be found in work by Yu, Diermeier and Kaufmann on party classification. Using *tf*idf*–representations for the 2005 Senate, Yu *et al.* reach a prediction accuracy on party membership of 69.7 per cent for the Senate, and yet the same estimates performed on House data yield a classification accuracy of 80.1 per cent.[45] One possible explanation for the higher accuracy in the House may be related to its germaneness requirements and increased leadership control. Indeed, House speeches are more constrained and thus reflect more consistent positions. Of course, House speeches are also usually much shorter. In related work Yu, Kaufmann and Diermeier have found that 62 per cent of the beginning sentences and 72 per cent of the ending sentences in House speeches clearly indicate the overall positions of the speeches (supporting or opposing the bill under debate).[46] In contrast, the percentages drop to 49 per cent and 35 per cent in the Senate speeches. From a text classification perspective, the positions of speeches with clear and consistent indicators are relatively easier to classify. As a result, the above differences between Senate speeches and House speeches make the ideology classification of Senate speeches more difficult. These findings suggest some intriguing possibilities for further research. For example, a surge in classification accuracy over time may indicate an increase in agenda control or changes in the

[44] The kappa coefficient is often used to measure inter-rater agreement in annotation. We followed the kappa computation procedure described at http://faculty.vassar.edu/lowry/kappa.html.

[45] Bei Yu, Stefan Kaufmann and Daniel Diermeier, 'Classifying Party Affiliation from Political Speech' *Journal of Information Technology and Politics*, 5 (2008), 33–48. The lower accuracy is a consequence of a smaller dataset.

[46] Bei Yu, Stefan Kaufmann and Daniel Diermeier, 'Exploring the Characteristics of Opinion Expressions for Political Opinion Classification', *Proceedings of the 9th Annual International Conference on Digital Government Research* (dg.o 2008) (Montreal, May 2008), pp. 82–9.

partisanship of members and their constituencies. Addressing these questions will require additional data collection that would allow for longitudinal cross-chamber studies.

A longitudinal approach may also help to answer the question whether the time period studied here has some unusual characteristics. For example, given that the period we analysed was characterized by sustained high economic growth, it is possible that political conflict may have shifted to other, non-economic, factors.[47] That said, a shift towards non-economic conflict would signal an important deviation from the interpretation offered by Poole and Rosenthal. One way to address this issue would be to analyse earlier congresses to see whether we find a similar importance of non-economic dimensions or whether we are witnessing a genuine shift. At this point, such an investigation is severely impeded by the limited availability of text in digital format.

Similarly, we can apply these methods to study language in different legislatures. In subsequent work, Høyland and Godbout have used SVMs to classify the party affiliation of members of the European Parliament (MEPs) based on their plenary speeches in the 5th and 6th Parliaments. Unlike the US Congress, the European Parliament (EP) contains as many as seven distinct party groups in a given session, and therefore the classification task is much more complex. Nonetheless, the overall classification accuracy of their model was relatively high for most parties, reaching as much as 77 per cent for the United Left and Nordic Green parliamentary group. The authors were also able to use SVM to compare the speech similarities of party-group members who represented different countries.[48] Although most parties displayed a high level of cohesion in their voting records during the 5th and 6th Parliaments, their analysis of legislative speech demonstrated that the same party-group members spoke about different issues in their planetary speeches. This was found to be especially true when comparing MEPs from old versus new member states who shared the same party and joined the European Union after the 2004 enlargement.

In summary, text classification results combined with feature analysis may provide a promising new methodology to study the content, changes and differences of ideological positions quantitatively. By means of some clever comparative designs, it may also be possible to disentangle some of the suggested influence factors of ideological positions, such as constituency pressure, party leadership influence and agenda control.

CONCLUSION

Perhaps the most important finding of the Poole–Rosenthal approach is that much of the voting behaviour in Congress can be explained by a stable, low-dimensional issue space. Indeed, for recent years (104th–106th Congress), a single dimension could account for about 90 per cent of all roll-call choices by members of Congress.[49] Poole and Rosenthal have argued that this dominant ideological dimension represents the traditional left–right ideological continuum associated with the government's role in the economy and economic redistribution,[50] which increasingly is subsuming intra-party conflict and voting on the second dimension.

[47] We thank an anonymous referee for pointing out this possibility.

[48] Bjørn Høyland and Jean-François Godbout, 'Predicting Party Group Affiliation from European Parliament Debates' (paper presented at the European Consortium for Political Research Meeting of the Standing Group on the European Union (Riga: Latvia, 2008)).

[49] Poole, *Spatial Models of Parliamentary Voting*.

[50] Except during the Era of Good Feelings (1817–25) and the period surrounding the Civil War (1853–76); Poole and Rosenthal, 'Congress'; Poole and Rosenthal, *Ideology and Congress*.

In this article, we have argued that analysing political speeches rather than voting behaviour can provide additional insights into the ideological positions of members of Congress. We rely on methods from text analytics which allows us to dispense with labour-intensive human coding. In addition, the approach was shown to be replicable, efficient and highly scalable. Our analysis of senate speeches provided strong evidence for consistent and common ideologies expressed in speech. The classification success was much higher than in comparable studies of consumer and voting behaviour. It also allowed us to gain some understanding of what separates the two ideologies at the conceptual level. First, consistent with Poole and Rosenthal, senators appear to separate on economic issues. However, such separation is expressed quite differently by the two sides, with Democrats frequently referring to corporations, special interests and the environment, while Republicans frequently mention the act of taxing (not the more technical 'taxation'). Secondly, in contrast to Poole and Rosenthal, terms related to values and moral issues are used prominently, especially by conservatives in the context of family and abortion. Thirdly, the speeches seem to be well designed to appeal to partisan constituencies, as in the use of 'gay' versus 'homosexual'. This may provide some evidence that senators consciously use 'signal' words to demonstrate their allegiance to core beliefs of more extreme constituents. Fourthly, moderate senators exhibit more nuanced speech patterns, which lowers their classification accuracy. Interestingly, these misclassifications add further support to the importance of some of the key features identified.

It is also interesting to note that part of our findings share some similarity with the work of cognitive linguists such as Lakoff.[51] Lakoff argues that the Republican and Democratic ideologies reflect different value systems deeply associated with personal views of morality. Our results are consistent with this approach, but further research is necessary. It would be particularly important to investigate whether the content of ideologies has shifted over time. While such an approach is currently not viable due to the limited availability of digitized text data, it marks an important next step in our understanding of political ideologies.

APPENDIX A

The task of a Support Vector Machine in the training phase is to find the two separating hyperplanes such that the margin is maximized. For illustration, consider the special case that each data point is represented by a pair of co-ordinates 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}.

The points x on the maximally separating hyperplane (the dotted line in the figure) satisfy the equation:[52]

$$w \bullet x + b = 0, \tag{5}$$

where w is a vector perpendicular to the hyperplane 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

[51] See, for example, George Lakoff, *Moral Politics: How Liberals and Conservatives Think* (Chicago: The University of Chicago Press, 2002).

[52] The large dot in the equation refers to the operation of the inner product of two vectors.

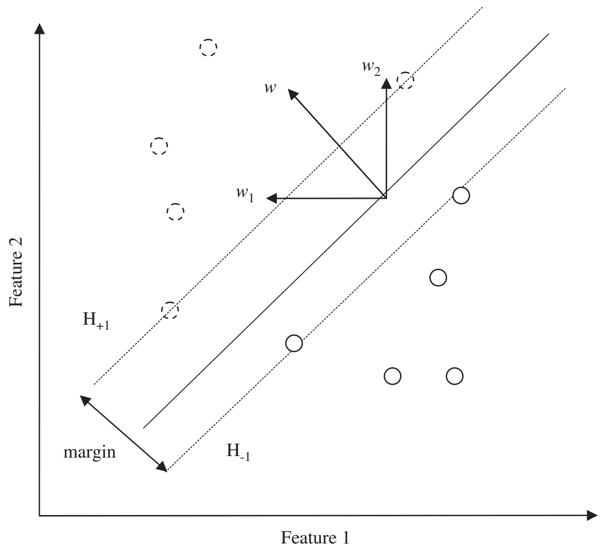

Fig. 1. Linear-separating hyperplanes with maximized margin

$y_i = +1$; 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 maximally separating hyperplane, given by the sign of $w \bullet x + b$.

Aside from its role in the classification of new data, the vector furnishes information about the informativeness of each feature in determining category membership. 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.

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

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \tag{6}$$

subject to the constraints

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, ..., l. \tag{7}$$

Only the support vectors have non-zero α_i values. For the other data points, $\alpha_i = 0$. In (6), 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.[53]

[53] Edda Leopold and Jörg Kindermann, 'Text Categorization with Support Vector Machines: How to Represent Texts in Input Space?' *Machine Learning*, 46 (2002), 423–44.

After training, the parameters w and b are given by (8) and (9).[54] Given a text example x, the linear decision function is the sign of $w \bullet x + b$.

$$w = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{8}$$

$$b = y_{sv} - w \bullet x_{sv}. \tag{9}$$

There are several efficient implementations of the SVM algorithm, such as LIBSVM and SVM^{light}.[55] We used the SVM^{light} package with its default setting in this study.[56]

APPENDIX B

The following table lists the twenty features with the most contributions to each side for all misclassified examples listed in Table 6. Conservative contributions are featured first. The features that are highlighted in bold font help us understand why these examples are misclassified.

TABLE B1 *Prediction Errors in the Moderate Test Case with Boolean Noun Features*

False conservative errors	False liberal errors
Bayh −1 0.716	**Alexander** +1 −0.019
lady beltway strings personality moneys payroll turf marriage prayers marines quote morass hometown guy exclusion motives adage infantry notification back	taxation bureaucracies appetite taxing adjournment businessmen congressmen designee strings lifestyle sanctity superintendent tactics tape infringement theft hill mandates cattle folks
childhood backgrounds prejudice gaps cynicism counselling racism minorities executives mayors wounds bridges eloquence gender shores specialist segregation sisters ignorance diseases	**homelessness sewage childhood backgrounds heating self-sufficiency counselling racism models disabilities infants graduates minorities artists starvation optimism hunger executives nurses trauma**
Breaux −1 0.456	**Collins** +1 −0.171
taxation bureaucracies businessmen standpoint lady personality micromanage taking moneys folks excuses mercy prayers quote nongermane politicians hometown exclusion nonsense contention	taxation bureaucracies disincentive adjournment prioritize standpoint designee beltway lifestyle tyranny bombing airmen tactics moneys tape infringement trucking theft titles mandates
hazards backgrounds prejudice breaks shelters planet descent epidemic downturn depression infants outreach mortgage minorities optimism caregivers giants treatments neglect suburbs	homelessness hazards backgrounds gaps breaks breast ecosystem shoreline heating cops epidemic medication ecosystems counterparts models disabilities depression outreach minorities optimism

[54] The abbreviation *sv* stands for an arbitrary support vector. In the SVM^{light} software package, the first support vector (according to its order in the input data) was used to compute b.

[55] Joachims, 'SVM^{light}'.

[56] Chang and Lin, 'Library for Support Vector Machines'.

TABLE B1 *(Continued)*

False conservative errors	False liberal errors
Carper −1 0.038	**Dewine** +1 −0.523
taxation taxing standpoint lady congressmen beltway strings personality taking airmen moneys theft mandates folks fitness payroll beef sailors commonsense usage	taxing unborn adjournment businessmen prioritize accusation congressmen designee hike sanctity arrogance tactics reasoning irresponsibility etc theft folks excuses payroll habit
hazards childhood backgrounds gaps breaks literacy epidemic counselling cutting-edge models infants outreach hunger executives nurses giants mayors disparities bridges stress	**homelessness mortality sewage childhood prejudice breast ecosystem shoreline planet cops epidemic medication counselling cutting-edge disabilities depression infants graduates outreach sediment**
Chafee −1 0.461	**McCain** +1 −0.315
taxation hike tactics titles sailors prayers egg criminal allotments entity concession relation snow malpractice recipient dilemma pitch monument tours mom	taxation appetite disincentive bureaucrat taxing adjournment businessmen prioritize standpoint congressmen designee spenders lifestyle micromanage tyranny bombing airmen tactics moneys tape
prejudice breast ecosystem shoreline outreach optimism nurses trauma analyses treatments neglect villages jurisdictions recession gallons sediments illnesses survivors treasures photograph	**mortality hazards rent sewage backgrounds prejudice gaps degradation breaks breast ecosystem shelters heating planet insecurity self-sufficiency cynicism epidemic medication ecosystems**
Conrad −1 0.219	**Smith** +1 −0.625
taxation appetite disincentive taxing adjournment businessmen rancher prioritize standpoint forefathers spenders taking bombing airmen tactics moneys tape irresponsibility cattle folks	taxation bureaucrat taxing unborn bureaucrats accusation lady forefathers lifestyle micromanage tyranny taking sanctity bombing superintendent grass tape titles hill mandates
rent revitalization recessions gaps heating planet cops descent epidemic medication racism downturn counterparts models depression graduates mortgage sediment optimism executives	**mortality revitalization childhood prejudice breast literacy planet insecurity orientation self-sufficiency descent epidemic medication counselling racism downturn counterparts relocation models disabilities**
Edwards −1 0.240	**Snowe** +1 −0.876
taxation hike tyranny bombing tape folks excuses bureaucracy commonsense nest prayers egg politicians hometown motives vacuum commissioners adversaries contacts back	bureaucracies disincentive unborn rancher standpoint designee personality lifestyle tyranny bombing airmen tape titles mandates tendency ranchers payroll habit turf bureaucracy
hazards rent sewage childhood breast explosions insecurity cops medication cutting-edge downturn graduates mortgage minorities executives nurses treatments bridges lung fires	**mortality hazards rent revitalization childhood backgrounds gaps degradation breast ecosystem shelters shoreline literacy heating planet self-sufficiency epidemic medication counselling ecosystems**
Hollings −1 0.702	**Specter** +1 −0.246
taxing lady forefathers personality superintendent trucking titles mandates folks	taxation adjournment businessmen designee beltway personality lifestyle tyranny sanctity

TABLE B1 *(Continued)*

payroll turf beef usage launch marriage liberals wheat politicians hometown criminal

hazards gaps shelters planet insecurity descent relocation models trauma giants analyses riders analysts locks waterways jurisdictions collaboration recession beat gallons

bombing airmen tactics moneys tape theft payroll turf bureaucracy sailors usage

mortality hazards rent childhood degradation breast shelters literacy epidemic medication counselling cutting-edge downturn disabilities graduates outreach mortgage minorities optimism executives

Lincoln −1 0.043

taxation disincentive adjournment prioritize forefathers tyranny superintendent tactics reasoning titles mandates folks fitness excuses payroll beef commonsense desires marriage prayers

homelessness mortality hazards backgrounds gaps breast ecosystem shelters heating insecurity orientation epidemic medication counselling ecosystems counterparts models disabilities depression infants

Nelson −1 0.560

taxation bureaucrat airmen grass infringement grab hill mandates cattle tendency folks ranchers payroll habit beef sailors commonsense marriage prayers marines

childhood gaps self-sufficiency descent epidemic counselling downturn counterparts models disabilities outreach optimism nurses treatments recession gallons specialist net illnesses obstacle

Pryor − 1 0.405

standpoint personality tyranny airmen theft titles mandates folks fitness sailors commonsense nest marriage prayers marines quote wheat foresight mankind egg

hazards rent prejudice breast cops epidemic medication counselling cutting-edge counterparts disabilities mortgage minorities artists optimism nurses perspectives stress analysts vaccines