Predicting Political Ideology from Congressional Speeches:

Can We Distinguish Ideological and Social Attributes?

Shlomo Argamon and Jonathan Dunn

Illinois Institute of Technology

Abstract

We describe a system for predicting the political ideology of authors of written texts, evaluated using the text of speeches from the *Congressional Record* for both chambers of the U.S. Congress from the 104th through the 109th congresses (Gentzkow & Shapiro, 2013). Political ideology is operationalized in three ways: first, as simple party membership; second, using aggregated special interest group ratings; third, using DW-NOMINATE scores (Poole & Rosenthal, 2007). Principal Components Analysis was used to reduce 134 special interest group ratings of members of congress into three robust dimensions: (i) Government and Institutions, (ii) Government and Individuals, and (iii) Government and Animals. Each dimension has a scalar value for each speaker, with the top end of the scale representing a liberal ideology and the bottom end of the scale representing a conservative ideology. These measures of political ideology are discretized and individual speeches are classified according to the ideological ratings of the speaker using an SVM classifier with a linear kernel (Fan, et al., 2008) using word-form bigrams and part-of-speech trigrams as features, calculated using the TF-IDF measure.

Together with these operationalizations of political ideology, the speeches are also classified according to social attributes of the speakers: Age, Sex, Race, Religion, Geographic Location, and Previous Military Service. This allows us to ask whether classification according to the speaker's ideology can be clearly distinguished from classification according to other properties. In other words, are social attributes a confounding factor for supposed textual classification by political ideology? We use two measures to look at this question: first, we try to predict social attributes using only special interest group ratings in order to see the predictive power of ideology measures alone for social attributes; second, we compare the predictive textual features across profiles in order to see if ideology is classified using the same textual features as social attributes. Finally, one weakness of text classification is that models often do not generalize to new datasets. In terms of classification by political ideology, the issue is that predictive features may reflect heavily topic-dependent items which are present during a given debate but do not reoccur in other situations. We try to prevent these problems and produce generalizable models by (i) training on data from the 108th and 109th congress and testing on data from the 105th congress, so that there is a gap between the training and testing data; and (ii) by training and testing with both chambers together, so that the models are not dominated by the topics of debate in a given chamber in a given year.

We find that both the generalizability and independence of classification by political ideology are called into question by the high correlation between predictive features for Race, Sex, Geographic Location, and Religion, on the one hand, and political ideology, on the other hand, suggesting a confound in which social attributes are mistaken for ideological attributes. It is not clear, however, which direction the confound applies.

1. Text Classification Methodology

The dataset consists of the complete text of speeches published in the *Congressional Record* for the 104th (starting in 1995) through the 109th (ending in 2007) congresses (text prepared and divided into speeches by Gentzkow & Shapiro, 2013). Speeches with fewer than 100 characters, often procedural, were removed. The 104th, 106th and 107th congresses were used for system configuration tests. These included features types (word-form, lemma, part-of-speech), context windows (unigrams, bigrams, trigrams), feature calculation measures (relative frequency, baseline relative frequency, and term frequency * inverted document frequency), learning algorithms (SVM with a linear kernel, SVM with an RBF kernel, Logistic Regression, J48, Naïve Bayes), and binning techniques for scalar class attributes (eventually using equal-frequency binning to create balanced classes). Reported results are trained on the 108th and 109th congresses, across the House and Senate, and tested on the 105th congress. Later congresses were used for training because they are more balanced for social classes (e.g., race and sex). All models were trained using balanced classes formed by removing randomly selected instances from majority classes. The text classification was performed using Weka (Hall, et al., 2009) to train and evaluate models, using the LibLinear and LibSVM packages (Fan, et a., 2008) for the SVM classifiers.

2. Previous Work Using Text to Study Political Ideology

Laver, et al. (2003) use word-form unigram relative frequencies to provide each word a probability of belonging to a specific policy-position group, using a pre-selected set of training documents. The collective probabilities of all the predictive words allows new documents to be classified according to the policy positions which they take. Although it focuses on specific policies rather than more abstract ideologies, this work can in practice be seen as classifying texts according to their political ideology.

Thomas, et al. (2006) classify U.S. Congressional speeches according to their opinion of the bill being debated using word-form unigram presence features. They also use same-speaker agreement (e.g., speaker identity) and different-speaker agreement (e.g., direct references) relations between speeches as soft constraints on the classification to improve performance.

Yu, et al. (2008a) measure the level of sentiment present in Senate speeches and other genres (e.g., movie reviews) and then compare the usefulness of different feature sets for detecting sentiment in these genres. They find, using an annotation-based study, that political speeches do not have a high level of sentiment, and that topics are more predictive of political ideology than sentiment. Yu, et al. (2008b) used a number of feature calculation types with word-form unigrams using SVM and Naïve Bayes classifiers to predict a speaker's party membership given a congressional speech, training and testing at different time periods and across chambers. Diermeier, et al. (2011) use an SVM classifier with word-form unigram frequencies to classify senators (using collected speech from a congress as a single document) as extremely liberal or conservative, as operationalized using DW-NOMINATE scores, in order to see which textual features are predictive of political ideology.

Koppel, et al. (2009) use Bayesian multi-class regression with word-form unigram frequencies to classify Arabic texts according to the organization which produced the text. This task could be viewed as classifying according to ideology, although it is not independent of social factors, as shown by the predictive features (e.g., "Israel" and "Palestine").

Sim, et al. (2013) use an automatically-constructed lexicon of ideological cue words and phrases to measure how ideologically-charged a given text is. The system focuses on ideological cues and how many non-cue "fillers" separate the cues, using the density of cues to represent degree of ideology present in a text.

Iyyer, et al. (2014) look for political ideology at the sentence level, framing the task as looking for political bias, by expanding the search for ideologically-charged phrases to syntactically related but non-linear constructions. This has the advantage of correctly classifying instances in which the speaker expresses a negative view of an ideologically-charged phrase, which would be detected by other systems as a positive view.

3. Operationalizing Political Ideology

For each speaker (833 in total) biographical details were gathered from the CQ Congress database (McCutcheon & Lyons, 2010). The first step in evaluating the profiling of ideology is to construct a measure of ideology. We used interest group ratings (e.g., the League of Conservation Voters) as such a measure, taking them for each speaker from the CQ Congress database. Additional interest group ratings were taken from the Votesmart.org database. In both cases, the most recent available rating was used (i.e., if a member has served from 1990 through 2003, then the ratings from 2003 would be used; this is not likely to be affected by change in an individual legislator's ideology over time, as Poole & Rosenthal, 2007, show that individual legislators change little over the course of their career). Interest group ratings range on scales between 0 and 100, with 100 indicating high ideological agreement with the group producing a given scale. The analysis started with 134 interest group ratings, some with relatively sparse coverage across speakers.

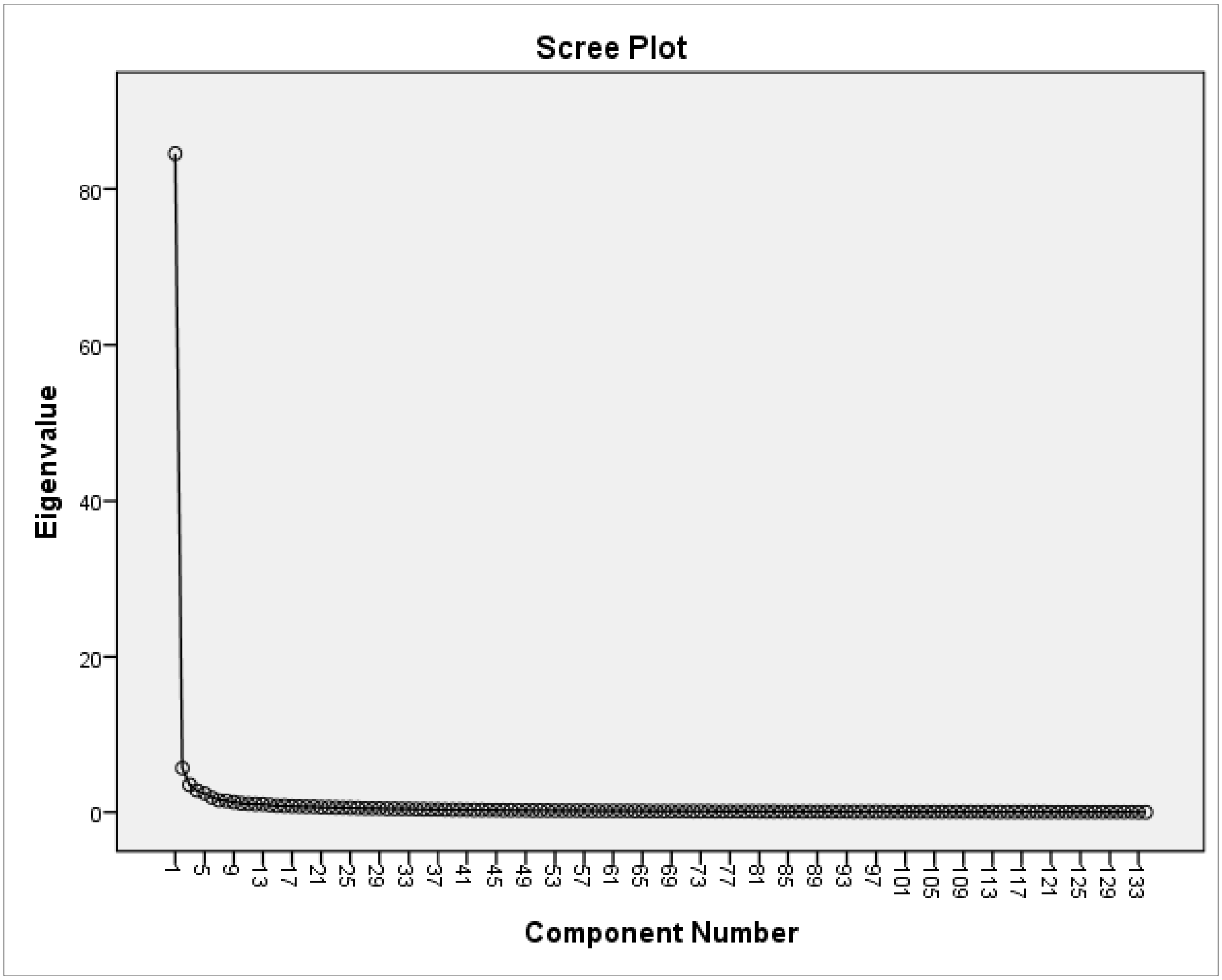
The first task was to reduce these ratings into a small number of robust and interpretable groups that represent a single dimension of the speaker's ideology. Principal Components Analysis was used to group interest ratings together. Missing values were replaced with the scale's mean value. The varimax rotation method was used with Kaiser normalization in order to create components with unique memberships. Thirteen components were identified with eigenvalues above 1 before rotation. Interest group ratings were included in a given component if (i) the component loading is above 0.400, (ii) the component loading is within 0.150 of the highest member of the component, and (iii) the component loading is at least 0.300 above the same rating's loading in a different component. Only 10 components had at least one member according to these criteria. To ensure adequate representation, we only included components in which no more than 100 speakers were not rated by any of the groups in the components. This last condition eliminated all but the first three components; the other components had relatively sparse coverage and were likely influenced by the use of mean values to replace missing instances. The three components, shown in Table 1 together with the variance explained by each, represent the relation between the Government and (1) Institutions (for example, labor unions and universities and businesses), (2) Humans (for example, immigrants and children and the needy, regardless of nationality), and (3) Animals. Together, these components account for 67.69% of the variation in interest group ratings. Each of these components is turned into a single score by taking the average of all member ratings.

Table 1. Ideology Components with Percent of Variance Explained

#	Name	Scale	Variance
1	Government and Institutions	100 = Liberal, 0 = Conservative	37.08%
2	Government and Humans	100 = Liberal, 0 = Conservative	26.48%
3	Government and Animals	100 = Liberal, 0 = Conservative	4.13%

Figure 1 below shows the scree plot for these components. The first component is the most significant, while the others quickly fall in their eigenvalues. Only three components were ultimately retained because of the distinctness and robustness factors described above.

Figure 1. Scree Plot For Interest Group Principle Component Analysis



The three ideology ratings representing these components, computed as described above, have the correlations shown in Table 2. Also shown are the correlations between the interest group-based ideological measures and the roll call-based measures from Poole & Rosenthal (DW-NOMINATE). Although all of the dimensions are related, Institutions is more related to the other interest group measures (0.830 and 0.735) than the others are to each other; thus, this is the major dimension of ideology in the interest group ratings. The three ideological ratings range from correlated (0.695) to highly correlated (0.830). This is, in part, because ideology in congress is often unidimensional (Poole & Rosenthal, 2007). However, the ideological dimensions are significantly unique here; further, Poole & Rosenthal examine roll call voting, while interest group ratings are more directly representative of ideological factions, especially those, like the Animals dimension, which fall outside of the normal political divide. Thus, it remains useful to examine all three dimensions. The correlations between the interest group measures and the roll call measures are all negative, simply because the roll call measures use higher numbers for conservative scores and the interest group measures use higher numbers for liberal scores. The correlations are calculated using Spearman's rho because the scales have different sensitivities but should produce a similar order of legislators. Poole & Rosenthal's first dimension, which captures the great majority of variation in roll call voting, correlates highly with the Institutions interest group measurement (0.896, when the negative is removed). The second roll call dimension, interestingly, is only correlated with the Animals interest group measure, and only to a lesser degree (0.278 when the negative is removed). Thus, we see that both types of ideology measurements are related, but still capture different dimensions of ideology. This means that it is worthwhile to examine all five measurements as operationalizations of ideology.

Table 2. Spearman Correlations Between Components and Roll Call Dimensions

	1st Dimension	**2nd Dimension**	**Institutions**	**Individuals**	**Animals**
1st Dimension	---	0.049	-0.896	-0.792	-0.633
2nd Dimension	0.049	---	-0.113	-0.074	-0.278
Institutions	-0.896	-0.113	---	0.830	0.735
Individuals	-0.792	-0.074	0.830	---	0.695
Animals	-0.633	-0.278	0.735	0.695	---

We begin by examining the interest group measures more closely. The first component, concerning the relation between Government and Institutions, contains 23 interest groups, some of which are shown in Table 3. First, this component has a large number of unions (11, or 48% of the component). This reflects a question of how involved the government should be in regulating and monitoring the relationships between businesses and industries, on the one hand, and individual workers, on the other hand. A higher score indicates agreement with a more active role for government. Second, this component has a number of groups concerned with the rights of minorities or less powerful members of society in relation to institutions, both public and private. Thus, the component contains interest groups seeking government involvement and protection of African Americans, Children, and the Elderly (6 groups, or 26%). Again, a higher score indicates agreement with a more active role for government in mediating the relationships between institutions and individuals.

Table 3. Top Five Interest Groups in Component 1, Government and Institutions

Component	**Interest Group**	**Loading**
1	Committee on Political Education of the AFLCIO	0.873
1	United Auto Workers	0.869
1	American Federation of State County Municipal Employees	0.864
1	Communication Workers of America	0.862
1	American Federation of Government Employees	0.862

The second component, concerning the relation between Government and Humans, contains 17 interest groups, some of which are shown in Table 4. The focus within this component is on the role of government in directly seeking to improve and protect the lives of individuals, such as women (in terms of reproduction, 3 groups), immigrants (1 group), farmers (1 group), the non-religious (1 group), and citizens of other nations (3 groups). A higher score on this ideological scale indicates support for a more active government role in the lives of individuals, directly rather than as mediated by institutions as in the first component.

Table 4. Top Five Interest Groups in Component 2, Government and Humans

Component	Interest Group	Loading
2	National Journal Liberal Composite Score	0.792
2	National Journal Liberal Economic Policy	0.774
2	Defenders of Wildlife Action Fund	0.769
2	National Journal Liberal Foreign Policy	0.767
2	National Journal Liberal Social Policy	0.731

The third component, concerning the relation between Government and Animals, contains 6 interest groups, some of which are shown in Table 5. The focus of this component is on the role of government in representing and producing policy in respect to non-humans. A higher score indicates agreement that government should have an active role in non-human life.

Table 5. Top Five Interest Groups in Component 3, Government and Animals

Component	Interest Group	Loading
3	Animal Welfare Institute	0.695
3	Humane Society of the US	0.593
3	Born Free USA	0.560
3	American Humane Association	0.560
3	American Society for the Prevention of Cruelty to Animals	0.560

4. Interactions Between Political Ideology and Other Speaker Attributes

Given these interest group ratings, individual and combined into ideological ratings, can we predict personal characteristics of the speaker? We viewed this as a classification task, with the personal characteristics as classes and the interest group ratings and ideological scores as features used to distinguish between classes. The personal characteristics include Age, Length of Service in Congress and First Election Year, Chamber of Congress, Political Party, Gender, Military Service, Race, Religion, and Occupation (before serving in congress). This is important because of the possible confounds between profiles of different attributes. In other words, if a particular ideology has the same predictive features as a non-ideological attribute (e.g., race), then classifications distinguishing the two become difficult. Thus, we need to make sure that successful ideological profiling is not simply an artifact of profiling other attributes. The point of this section is to investigate possible relations between the ideological dimensions and the social dimensions.

The relation between age and interest group ratings was first examined using the Pearson R correlation. Table 6 shows all ratings which correlate significantly with Age, with significant defined as $p < 0.01$ (2-tailed) and a correlation coefficient above 0.2. Only four ratings had any significant correlation with age, two positive and two negative. The two positive correlations, in which younger members of congress have higher ratings (age is represented using the year of birth) advocate limited immigration, border fences, and a neo-conservative foreign policy. The two negative correlations, in which older members of congress have higher ratings, advocate for farmers and Arab Americans. Given the small number of weak correlations, it is worth asking whether a discrete approach to this personal attribute of speakers would better support classification.

Table 6. Scalar Age and Interest Group Ratings using Pearson's R

Group	Correlation	Direction
Americans for Immigration Control	0.240	Positive
Keep America Safe	0.213	Positive
National Farmers Union	0.203	Negative
Arab American Institute	0.201	Negative

We turned age into a discrete attribute by grouping speakers into equal-frequency bins (in order to produce balanced classes) and evaluating the resulting classes using the logistic regression algorithm (10 fold cross-validation) using various feature sets. The two variables tested are (i) the number of bins, testing what age groups are formed by ideological ratings and (ii) the combination of features, testing what information is added by the various measures. The results are shown in Table 7, using Accuracy to represent performance. The majority is surpassed most in the 2-3 bin range, showing that fine distinctions in age are not possible. The ideological measures on their own (the two roll call dimensions and the three interest group measures) perform better than the interest groups ratings on their own and better than everything together. However, the relationship between age and ideology is not strong enough to justify further investigations.

Table 7. Discrete Age by Accuracy of Feature Sets

Bins	Majority Baseline	Interest Groups	Dimensions	Everything
2	51.1%	59.7%	63.26%	60.5%
3	33.7%	42.3%	43.5%	43.2%
4	25.9%	31.6%	35.1%	31.8%
5	20.8%	25.6%	28.5%	25.8%

Next, we examined the correlation between years of service in congress and ideology ratings. The scalar version was examined using Pearson's R, with the results shown in Table 8 below (and with significance defined as above; the lowest interest groups were removed to save space). Length of service had a comparatively large number of relatively weak correlations. Longer service in congress is consistently but weakly correlated with liberal scores, and shorter service in congress is consistently but equally weakly correlated with conservative scores.

Table 8. Scalar Length of Service and Interest Group Ratings using Pearson's R

Group	Correlation	Direction
Council for a Livable World	0.238	Positive
League of United Latin American Citizens	0.236	Positive
Citizens for the Rehabilitation of Errants	0.234	Positive
American Civil Liberties Union	0.232	Positive
American Immigration Lawyers' Association	0.232	Positive
Freedom Works	0.248	Negative
American Family Association	0.235	Negative
Citizens Against Government Waste	0.231	Negative
Americans for Immigration Control	0.217	Negative
Concerned Women for America	0.217	Negative

To help understand this trend, a discrete classification was tested using various equal-frequency bins (with logistic regression and 10-fold cross-validation as before). The results, shown by

Accuracy, are in Table 9 below. We see performance somewhat above the majority baseline, especially 2-3 bins, showing again an inability to make fine-grain distinctions. For 2-3 bins, the interest group ratings hold much of the information that is allowing the classification, with little added with everything included. Some of this information is lost using only the ideology dimensions (which includes both roll call and interest group measures). Again, then, there is a slight relationship between length of service (related to age) and ideology.

Table 9. Discrete Length of Service and Interest Group Ratings using Logistic Regression

Bins	Majority Baseline	Interest Groups	Dimensions	Everything
2	54.9%	62.5%	58.3%	63.0%
3	35.8%	46.0%	49.0%	48.0%
4	32.4%	35.8%	40.2%	37.0%
5	21.3%	31.6%	31.4%	32.6%

We next looked at the classification of members of congress by Chamber. Balanced classes were produced by performing a biased resampling. The results, in Table 10, showed that the distinction is easy to draw with the interest groups, both with balanced and unbalanced classes (87.39% vs. 94.59% respectively). One explanation for this could be that the roll call measures, calculated separately on the House and Senate by necessity, provide information that allows the chambers to be distinguished. However, this is clearly not the case because the classifier performs slightly better without the roll call information as features, so that the necessary information is present with only the interest group ratings.

Table 10. Classification of Chamber Membership by Accuracy

	Majority Baseline	Interest Groups	Dimensions	Everything
Unbalanced	81.39%	84.03%	80.79%	87.39%
Balanced	52.8%	94.95%	61.70%	94.59%

Given this higher performance with the interest group ratings than with only the ideology measurements, what specific groups distinguish the House and the Senate? To test this, we used the ClassifierSubsetEval algorithm in Weka (with 10 fold cross-evaluation) with the BestFirst search method. Using this method, the top features for distinguishing between the two chambers are shown in Table 11, sorted by the number of folds in which they were selected for classification (only those used for at least 6 folds are shown). One possibility is that these ratings lack representation for a significant portion of either the House or the Senate. However, this is not the case. Another interpretation is that one chamber took action or did not take action on an issue important to these groups and so many members received high or low ratings regardless of ideology. This is unlikely, however, given the span of congresses and the mix of issues represented by these interest groups.

Table 11. Top features for distinguishing between House and Senate

Interest Group	# Folds
American Civil Liberties Union	10
Americans for Immigration Control	10
United Food and Commercial Workers	7
National Journal, Liberal: Social Policy	6
Public Citizens Congress Watch	6

The classification by political party, shown in Table 12, did not need to be balanced, as the two parties are already evenly matched. Performance was high with all feature sets (from 97.8% to 99% accuracy) In addition, classification using only the dimensions has slightly higher performance than the full set of features which includes the interest group ratings, so that the party distinction is embedded in the ideological scales. For this reason, it is not necessary to ask which individual interest groups allow the distinction to be made.

Table 12. Classification of Political Party by Accuracy

	Majority Baseline	**Interest Groups**	**Dimensions**	**Everything**
Accuracy	54.0%	97.8%	99.0%	98.4%

For classification by sex, the dataset is heavily biased towards males (726 vs. 107 instances). Thus, both balanced and unbalanced tests are shown in Table 13. Balancing is again achieved using a biased resampling, and classification is done with logistic regression and 10-fold cross-validation. Without balancing, the performance is below the baseline or at the baseline (hence, the classifier simply guesses the majority). With balancing, performance surpasses the baseline but not by much (e.g., 60% for the dimensions alone with a 51.4% baseline). This tells us that ideology is slightly related to sex, but also that the ideological measurements outperform the individual interest group ratings, so that there is no need to look further at the groups.

Table 13. Classification of Sex by Accuracy

	Majority Baseline	**Interest Groups**	**Dimensions**	**Everything**
Unbalanced	87.1%	81.1%	87.1%	80.3%
Balanced	51.4%	57.2%	60.0%	59.1%

The classification according to prior military service, in Table 14, is a moderately balanced class; results are shown for both unbalanced and balanced classifications. The performance on the unbalanced classes was below the majority baseline, while the balanced classes were above the baseline. The first question is whether military service is related to a different attribute. For party, 29% of Democrats served in the military versus 33% of Republicans. For chamber, 28% of the House and 42% of the Senate served in the military. For gender, 1% of women and 35% of men served in the military. Given these differences, it is possible that the classification according to military service is related to chamber and gender classification.

Table 14. Classification of Military Service By Accuracy

	Majority Baseline	**Interest Groups**	**Dimensions**	**Everything**
Unbalanced	69.1%	65.4%	68.9%	64.8%
Balanced	51.9%	66.3%	60.0%	67.5%

Since the performance with the interest groups alone is both above the baseline and higher than the dimensions, it is important to ask which interest groups support the classification. We test for the features which help to classify military service, as above, using the ClassifierSubsetEval algorithm. Three interest group ratings are chosen in more than 6 folds, as shown in Table 15. One of these, the ACLU, was also a predictor for chamber classification, perhaps reflecting the fact that a much higher percentage of the senate has served in the military in this dataset.

Table 15. Top features for distinguishing between Military Service and No Military Service

Feature	# Folds
American Civil Liberties Union	8
Concord Coalition	8
Democrats for Life of America	6

The classification according to race, in Table 16, is heavily balanced towards white members of congress. Thus, performance on the unbalanced classes is far below the baseline of 87.9%. Because there are four minority races, the balanced classes combine all minorities into a single category before applying biased resampling. The non-white category includes African Americans, Asian Americans, Native Americans, and Hispanics together; this categorization is not ideal, but it also represents the only way to build balanced classes for this dataset. When the classes are balanced, the performance with the ideological dimensions alone reaches 75.4% accuracy; the weakness here is that the balanced classes cause a small dataset, with only 208 speakers. The first question about the features used to classify by race is the relation between race and party membership: 95% of African Americans, 78% of Asian Americans, and 84% of Hispanics are members of the Democratic party. Second, in terms of chamber, 95% of African Americans, 78% of Asian Americans, and 91% of Hispanics are members of the House. Third, in terms of gender, 70% of African Americans, 78% of Asian Americans, and 78% of Hispanics are men. Thus, race is connected with party, chamber, and gender. Because the ideological measures perform as well as the interest groups, we do not need to investigate individual interest groups.

Table 16. Classification of Race using Interest Group Ratings

	Majority Baseline	Interest Groups	Dimensions	Everything
Unbalanced	87.9%	79.8%	88.7%	81.1%
Balanced	51.9%	75%	75.4%	75.0%

The classification by religion, in Table 17, is problematic because of the large number of classes, some of which are more closely related than others. Further, the religions differ on several dimensions; for example, the theological teachings of a group may be quite different even though they represent similar social groups. We do not attempt to classify with the default classes; rather, the classes are combined and balanced. The problem with combining the groups is the large number of overlaps (e.g., many closely related groups of Protestants). All groups with at least 40 members were kept and balanced. This left eight balanced classes: Protestant, Roman Catholic, Jewish, Methodist, Presbyterian, Episcopalian, Baptist, and Not Specified; the balanced dataset is reduced to 371 members of congress. The performance is low throughout, with the classification by only ideological dimensions performing the highest with 23.1% accuracy. Given the low performance of classification by interest groups, no further investigation is necessary.

Table 17. Classification of Religion by Accuracy

	Majority Baseline	Interest Groups	Dimensions	Everything
Balanced	14.0%	16.4%	23.1%	16.1%

Classification by occupation, in Table 18, is similarly faced with a large number of unbalanced classes. All classes with below 90 members were discarded, leaving Law, Education, Public Service (including congressional aides), and Business/Banking. These were balanced as above, leaving a dataset of 384 members of congress. The results are near the majority baseline for all feature sets. The ideology measures alone have the highest performance of 33% accuracy, leaving no need to investigate the interest group ratings further.

Table 18. Classification of Occupation by Accuracy

	Majority Baseline	**Interest Groups**	**Dimensions**	**Everything**
Balanced	26.3%	27.3%	33.0%	26.0%

The point of this section has been to explore (i) possible relations between the classes which we are using to profile speakers and (ii) possible confounds in the dataset. We can conclude that there are several such confounds: several groups are very under-represented in this dataset (e.g., women and minorities). Further, these under-represented groups tend to share some ideological properties (e.g., the majority are Democrats). Thus, the classifications for these groups will be weakly trained and, perhaps more importantly, the features that are predictive of one group will, by virtue of their connection, also be predictive for another group. This is because, with few and imbalanced members of a minority present, characteristics of a few speakers will be mistakenly generalized as characteristics all of members (e.g., if most African Americans in the dataset are Democrats, the profile for Democrats will erode the profile for African Americans). Part of this problem can be solved with larger and more representative datasets which allow the adequate training and separation of these groups.

Part of this problem, however, will remain because there will always be a certain level of erosion between profiles, where one author attribute comes to dominate and erase other attributes. For example, level of education (which is not examined for this dataset, given the fact that most members of congress have the same level of education) tends to erode other profiles like socio-economic status and geographic origin as authors move toward a written standard. The best approach to this problem, in addition to using larger amounts of data, is to look at abstract features which authors are not consciously aware of and which cannot be used for stylistic changes. The difficulty is finding which features remain constant in standardized writing.

5. DW-NOMINATE Across Chambers

The DW-NOMINATE measure produces independent scales for each voting body; in other words, the scores for the House and the Senate are not directly comparable. Further, the scores for each congress are independent (although Poole & Rosenthal, 2007, show that there is little variation in individual legislators across congresses). For the purposes of this paper we have trained and tested on the House and the Senate together, for two reasons: (1) It gives us more textual data, thus allowing better modeling, especially of unbalanced classes (e.g., there are few speeches by minorities, and combining chambers helps to increase the number); (2) It gives us a more generalizable model that is less specific to the issues being debated in a given chamber during a given congress. The problem, however, is that combining the chambers in this way forces us to use independent DW-NOMINATE scores as if they were equivalent.

In order to see if we are justified in combining the House and Senate measures in this way, the Pearson's correlations between the two dimensions of the DW-NOMINATE scores and the special interest group ratings are shown for the House only, in Table 19, and for the Senate only, in Table 20. The special interest group ratings are consistent across both chambers, allowing us to use them as a baseline for comparing the DW-NOMINATE scores across the chambers. The correlations between the DW-NOMINATE dimensions and the special interest group ratings are quite similar in both chambers: -0.954 between the first dimension and the Government and Institutions component in the House and -0.958 in the Senate. The second dimension is not as consistently correlated between chambers: -0.049 between the second dimension and the Government and Individuals component in the House, but 0.224 in the Senate. Thus, the second dimension, compared across both chambers, is not a reliable measure, even though the first dimension is. However, given the limited ability of the second dimension to explain variations in roll call voting behavior, we opt to continue using both chambers in order to increase the data available for training.

Table 19. Pearson's Correlations Between Ideology Measures, House Only

	DW-N.1	**DW-N.2**	**SIG: Inst.**	**SIG: Ind.**	**SIG: Ani.**
DW-N.1	1	-.055	-.954	-.887	-.689
DW-N.2	-.055	1	-.017	-.049	-.381
SIG: Inst.	-.954	-.017	1	.887	.751
SIG: Indv.	-.887	-.049	.887	1	.635
SIG: Ani.	-.689	-.381	.751	.635	1

Table 20. Pearson's Correlations Between Ideology Measures, Senate Only

	DW-N.1	**DW-N.2**	**SIG: Inst.**	**SIG: Ind.**	**SIG: Ani.**
DW-N.1	1	-.230	-.958	-.938	-.758
DW-N.2	-.230	1	.284	.224	.019
SIG: Inst.	-.958	.284	1	.951	.782
SIG: Indv.	-.938	.224	.951	1	.776
SIG: Ani.	-.758	.019	.782	.776	1

Given that the text classification models used discretized class variables, it is also useful to look at the correspondence between DW-NOMINATE bins and the Government and Institutions component across chambers. Here we again see that DW-NOMINATE can be used as a single score across chambers without major problems: 94% of members of the House with high DW-NOMINATE 1st dimension scores also have low scores for the Government and Institutions component, compared with 100% in the Senate. Similarly, 95% of members of the House with low DW-NOMINATE 1st dimension scores also have high scores for the Government and Institutions component, compared with 89% in the Senate. These correspondences are close enough to justify combining the measures in order to increase the data set.

Table 21. Comparison of Category Memberships in House and Senate

Categories	House	Senate
% of Members in the High category for DW-NOMINATE 1	52%	44%
% of High-DW-NOMINATE 1 Members in Low category for Institutions	94%	100%
% of High-DW-NOMINATE 1 Members in High category for Institutions	6%	0%
% of Low-DW-NOMINATE 1 Members in Low category for Institutions	5%	11%
% of Low-DW-NOMINATE 1 Members in High category for Institutions	95%	89%

6. Performance of Social and Political Profile Building

Profiles for each speech for each of the thirteen attributes were built using the SVM classifier described above. In training the models, speeches from a period of four years were used and resampled with bias towards balancing a given class. In other words, for a class, like Sex, which is greatly imbalanced, only a portion of the speeches are used for training in order to train with balanced classes. Thus, each class is fully balanced for training. Numeric features were discretized using equal-frequency binning to create balanced classes. All testing is done on speeches from a period of two years (and separated by two years from the training data) and no speeches were removed (e.g., the models are tested on unbalanced data). Features were scaled between -1 and 1. Results are reported by accuracy of each class value within a profile.

The first social profile is age, shown in Table 22, which was discretized into three equal-frequency bins. As with other numeric classes, the number of bins was selected based on the distribution of the class variable. The class spans chosen by the binning process are not necessarily generational, the divide along which we would expect linguistic differences, because the older and younger generations are under-represented. Nevertheless, this is an interesting result given the large number of speeches and speakers involved.

Table 22: Age Profile By Accuracy and Class

Class	Instances	Accuracy
Before-1939	32,841	50.40%
1939-1947	25,915	24.49%
1947-Present	18,748	44.74%

Geographic location, shown in Table 23, is based on the state of a speaker's service. Thus, someone born and raised in the South could come to be elected in the North while maintaining regional characteristics of the South. The division into regions is based on state location (e.g., Ohio is a mid-western state). In this data set, the Northern speakers are more numerous and more accurately identified than the others, although performance is low throughout.

Table 23: Geographic Profile By Accuracy and Class

Class	Instances	Accuracy
Northern	20,338	33.41%
Southern	18,409	26.36%
Midwestern	19,475	25.26%
Western	19,282	28.92%

Previous military service, shown in Table 24, profiles a binary distinction between those who have served and those who have not. All forms of military service are included (e.g., active service for 20 years is the same as reserve service). The identification of speakers who have not served in the military is more accurate than of those who have served.

Table 24: Military Experience Profile By Accuracy and Class

Class	*Instances*	*Accuracy*
Yes	31,389	52.98%
No	46,115	55.07%

Race, shown in Table 25, is reduced to a distinction between White speakers and Non-White speakers (e.g., minorities and non-minorities in the U.S. context). The class of minorities is greatly under-represented in this test set; the training set was equally unbalanced and had to be greatly reduced to create balanced classes for training. Still, the overall accuracy of Race identification is over 60%.

Table 25: Race Profile By Accuracy and Class

Class	**Instances**	**Accuracy**
White	71,554	63.39%
Non-White	5,950	59.96%

Religion profiles, shown in Table 26, have a four-way distinction: Catholic Christian, Non-Catholic Christian, Jewish, and Other. The data set is dominated by Non-Catholic Christian speakers with a small number of Jewish and Other speakers. Nonetheless, the highest accuracy is found in those two categories (35%). While this accuracy is low, it is important to remember that many of these attributes are not related to the topics of discussion and may not be a salient part of the speaker's identity. In other words, given non-religious discourse, this is a decent classification result.

Table 26: Religion Profile By Accuracy and Class

Class	**Instances**	**Accuracy**
Catholic	22,024	25.21%
Non-Catholic	47,247	27.77%
Jewish	6,177	35.47%
Other	2,056	35.50%

Sex, shown in Table 27, is unbalanced in this data set, especially within the Senate. Women are more accurately identified than men (67%). This accuracy is not as high as other studies (such as Mukherjee & Liu, 2010). However, it has often been suggested that the linguistic differences between men and women are actually differences between people with power and those without power (Lakoff, 2000). In this case, all speakers are powerful individuals.

Table 27: Sex Profile By Accuracy and Class

Class	Instances	Accuracy
Male	70,000	48.58%
Female	7,504	67.05%

The profile for chamber membership, shown in Table 28, reflects the social standing of the speaker, in the sense that the Senate is more elite and patrician than the House. Thus, this is the best measure available of the socio-economic distinctions possible within an already well-educated and elite group of speakers. The classification here is quite accurate (90% for the House). This is in spite of the fact that superficial features which separate the chambers have been removed (e.g., ”Mr. Speaker” vs. ”Mr. President”). While a portion of this classification is likely to be based on differences in topics being debated in the chambers, this portion is small given that the speeches used as training came several years after the speeches used for testing, so that no particular legislation is common across both periods (thus eliminating differences in agenda as a major indicator of chamber).

Table 28: Chamber Profile By Accuracy and Class

Class	Instances	Accuracy
House	44,988	90.70%
Senate	32,516	88.68%

The first ideological profile, Political Party Membership, is shown in Table 29, making a binary distinction between Democrats and Republicans. The accuracy for detecting Democrats is 60.41%, despite being trained on data separated by the testing data by several years (and, in this case, training on speeches given with a president belonging to a different party than in the testing speeches). While this accuracy is not as high as that reported elsewhere (Yu, et al., 2008a, 2008b.

Table 29: Political Party Profile By Accuracy and Class

Class	Instances	Accuracy
Democratic	32,531	60.41%
Republican	44,231	48.96%

The profile using the first dimension of the DW-NOMINATE scores (the dimension which captures the liberal / conservative divide) is shown in Table 30. For this scale, higher scores indicate a more conservative ideology. Thus, the highest accuracy is for the more liberal group of legislators, at 63.18%. The use of two bins for this profile was based on the clearly bimodal distribution.

Table 30: DW-NOMINATE 1st Dimension Profile By Accuracy and Class

Class	Instances	Accuracy
Below 0.239	36,440	63.18%
Above 0.239	41,064	48.56%

The profile using the second dimension of the DW-NOMINATE scores (which captures the influence of issues outside of the liberal / conservative divide) is shown in Table 31. Accuracy here is much lower than for the first dimension, likely because these issues are secondary to political speech and not discussed as frequently. Three bins were chosen for this profile because the distribution was even across the range of scores.

Table 31: DW-NOMINATE 2nd Dimension Profile By Accuracy and Class

Class	Instances	Accuracy
Below -0.1795	27,583	39.55%
-0.1795 to 0.1795	26,246	29.24%
Above 0.1795	23,675	43.87%

The profile using the first component of special interest group ratings, Government and Institutions, is shown in Table 32. Here higher scores indicate higher agreement with liberal special interest groups. Thus, again, the highest performing class is that containing the most liberal legislators. The use of two bins for this profile was based on the clearly bimodal distribution.

Table 32: SIG: Institutions Profile By Accuracy and Class

Class	Instances	Accuracy
Below 41.8	40,458	48.08%
Above 41.8	37,046	63.72%

The profile using the second component of special interest groups ratings, Government and Individuals is shown in Table 33. The performance here is similar but not identical to the performance on the first special interest group measure. The use of two bins for this profile was based on the clearly bimodal distribution.

Table 33: SIG: Individuals Profile By Accuracy and Class

Class	Instances	Accuracy
Below 43.5	36,841	49.24%
Above 43.5	40,037	60.96%

The profile using the third component of the special interest groups ratings, Government and Animals, is shown in Table 34. Three bins were chosen for this profile because the distribution of scores was relatively even. The performance is less than the other interest group ratings, in part because the issues represented and the ideological community formed by those issues are not very salient in congress (e.g., these issues are marginal).

Table 34: SIG: Animals Profile By Accuracy and Class

Class	Instances	Accuracy
Below 20.3	26,227	32.95%
20.3 to 68.9	24,100	36.78%
Above 68.9	26,551	45.47%

The profiles discussed in this section range from high performance (e.g., chamber classification) to fairly low performance (e.g., geographic location classification). These classifications are interesting on their own, given the large number of authors involved, the time gap between training and testing data, and the use of new classification attributes (e.g., special interest group ratings). However, the primary purpose of the classifications is to produce a meta-profile of each speech consisting of its predicted class memberships for each of these attributes.

7. Feature Ranks Across Social and Political Profiles

We use Weka to produce feature ranks for each profile (e.g., classification by party membership) using InfoGain, which gives us a measure of how informative each feature is for the classification. This measure is comparable across profiles, so that we can directly compare the usefulness of a given feature for, for example, both party and sex classification. Table 35a shows the Pearson correlations between the InfoGain measures for each of the 21,400 textual features used in the system across all profiles. The question is whether the predictive features for each are related, so that a small number of features allows many of the classifications.

Table 35a. Pearson Correlations Between InfoGain Measures For All Features Across Classes

	Cham.	**Age**	**Geo.**	**Mil.**	**Race**	**Relig.**	**Sex**	**Party**	**DwN. 1**	**DwN. 2**	**Inst.**	**Ind.**	**Ani.**
Chamber	1	.638**	.253**	.318**	.654**	.500**	.236**	.376**	.426**	.431**	.398**	.403**	.404**
Age	.638**	1	.358**	.451**	.592**	.535**	.278**	.337**	.444**	.349**	.377**	.400**	.314**
Geography	.253**	.358**	1	.263**	.360**	.449**	.345**	.189**	.495**	.594**	.499**	.486**	.595**
Military	.318**	.451**	.263**	1	.427**	.391**	.224**	.043**	.179**	.312**	.181**	.181**	.207**
Race	.654**	.592**	.360**	.427**	1	.639**	.410**	.136**	.426**	.587**	.448**	.438**	.515**
Religion	.500**	.535**	.449**	.391**	.639**	1	.440**	.191**	.570**	.591**	.571**	.571**	.594**
Sex	.236**	.278**	.345**	.224**	.410**	.440**	1	.172**	.491**	.452**	.523**	.511**	.552**
Party	.376**	.337**	.189**	.043**	.136**	.191**	.172**	1	.402**	.144**	.377**	.382**	.290**
Dw-Nom. 1	.426**	.444**	.495**	.179**	.426**	.570**	.491**	.402**	1	.608**	.948**	.979**	.844**
Dw-Nom. 2	.431**	.349**	.594**	.312**	.587**	.591**	.452**	.144**	.608**	1	.662**	.638**	.785**
SIG: Instit.	.398**	.377**	.499**	.181**	.448**	.571**	.523**	.377**	.948**	.662**	1	.970**	.897**
SIG: Indiv.	.403**	.400**	.486**	.181**	.438**	.571**	.511**	.382**	.979**	.638**	.970**	1	.881**
SIG: Anim.	.404**	.314**	.595**	.207**	.515**	.594**	.552**	.290**	.844**	.785**	.897**	.881**	1

*** Correlation is significant at the 0.01 level, two-tailed.*

According to this measure, many of the profiles are related, ranging from a correlation of 0.979 between DW-NOMINATE 1st dimension and the Government and Individuals component (which we would expect to be correlated) and a correlation of 0.043 between party membership and previous military experience (which we would not expect to be correlated). A possible confounding factor here is that many features are not useful for any classification, and thus artificially increase the correlation. We remove this confound by removing all features which fall below the threshold of 0.01 InfoGain for every class (e.g., removing useless features), for a subset of approximately 5,000 features.

The correlations in Table 35b are more accurate, given that correlations between low-ranked features do not overpower useful features. However, the results are quite similar: for example, the correlation between Chamber and Age predicting features goes from 0.638 to 0.662. Some correlations are reduced, such as that between Chamber and SIG: Institutions, from 0.398 to 0.297. Others are increased, such as that between Chamber and Military experience, from 0.318 to 0.526.

Table 35b. Pearson Correlations Between InfoGain Measures For Features Above Threshold

	Cham.	**Age**	**Geo.**	**Mil.**	**Race**	**Relig.**	**Sex**	**Party**	**DwN.1**	**DwN.2**	**Inst.**	**Ind.**	**Ani.**
Chamber	1	.662**	.258**	.526**	.700**	.573**	.265**	.146**	.330**	.469**	.297**	.305**	.357**
Age	.662**	1	.336**	.622**	.662**	.581**	.300**	.264**	.373**	.335**	.286**	.313**	.255**
Geography	.258**	.336**	1	.299**	.389**	.482**	.350**	.215**	.524**	.589**	.507**	.504**	.597**
Military	.526**	.622**	.299**	1	.620**	.553**	.249**	.094**	.232**	.384**	.218**	.220**	.252**
Race	.700**	.662**	.389**	.620**	1	.737**	.476**	.073**	.469**	.628**	.470**	.472**	.530**
Religion	.573**	.581**	.482**	.553**	.737**	1	.546**	.177**	.609**	.659**	.604**	.606**	.637**
Sex	.265**	.300**	.350**	.249**	.476**	.546**	1	.211**	.602**	.534**	.628**	.625**	.628**
Party	.146**	.264**	.215**	.094**	.073**	.177**	.211**	1	.291**	.132**	.266**	.274**	.220**
DwN.1	.330**	.373**	.524**	.232**	.469**	.609**	.602**	.291**	1	.676**	.939**	.973**	.860**
DwN.2	.469**	.335**	.589**	.384**	.628**	.659**	.534**	.132**	.676**	1	.717**	.700**	.819**
SIG: Inst.	.297**	.286**	.507**	.218**	.470**	.604**	.628**	.266**	.939**	.717**	1	.965**	.909**
SIG: Ind.	.305**	.313**	.504**	.220**	.472**	.606**	.625**	.274**	.973**	.700**	.965**	1	.898**
SIG: Ani.	.357**	.255**	.597**	.252**	.530**	.637**	.628**	.220**	.860**	.819**	.909**	.898**	1

*** Correlation is significant at the 0.01 level, two-tailed.*

In general, we should expect that the features for classifying ideology across the different operationalizations (party membership, special interest group ratings, and DW-NOMINATE scores) should be highly correlated, and that the features for classifying social properties should not be correlated with ideology, or with each other (at least, in most cases). However, looking at both sets of correlations we see that this expectation is not always met.

First, we look at the correlations between feature information within ideological attributes. DW-NOMINATE, 1st dimension, is highly correlated with the special interest group ratings (0.939, 0.973, and 0.860), although less correlated with the 2nd dimension (0.676), although as mentioned above that measure is problematic here. The correlations between the special interest group ratings is also high, with that between Institutions and Individuals reaching 0.965. Thus, these operationalizations of political ideology are quite similar and produce similar classifications. However, the relations between party membership and these ratings is low (ranging from 0.132 to 0.291). This is somewhat unexpected. When we look below at the individual predictive features for each class, though, it will become evident that party membership crosses each of the other ideological measures, which members falling on both sides.

Second, we look at the correlations between social attributes and ideological attributes. DW-NOMINATE's 1st dimension correlates between 0.232 (Military Service) and 0.609 (Religion) with social attributes. Setting 0.5 as the mark of meaningful correlation, then, Geographic Location, Religion, and Sex are meaningfully correlated with the 1st dimension (the first two are also some of the lowing performing classifications). Looking at the Government and Institutions component, it is meaningfully correlated with Geographic Location, Religion, and Sex again. This leaves several social attributes which are relatively independent of ideology: Age, Military Experience, and Race (although meaningfully correlated with Government and Animals). Party classification, strangely, is not correlated with any other class at a level above 0.291. Thus, Party membership is the class most independent of other predictive textual features.

Table 36 shows the top two features for classification by chamber. These features are largely procedural or a matter of senate convention (e.g., "the senator from…"), and thus serve as a sort of calibration for examining the predictive features.

Table 36. Top 20 Features for Chamber Classification

InfoGain	Feature	InfoGain	Feature
0.1889387	the gentleman	0.0503596	consent that
0.1502774	gentleman from	0.0488837	balance of
0.0948060	the senate	0.0474051	NN CC JJ
0.0888486	the senator	0.0469980	IN EX PRP
0.0816476	senator from	0.0461844	NN NN
0.0656716	unanimous consent	0.0460233	NN DT VBZ
0.0604919	ask unanimous	0.0460119	myself such
0.0533602	yield myself	0.0459371	the balance
0.0505477	DT DT JJ	0.0457777	such time
0.0504252	may consume	0.0453731	DT DT VBZ

The top features for Age classification are shown in Table 37. First, several of the top features from the senate are present here (overall, the feature ranks are correlated 0.662). Second, non-sparse features such as Type / Token ratios are useful for age classification (making up 6 of the top 20 features).

Table 37. Top 20 Features for Age Classification

InfoGain	Feature	InfoGain	Feature
0.0110068	the senator	0.0025371	Std. Dev. Word Length All
0.0100347	senator from	0.0021858	IN EX PRP
0.0049544	Speech Length Lexical	0.0019983	Type/ Token Lexical
0.0049204	the gentleman	0.0019782	the gentlewoman
0.0044763	Speech Length All	0.0019681	yield the
0.0041686	gentleman from	0.0019091	necessarily absent
0.0037453	Type / Token All	0.0018412	Avg. Word Length Lexical
0.0034315	distinguished senator	0.0017925	NN NN
0.0030735	the senate	0.0017688	IN JJ CC
0.0026559	the distinguished	0.0017576	h. r.

Table 38 shows the top features for geographic location, which is a fairly lowing performing classification. With the exception of the phrase "water resources", there is no intuitive reason why these features are predictive.

Table 38. Top 20 Features for Geographic Classification

InfoGain	Feature	InfoGain	Feature
0.0030630	the senator	0.0025878	senate proceed
0.0030217	reconsider be	0.0025798	we make
0.0030193	proceed to	0.0025682	be laid
0.0029305	laid upon	0.0025551	we came
0.0029030	NNS MD TO	0.0025432	Std. Dev. Word Length Lexical
0.0028358	senator from	0.0024842	we made
0.0028057	whether or	0.0024577	CC VBZ VBG
0.0027487	NN NNS VB	0.0024318	VB CC RB
0.0027301	PRP VBD VBP	0.0023926	CC VBN TO
0.0027158	necessarily absent	0.0023804	water resources

Table 39 shows the top 20 features for classifying previous military experience, which as we saw above is largely independent of ideology classification. Several of these features are intuitively related to military affairs: “ordered to”, “the armed”, “of defense.”

Table 39. Top 20 Features for Previous Military Service Classification

InfoGain	Feature	InfoGain	Feature
0.0028947	the senator	0.0010527	other side
0.0024080	the gentleman	0.0010511	ordered to
0.0022917	senator from	0.0010312	this house
0.0017868	gentleman from	0.0010174	sure that
0.0012916	side of	0.0010103	Speech Length All
0.0012627	the aisle	0.0010002	Speech Length Lexical
0.0012391	with regard	0.0009939	the armed
0.0011870	this congress	0.0009183	necessarily absent
0.0011703	yield the	0.0009061	regard to
0.0011114	year ago	0.0008969	of defense

Table 40 shows the top features for classifying by race. Many of these are related to chamber classification, partially because the number minorities in the senate is quite low, so that a speech from the senate is likely given by a white speaker. The overall correlation between the two is 0.700.

Table 40. Top 20 Features for Race Classification

InfoGain	Feature	InfoGain	Feature
0.026006	the senator	0.008957	SpeechLengthAll
0.022229	senator from	0.007961	yield the
0.019807	the senate	0.007747	floor
0.017217	unanimous consent	0.006804	i ask
0.016050	ask unanimous	0.005865	Type / Token All
0.014877	consent that	0.005851	DT DT VBZ
0.013879	the gentleman	0.005848	senate proceed
0.011691	gentleman from	0.005765	may consume
0.009355	Speech Length Lexical	0.005706	i rise
0.009264	proceed to	0.005544	such time

Table 41 shows the top features for religion classification, another low-performing class. Several of these are procedural: “unanimous consent”, “ask unanimous”, “consent that”, “printed in”, “be printed”, “I ask”. Others are related to senate classification (“the senate”), again because minority groups (Non-Protestants here) are less common in the senate.

Table 41. Top 20 Features for Religion Classification

InfoGain	Feature	InfoGain	Feature
0.007946	unanimous consent	0.004454	printed in
0.007929	ask unanimous	0.004305	be printed
0.007464	the senate	0.004087	gentleman from
0.007254	senator from	0.003958	i ask
0.007053	consent that	0.003646	Type / Token Lexical
0.007045	Speech Length Lexical	0.003584	senate proceed
0.006985	the senator	0.003415	CC VBZ NNS
0.006932	Speech Length All	0.003301	h r
0.005748	the gentleman	0.003276	Type / Token All
0.004486	proceed to	0.003223	Std. Dev. Word Length Lexical

Table 42 shows the top features for sex classification. Non-sparse features are a significant part of the predictive features (4 of the top 20), and classification is again related to chamber membership (there are few women in the senate).

Table 42. Top 20 Features for Sex Classification

InfoGain	Feature	InfoGain	Feature
0.006115	Speech Length Lexical	0.002828	rise today
0.005804	consent that	0.002732	Type / Token Lexical
0.005624	unanimous consent	0.002728	reconsider be
0.005432	ask unanimous	0.002724	Type / Token All
0.005282	i rise	0.002654	colleagues to
0.005123	Speech Length All	0.002639	NN NNS VB
0.004368	my colleagues	0.002605	laid upon
0.003072	proceed to	0.002572	whether or
0.003034	the senate	0.002498	senate proceed
0.003033	to help	0.002469	you that

Table 43 shows the top 20 features for party classification, which is relatively successful overall (60% for Democrats). However, the feature rank is not related to any other class. They are also either non-sparse features or purely grammatical, and thus difficult to interpret.

Table 43. Top 20 Features for Party Classification

InfoGain	Feature	InfoGain	Feature
0.8936144	Std. Dev. Word Length All	0.0733243	NN CC EX
0.8609651	Std. Dev. Word Length Lexical	0.0720788	NN CC JJS
0.7507402	Avg. Word Length All	0.0716731	NN EX VBG
0.6833442	Avg. Word Length Lexical	0.0698531	NN IN
0.6317203	Type / Token All	0.0697886	NN CC RB
0.5082282	Type / Token Lexical	0.0677162	MD VB CC
0.3691924	Speech Length All	0.0676686	NN EX VBD
0.2668501	Speech Length Lexical	0.0669366	MD VB WP
0.0792967	NN EX VBP	0.0661407	NN IN DT
0.0745606	NN CC JJ	0.0659264	NN EX VBZ

Table 44 shows the top features for classifying by discretized DW-NOMINATE, 1st dimension, scores. Some of these are intuitively related to political affairs: "this administration", "the administration", "the Bush", "the Republican". A number of non-sparse features and grammatical features, difficult to interpret, are also useful. Table 45 shows the distribution of DW-NOMINATE categories across party lines, showing that it is largely but not entirely related to party.

Table 44. Top 20 Features for DW-NOMINATE 1st Dimension Classification

InfoGain	Feature	InfoGain	Feature
0.0104585	consent that	0.0047024	the republican
0.0098262	ask unanimous	0.0046263	TO NNS CC
0.0085592	unanimous consent	0.0045641	CC VBZ NNS
0.0072208	Type / Token All	0.0043849	senate proceed
0.0069978	Speech Length All	0.0043599	it is
0.0065689	Speech Length Lexical	0.0043450	the bush
0.0060247	this administration	0.0043372	in the
0.0057780	Type / Token Lexical	0.0042725	NN NN IN
0.0052560	proceed to	0.0042513	reconsider be
0.0052518	the administration	0.0041239	laid upon

Table 45. Number of DW-NOMINATE 1st Dimension Speakers by Party Membership

Category	Republican	Democrat
Low	33	377
High	417	4

Table 46. Top 20 Features for DW-NOMINATE 2nd Dimension Classification

InfoGain	Feature	InfoGain	Feature
0.0071454	ask unanimous	0.0035051	whether or
0.0068227	unanimous consent	0.0033829	NNS MD TO
0.0063181	proceed to	0.0033710	PRP VBD VBP
0.0061880	consent that	0.0031360	any statements
0.0048255	the senate	0.0029449	consideration of
0.0043414	reconsider be	0.0029184	CC VBZ VBG
0.0043347	senate proceed	0.0029047	gentleman from
0.0041289	be laid	0.0028662	we make
0.0041070	laid upon	0.0028630	CC VBN TO
0.0040560	the gentleman	0.0028433	we made

Table 46 shows the top features for classifying by DW-NOMINATE, 2nd dimension, a somewhat problematic measure for this dataset, as discussed above. The distribution is more spread across party lines, as shown in Table 47.

Table 47. Number of DW-NOMINATE 2nd Dimension Speakers by Party Membership

Category	Republican	Democrat
Low	138	98
Middle	197	116
High	115	167

Table 48 shows the top features for the Government and Institutions component classification, with distribution across party lines as shown in Table 49. Several of these are political in nature, and others purely linguistic.

Table 48. Top 20 Features for SIG: Institutions Classification

InfoGain	Feature	InfoGain	Feature
0.0141944	consent that	0.0062523	TO NNS CC
0.0139459	ask unanimous	0.0061281	authorized to
0.0123540	unanimous consent	0.0059758	CC VBZ NNS
0.0082628	Speech Length All	0.0058284	NN NN IN
0.0079231	Type / Token All	0.0057083	senate proceed
0.0076904	Speech Length Lexical	0.0055061	reconsider be
0.0071366	Type / Token Lexical	0.0053174	laid upon
0.0070298	proceed to	0.0052605	NNS MD TO
0.0064754	be authorized	0.0051382	be laid
0.0064053	this administration	0.0051354	the republican

Table 49. Number of SIG: Institutions Speakers by Party Membership

Category	Republican	Democrat
Low	419	5
High	31	376

Table 50 shows the top features for classifying by the Government and Individuals component, with the distribution across party lines shown in Table 51. Many of these ("the Bush") are political in nature, others are not.

Table 50. Top 20 Features for SIG: Individuals Classification

InfoGain	Feature	InfoGain	Feature
0.0133613	consent that	0.0058363	TO NNS CC
0.0130095	ask unanimous	0.0057535	authorized to
0.0115816	unanimous consent	0.0056963	CC VBZ NNS
0.0081677	SpeechLengthAll	0.0055824	the republican
0.0080616	TypeTokenAll	0.0055351	the administration
0.0079821	SpeechLengthLexical	0.0052760	NN NN IN
0.0070991	TypeTokenLexical	0.0052060	a m
0.0068672	this administration	0.0051866	senate proceed
0.0062862	proceed to	0.0050723	it is
0.0061032	be authorized	0.0049875	the bush

Table 51. Number of SIG: Individuals Speakers by Party Membership

Category	Republican	Democrat
Low	444	0
High	15	372

The top features for classification by the Government and Animals component is shown in Table 52, with the distribution of speakers across parties in Table 53. This component is less closely aligned with political parties than the other special interest group components. Many of the features, such as "unanimous consent" and "consent that" have been features for previous classifications, in this case Sex, Race, and Religion among others.

Table 52. Top 20 Features for SIG: Animals Classification

InfoGain	Feature	InfoGain	Feature
0.0159138	ask unanimous	0.0064039	CC VBZ NNS
0.0157077	consent that	0.0063885	be laid
0.0149636	unanimous consent	0.0062986	NNS MD TO
0.0095516	proceed to	0.0061783	whether or
0.0076778	Speech Length All	0.0060593	PRP VBD VBP
0.0075718	senate proceed	0.0059376	with state
0.0074046	Speech Length Lexical	0.0057283	we came
0.0070626	reconsider be	0.0057122	a. m.
0.0068713	laid upon	0.0057110	Std. Dev. Word Length Lexical
0.0067107	TO NNS CC	0.0056656	Type / Token All

Table 53. Number of SIG: Animals Speakers by Party Membership

Category	Republican	Democrat
Low	239	20
Middle	144	143
High	67	218

7. Discussion

The purpose of this paper has been to test the classification of texts according to the political ideology of the speaker, using data from the *Congressional Record.* We have used three different operationalizations of political ideology: party membership, aggregate special interest group ratings, and DW-NOMINATE scores. Special interest group ratings and DW-NOMINATE scores are very highly correlated across speakers and share many of the same predictive textual features. Party is closely related to both, although the party classification was not highly correlated with the others in terms of which features are predictive. For each of these operationalizations of political ideology, the classification accuracy was moderately successful, given that the system operates at the level of individual texts, many of which are short and dominated by formulaic material.

Given this, we wanted to test both the generalizability and the independence of the ideological classifications. First, in order to be meaningful the classifier needs to be finding generalizations that reach across time and particular issues, rather than simply finding key slogans from a particular debate. We tested generalizability in two ways: (1) by separating the training and testing sets by a whole congress, (2) by training and testing on both chambers together. Both of these choices mean that the classifier needs to find textual features which go beyond specific debates and election cycles. The results of this test are inherent in the profiling accuracy: while moderately successful, the results are not as high as in other work. Thus, we see that the more we separate the training data and the testing data, the worse the features perform.

Second, in order to be meaningful, the classifier needs to detect the political ideology of the speaker but not confuse ideology with other non-ideological attributes. We tested this by classifying texts according to non-ideological social attributes of the speakers: Age, Sex, Race, Religion, Geographic Location, and Previous Military Experience. This is important because social factors influence an individual's linguistic patterns. Thus, we need to be sure that the features which we believe are predictive of ideology are independent of non-ideological classifications, especially given a skewed data set such as this one. What we find is that the linguistic features which are predictive of ideology are also predictive of Geographic Location, Religion, Sex, and, to a lesser degree, Race. There are two ways to interpret this correlation: First, it is possible that what we think is ideology is in fact a confound for other social attributes. For example, if most minorities in congress are Democrats, and if minorities have distinct linguistic patterns, then those linguistic patterns will be detected by the ideology classifier even though they are unrelated to ideology. Second, it is possible that the confound goes in the opposite direction: given the standardized language used in these speeches, perhaps there are no linguistic differences between speakers along social variables. Thus, in this interpretation, the classification by religion, sex, race, and geographic location is just a by-product of the fact that individuals in one class tend to belong to a single ideological group.

These two questions, generalizability and independence, are essential for using text as data for studying political ideology. This paper has raised the issues and provided evidence that is potentially problematic for both. However, the results here are not sufficient to answer these questions. For that we need a larger study which uses text from multiple sources (e.g., the Canadian or European parliaments) in order to truly compare generalizability and independence. Further, we need data from non-official political speech in order to see the degree to which linguistic patterns reflect both social and ideological attributes, using genres which are less formal and thus more likely to show socially-patterned variations.

Works Cited

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*, 42(01), 31–55.

Fan, R., Chang, K., Hsieh, C., Wang, X., & Lin, C. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9, 1871–1874.

Gentzkow, M., & Shapiro, J. (2013). *Congressional Record for 104th-109th Congresses: Text and Phrase Counts. ICPSR33501-v2*. Ann Arbor, MI.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), 10. doi:10.1145/1656274.1656278

Iyyer, M., Enns, P., Boyd-graber, J., & Resnik, P. (2014). Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* Stroudsburg, PA: Association for Computational Linguistics. 1113–1122.

Koppel, M., Akiva, N., Alshech, E., & Bar, K. (2009). Automatically Classifying Documents by Ideological and Organizational Affiliation. *2009 IEEE International Conference on Intelligence and Security Informatics*, 176–178.

Lakoff, R. (2000). *The Language War.* Oakland, CA: University of California Press.

Laver, M., Benoit, K., Garry, J., & Trinity, K. B. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *The American Political Science Review*, 97(2), 311–331.

McCutcheon, C., & Lyons, C. (2010). *CQ's politics in America [Several Years] The [104th …. 109th] Congress.* Washington, DC: Congressional Quarterly.

Mukherjee, A., & Liu, B. (2010). Improving Gender Classification of Blog Authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Stroudsburg, PA: Association for Computational Linguistics. 207–217.

Poole, K., & Rosenthal, H. (2007). *Ideology and Congress.* Edison, NJ: Transaction Publishers.

Sim, Y., Acree, B., Gross, J., & Smith, N. (2013). Measuring ideological proportions in political speeches. *Proceedings of EMNLP 2013.* 91–101.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of EMNLP 2006.*

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, 5(1), 33–48.

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 2008 International Conference on Digital Government Research.*