

Classifying Party Affiliation from Political Speech

Bei Yu
Stefan Kaufmann
Daniel Diermeier

Abstract:

In this paper we discuss the design of party classifiers for Congressional speech data. We then examine the party classifiers' person-dependency and time-dependency. We found that party classifiers trained on 2005 House speeches can be generalized to the Senate speeches of the same year, but not vice versa. The classifiers trained on 2005 House speeches perform better on Senate speeches from recent years than older ones, which indicates the classifiers' time-dependency. This dependency may be caused by changes in the issue agenda or the ideological composition of Congress.

Keywords: machine learning, text classification, generalizability, ideology, evaluation

Notes:

Bei Yu (bei-yu@northwestern.edu) is a postdoctoral fellow in the Ford Motor Company Center for Global Citizenship, Kellogg School of Management and Northwestern Institute on Complex Systems (NICO), Northwestern University.

Stefan Kaufmann (kaufmann@northwestern.edu) is an assistant professor in the Department of Linguistics at Northwestern University.

Daniel Diermeier (d-diermeier@kellogg.northwestern.edu) is the IBM Distinguished Professor of Regulation and Competitive Practices in the Department of Managerial Economics and Decision Sciences (MEDS), Ford Motor Company Center for Global Citizenship, Kellogg School of Management and Northwestern Institute on Complex Systems (NICO), Northwestern University.

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

Introduction

Political text has been an underutilized source of data in political science, in part due to the lack of rigorous methods to extract and process relevant information in a systematic fashion. Recent advances in text mining and natural language processing techniques have provided new tools for analyzing political language in various domains related to digital government initiatives and political science research (Laver, Benoit and Garry 2003; Quinn et al. 2006; Diermeier et al. 2007; Evans et al. 2005; Thomas, Pang and Lee 2006; Kwon et al. 2006). Some of the texts available in this domain are well-prepared speech or formally written texts, such as the Congressional record, party manifestos, or legislative bills. Some are less formal, such as email feedback on government policy from the general public as well as newsgroup discussions and blogs on political issues.

Automatic text classification is a widely used approach in the computational analysis of text. In the context of political speech a common goal, especially among computer scientists, has been the construction of general-purpose political opinion classifiers because of their potential applications in e-Rulemaking and mass media analysis (Shulman 2005; Agrawal et al. 2003; Kwon et al. 2006; Thomas, Pang and Lee 2006). The goal of political opinion classification is to correctly sort political texts depending on whether they support or oppose a given political issue under discussion. This task is closely related to the sentiment classification work which has been in progress for more than ten years (Esuli, 2006), and most of which has focused on commercial domains such as customer reviews. Opinion classifiers have achieved good classification accuracies (>80%) in some text domains with strong expressive content, such as movie and customer reviews (Pang, Lee and Vaithyanathan 2002; Dave, Lawrence and Pennock 2003; Hu

and Liu 2004). In the political context, this line of research is trying to apply the same methodology to political text. A potential difficulty facing this approach is that in political texts, especially professional political speech, opinions are usually expressed much more indirectly. To illustrate, we may quote from expressive movie reviews and the more deliberative congressional speech for comparison. Below are a few opening sentences from sample movie reviews¹.

“Kolya is one of the richest films I’ve seen in some time.”

“Today, war became a reality to me after seeing a screening of Saving Private Ryan.”

“Let’s face it: since Waterworld floated by, the summer movie season has grown very stale.”

However, no similarly expressive language can be found in the following comment on the Partial Birth Ban Act², despite the fact that the issue was highly emotional and controversial. Nevertheless, an educated reader can easily infer that this speaker is opposing the bill. The message conveyed is one of annoyance and “waste of time,” presumably because more important issues do not get tackled during the available time to debate.

“Mrs. MURRAY. Madam President, here we are, once again debating this issue. Since we began debating how to criminalize women's health choices yesterday, the Dow Jones has dropped 170 points; we are 1 day closer to a war in Iraq; we have done nothing to stimulate the economy or create any new jobs or provide any more health coverage. But here we are, debating abortion in a time of national crisis.”

¹ The movie reviews are downloaded from <http://www.cs.cornell.edu/People/pabo/movie-review-data/> (last visit: October 31, 2007)

² The Congressional speech data are downloaded from <http://thomas.loc.gov/> (last visit: October 31, 2007)

In related work (Yu, Kaufman, and Diermeier 2008) we have investigated whether the opinion classification approach favored by computer scientists offers a promising direction for the study of political speech. We found that standard methods that work well in opinion classification face a number of difficulties in this new domain. First, political speech uses far fewer of the sentiment words - typically adjectives or adverbs – that have been found to be most indicative of opinion in, say, movie reviews. Instead, opinion in political speeches tends to be expressed by the choice of nouns. Second, nouns which carry no political meaning in common usage may do so in the context of a particular debate. For example, in the debate on the Partial Birth Ban Act, opponents of the bill frequently used medical or technical terms rather than the more emotive term “abortion.” The use of medical terms in general does not signal a particular political position, but it does signal a pro-choice position (is understood as such) in the context of this. Third, classification success as measured against both voting decisions of the speakers and manual annotations of the speeches is worse than in the case of consumer reviews.

In this paper, we propose an alternative approach based on the concept of political ideology. In a political setting, a person’s opinion on a given issue can be expected to depend on his or her underlying ideology rather than common standards as may be more typical of commercial speech. Ideologies give structure to an individual’s view on various issues. Intuitively, an ideology expresses a view of which issue positions go together, the “knowledge of what-goes-with-what” (Poole 2003). In other words, ideology will shape each individual’s views on given issues, and these influences will be identifiably different for, e.g. Liberals and Conservatives (see Figure 1).

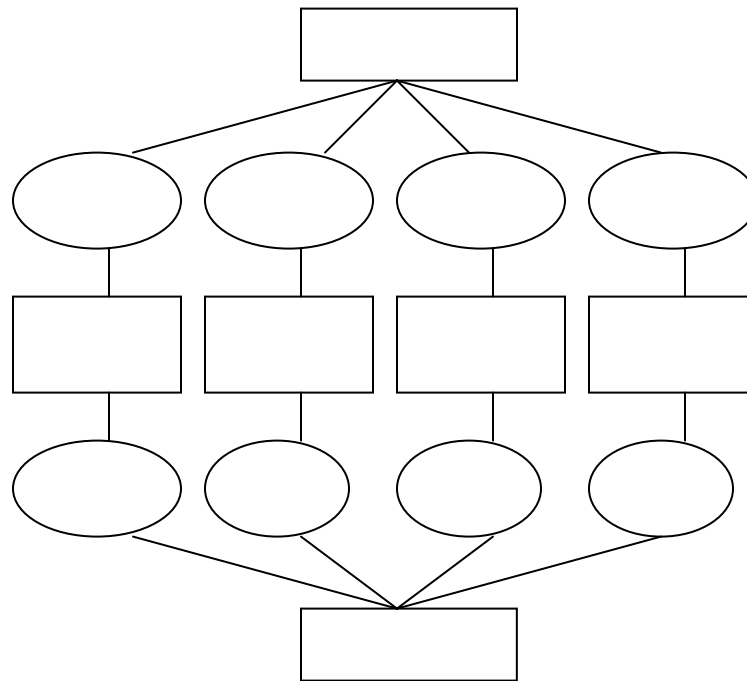


Figure 1: The relation between ideology and opinions on various issues

For our purposes, the importance of political ideology suggests a new research orientation. Rather than classifying isolated opinions, this approach would focus on classifying the underlying ideology of the person who holds the opinion. Underlying this approach is the hypothesis that ideologies give coherence to a person's opinions and attitudes, so that once we have properly identified a person's ideology, we may be able to predict his or her opinions on new or modified issues. In a highly influential essay, Converse (1964) viewed ideologies as "belief systems" that constrain the opinions and attitudes of individuals.

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

For example, we know that in the U.S., liberal lawmakers favor fewer regulations of personal behavior and higher levels of income redistribution. We also know that conservatives typically favor more regulations of private personal behavior and fewer economic restrictions. The coherence is particularly striking if we restrict attention to issues of morality, culture and the like. A legislator who is voting to oppose gun control is also likely to limit abortion rights and vice versa. We can, of course, imagine a libertarian position which favors lower restrictions in both the economic and the personal domains -- e.g., one which opposes labor regulations and restrictions on marijuana use. These positions, however, are not represented in Congress to a significant degree, nor do they resonate widely in public discourse.³

While ideology is a potentially promising organizing principle of political opinions, at least among political elites, it creates new challenges. Most importantly, ideology is not directly observable, which makes ideology identification and measurement difficult. Consequently, scholars have employed different strategies, ranging from survey responses to statistical estimates based on voting records. Poole and Rosenthal (1997) find that over the history of the U.S. Congress a two-dimensional spatial model (estimated with D-NOMINATE scores) can correctly classify about 85 percent of the individual voting decisions of each member of Congress. Moreover, for most periods of American history, a single dimension is sufficient.

³ Understanding why certain ideologies resonate is an interesting research question in itself. For some recent approach from the perspective of cognitive linguistics, see Lakoff (2002).

Recently, these approaches have been extended to political speech, as both voting and speech can be understood as expressions of a common underlying belief system (Monroe and Maeda 2004; Laver, Benoit and Garry 2003; Diermeier et al. 2007). Indeed, one may argue that speech is a richer kind of data, since speech during a Congressional debate is less constrained by institutional rules compared to voting. With the digitization of government documents, large volumes of congressional records (from the 101st Congress to date) have become publicly accessible through the Thomas database⁴, which provides ideal data for ideology analysis in speech. The goal is to use text classification as an analytical tool to probe whether ideology constrains political speech as well as other kinds of political expression.

The use of text classification as an analytical tool is not unique to the political science domain. Humanist scholars have been employing it for many years, most importantly in the area of identifying literary *style*. Craig (1999) once explained the connection between authorship attribution and stylistic analysis as two sides of a coin - you have learned something about the authors' stylistic differences if you can tell them apart. Similarly, if we achieve high accuracy in ideology classification, we can surmise that the classifier has learned something significant about the patterns that make texts conservative or liberal. We can then extract these patterns to see if they make sense in the political science context. Currently, the text data explored in related studies are mostly formal discourse, such as Senatorial speech (Diermeier, Godbout, Yu, and Kaufmann 2007), Supreme Court briefs (Evans et al. 2005), and party manifestos (Laver, Benoit, and Gary 2003). These studies all achieve high classification accuracy on their data sets, which suggests that detectable patterns associated with ideological orientation do exist at least in these formal genres of political discourse.

⁴ The URL for the database is <http://thomas.loc.gov/> (last accessed 10/30/2007).

As an example, in a previous study (Diermeier, Godbout, Yu, and Kaufmann 2007) we used the signs of Senators' D-NOMINATE scores to label ideology categories (liberal or conservative) of Senatorial speeches from the 101st-108th Congresses⁵. Speeches of the 25 most conservative and the 25 most liberal Senators (as measured by their D- NOMINATE scores) in each of the 101st-107th Congresses were selected as training data, and the 50 corresponding "extreme" Senators in the 108th Congress were used as test data. We used an SVM algorithm to train an ideology classifier and observed high classification accuracy both within the training set (through 5-fold cross validation) and on the test set. The purpose of using the 108th Senatorial speech as the test set was to examine whether classifiers trained on speeches on old issues can predict the positions on new issues, as implied by the notion of ideologies as a belief *system*.

In addition to classifying "extreme" Senators correctly, our approach also allowed us to explore why this persistence across different Congresses occurs and whether it indeed reflects coherence in belief systems. Using feature analysis, we found that the key issues discussed by liberals are energy and the environment, corporate interests and lobbying, health care, inequality and education. For conservatives, the key issues discussed are taxation, abortion, stem cell research, family values, defense, and government administration. Furthermore, the two sides often choose different words to represent the same issue. For example, among the adjectives most indicative of Democratic positions we find the word *gay*, whereas for Republicans we find the word *homosexual*.

While these results are encouraging, we need to verify whether they are indeed indicative of an underlying ideology. While we cannot observe ideologies directly, the concept of ideologies as coherent and constraining belief systems has various testable implications. First,

⁵ Appendix A describes the details of downloading and processing the Senatorial speech data from the government website www.thomas.gov.

ideologies need to be fairly stable across issues and over time. Empirically, this means that a hypothesized ideology needs to reliably predict positions on other issues and in future periods. Second, while ideologies will be held by specific persons, they cannot be overly person-specific. The concept would lose its usefulness in political discourse if every person had their own ideology. Rather, ideologies are considered as applying to groups of people, e.g., members of the same political party or movement. In other words, knowing the position of one conservative Senator should make it easier to predict the positions of other conservative Senators than Liberal ones.

A limitation of our existing results is that it was difficult to evaluate these characteristics within the Senatorial speech data alone, since it was impossible to control all three sources of variation - person, issue, and time - within the same data set. For example, most of the Senators in the 108th Congress were also Senators in previous Congresses. While our classifier performed well on the speeches of those Senators that were new in the 108th Congress (4 out of 5 are correctly classified), that sample is too small to draw reliable inferences. On the other hand, removing from the training data those speeches that were given by speakers who were still Senators in the 108th Congress resulted in a lack of speeches from recent years in the training set. Hence the person and time factors cannot be separated in a satisfactory way. Previous work (e.g. Quinn et al. 2006) has shown that the issues discussed in Congress vary substantially from year to year. While this suggests that our estimates do a good job in identifying ideology across time and (if the Quinn et al. results are correct) over issues, it does not constitute a direct test.

Our goal in the present study was to control the person and time factors by using speeches from both House and Senate. Obtaining the 2005 House speech data from Thomas et al. (2006), we first tested our ideology classifiers' generalizability across House representatives and

Senators of the same year (2005). We ran a cross-evaluation consisting of two tests. In the first test, we trained ideology classifiers on speeches of 2005 House representatives and tested these classifiers on speeches from the 2005 Senate. In the second test, we switched the training data and the test data. If high prediction accuracies are observed in the cross-evaluation, it is evident that the ideology classifiers trained on one group of legislators can be generalized to the other group.

We then tested the cross-time generalizability of our approach by using speeches from different years in the House and the Senate for training and testing. For example, we trained ideology classifiers on 2005 House data and tested these classifiers on Senate data from 2005 and other years. Stable prediction accuracies over time will provide evidence that the ideology classifiers can be generalized to speech data in different periods, otherwise the classifiers are time-dependent.

One potential difficulty for this approach lies in the fact that roll-call based measures, such as the D-NOMINATE scores, may not be directly comparable across chambers due to the fact that the each chamber may have decided on a different universe of bills. To avoid this problem, we use party membership (Democrat and Republican) as our classification categories. Previous work on voting behavior (e.g. Poole and Rosenthal 1997) has shown that party affiliation is a reasonably reliable measure for ideological orientation, especially for legislators with extreme positions as analyzed in Diermeier, Godbout, Yu, and Kaufmann (2007).

This paper is organized as follows. We firstly introduce the text classification process, the text classification methods, and the evaluation measures used in this study. Then we report a series of generalizability evaluation experiments and results. Before concluding, we discuss

specific challenges in evaluating classifier generalizability and its relationship to data assumption violations in text classification experiment design.

The text classification process

As in other domains, a political text classification problem involves data cleaning and preparation, knowledge discovery, and interpretation and evaluation steps. It is often an iterative process with multiple rounds of experiments (see Figure 2). For text classification, firstly a sample set of text data is drawn from a large text collection of interest. For example, we may choose the speeches of the 108th Senate as a sample set of the whole Congressional speech collection. Each document in the sample is then mapped to a numerical document vector, usually a vector of counts of certain linguistic patterns, such as occurrences of words and phrases. Furthermore, each document in the sample is labeled as belonging to one of the categories that define the classification task. In some cases, this categorization is subjective, e.g., based on the judgments of human. However, in this study we used an objective criterion, viz. the speakers' party affiliation.

Once all sample documents are associated with vector representations and category labels, we designate a classification method to train a classifier on the sample data. Cross validation or hold-out tests are often used to estimate the classifier's generalization error, which is the expected error rate when the classifier is used to classify new data. After all, the classifier is meant to classify the whole political text collection from which the sample data set was drawn.

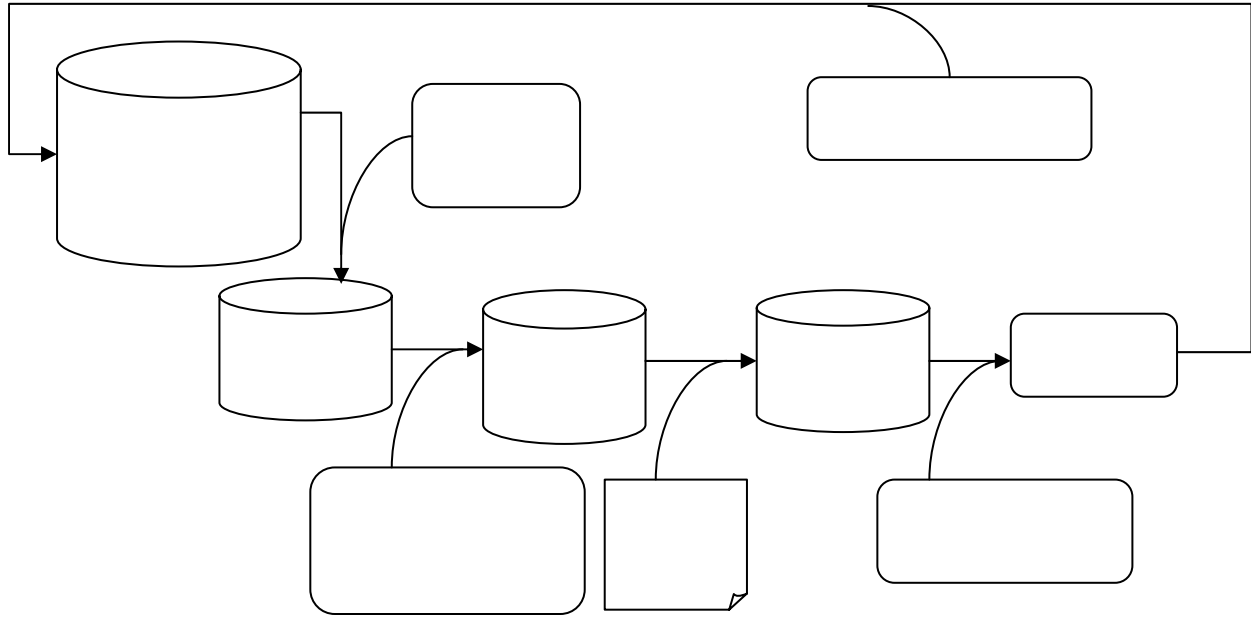


Figure 2: The text classification process

Ideology classification experiment design

Figure 2 also shows that there are many choices to make in the design of text classification experiments, such as the sampling method, the text representation model, the label acquisition, the classification methods, and the evaluation metric. Without any prior knowledge of the particular classification problem, we start with the simplest text representation, the Bag-of-Words (BOW) approach, which maps each document to a vector of word occurrence counts in that document. Rare words (frequency<3) and overly common words (the 50 most frequent ones in the data set) are removed from the vocabulary represented in these vectors.

For classification applications, some classes are easy to separate for most algorithms. In many cases, however, the data sets have some characteristics which favor some methods over others. Therefore it is common to try multiple algorithms on a new data set. In this study we

chose Support Vector Machines (SVMs) and naïve Bayes (NB) algorithms to train ideology classifiers. According to a number of classification algorithm comparison studies, naïve Bayes and SVM are among the most widely used text classification methods (Sebastiani 2002; Dumais et al. 1998; Joachims 1998, Yang and Liu 1999). Existing comparison results show that SVMs are among the best text classification methods to date. Naïve Bayes is a highly practical Bayesian learning method (Domingos and Pazzani 1997). It is a simple but effective method, often used as a baseline algorithm. SVM and naïve Bayes are also the most popular classification algorithms in current political text classification studies (Kwon et al. 2006; Thomas, Pang and Lee 2006; Evans et al. 2005).

We used the SVM-light package⁶ and its default parameter settings as the implementation of the SVM algorithm in this study. SVM allows for the use of various kinds of word frequency measures in calculating document vectors, resulting in different models for the same data set. We combined SVM with three different frequency measures. The first one is “svm-bool,” which uses simple presence or absence of each vocabulary word in the document. The second one is “svm-ntf,” which uses the normalized word frequency. The third one is “svm-tfidf,” which uses word frequency weighted by inverse document frequency.

We implement two variations of naïve Bayes algorithms which were described in Mitchell (1997). The first one uses word presence and absence as feature value (“nb-bool”). The second one uses word frequency (“nb-tf”). These two methods are also called the multi-variate Bernoulli model and the multinomial model, respectively (McCallum and Nigam 1998).

Table 1 summarizes the five classification methods used in this study. For a given training data set, each method will generate a different classifier. We evaluated the five classifiers’ person-dependency and time-dependency in parallel.

⁶ This software can be downloaded from <http://svmlight.joachims.org/>.

Table 1: Variations of SVM and naïve Bayes classification methods

		Feature values			
		presence/absence	frequency	normalized frequency	idf-weighted frequency
Algorithms	SVM	svm-bool	n/a	svm-ntf	svm-tfidf
	naïve Bayes	nb-bool	nb-tf	n/a	n/a

Cross-validation and hold-out tests are the usual methods for classification result evaluation. *N-fold cross-validation* partitions a data set into N folds and runs classification experiment N times, each time using one fold of data as the test set and training the classifier on the remaining N-1 folds. The classification accuracy is averaged over the results of the N runs. In a *hold-out test*, the data set is divided into a training subset and a test subset. A *leave-one-out* test is a special case of N-fold cross-validation, where N equals the number of documents in the whole data set. For small data sets, an arbitrary train/test split might result in both small training and test sets, potentially yielding varied results for different ways of splitting. Therefore leave-one-out evaluation is often used for small data sets. We used both leave-one-out cross-validation and hold-out tests in our study.

Evaluation of ideology classifiers' time and person dependency

In the introduction, we briefly discussed the ideology classification results of our previous study, in which we demonstrated that SVM-based ideology classifiers trained on speeches from the 101st-107th Senate can effectively predict the ideologies of speeches from the 108th Senate as measured by D-NOMINATE scores as well as their party affiliation. In this section, we discuss a

series of experiments designed to evaluate the ideology classifiers' person-dependency and time-dependency.

Our first experiment is intended to test whether our inferred ideology classifiers exhibit too much person-dependency, i.e., whether they are essentially person classifiers. Recall that in the Congressional context the notion of ideology should properly be understood as a *shared* belief system. Our approach is to design an experiment which (to the extent possible) keeps time and issues constant while varying the set of individuals. Specifically, we exploit the bicameral structure of the U.S. Congress, using one chamber as the training set and the other as the test set. To control for issue similarity we only use data from one year. While this does not perfectly control issue similarity –the two chambers set their own agendas- due to the fact that both chambers have to agree on each proposed bill for it to become law, we can expect substantial overlap between the two agendas. The task is to correctly classify party affiliation.

We use the 2005 Congressional speeches in the House⁷ and the Senate, here labeled as “2005House” and “2005Senate.” In addition to within-chamber validation tests, we also run a cross evaluation which consists of two tests: 1) training classifiers on the “2005House” data and testing them on the “2005Senate” data; and 2) training classifiers on the “2005Senate” data and testing them on the “2005House” data. By this design, we ensure that the training and test data are produced by two groups of speakers without overlap, yet that the issues under discussion are highly similar because the speeches were given in the same Congress in the same year.

⁷ We used the 2005 House debate corpus from (Thomas et al., 2006) as the “2005House” data set. This corpus includes the 2005 House debates on 53 controversial bills. Controversial bills are defined as the losing side (according to the voting records) generated at least 20% of the speeches. Thomas et al. (2006) split the selected debates into three subsets (training, test and development). We merge the three subsets into one whole data set to maximize the amount of data to use. In the whole data set 377 House representatives have speeches included in the corpus. We concatenated each speaker's speeches as one document. Thus we have 377 examples in the “2005House” data set.

There are three possible findings. First, neither direction leads to high classification accuracy. In that case we would have to conclude that our classifier is too connected to individual or chamber characteristics. The critical feature of cross-person accuracy would be lacking. Second, classification leads to high accuracy in both directions. In that case we have evidence for having identified features of party ideology that operate at the group level. Third, the classification works in one direction, but not in the other. This is an important case, which we also encountered in Diermeier, Godbout, Yu, and Kaufmann (2007). In that study, we found that using the speeches of ideologically extreme Senators as test data allowed us to classify moderate Senators well, but not vice versa. We interpreted this as evidence that the ideology of extreme Senators is more well-defined compared to the more “blurry” or mixed ideology of moderates. We can test this hypothesis in the current cross-chamber design. As the House is commonly believed to be more partisan than the Senate, this would imply that training on the House data should predict Senate data much better than vice versa. Any other findings (better accuracy in the reverse case or the same accuracy) would cast doubt on this hypothesis.

We firstly train SVM and NB classifiers on the “2005House” data and test the classifiers on the “2005Senate” data. We then switch the training and testing data and repeat the experiment.

Table 2 lists the results of the “2005 House to Senate” experiment. The first column shows the five classifiers’ leave-one-out cross validation accuracies on “2005House.” The accuracies range from 70% to 80%. The second column shows these classifiers’ prediction accuracies on “2005Senate.” Three classifiers (“svm-bool”, “svm-tfidf”, and “nb-tf”) achieve over 80% prediction accuracies, which demonstrates that they are not likely person-dependent. Appendix B presents three tables which list the most discriminative word features learned by the three classifiers respectively. Similar to the feature analysis result in (Diermeier, Godbout, Yu,

and Kaufmann 2007), these features indicate the key issues discussed by liberals/democrats and conservatives/republicans. The “nb-bool” classifier performs worse than the majority baseline. The svm-ntf classifier is better than the majority baseline⁸, but not as successful as the other three methods.

Table 2: 2005 "House to Senate" classification accuracies (in percent)

	2005 House cross validation	2005 Senate prediction
majority baseline	51.5	55.0
svm-bool	75.1	88.0
svm-ntf	69.8	63.0
svm-tfidf	80.1	81.0
nb-bool	77.9	50.0
nb-tf	78.7	83.0

Table 3 lists the results of the “2005 Senate to House” experiment. The first column shows the five classifiers’ leave-one-out cross validation accuracies on “2005Senate.” Svm-ntf still performs the poorest among the five classifiers. Its performance is almost the same as the majority baseline. The cross-validation accuracies for the other four classifiers range from 70% to 86%, similar to the range in the “2005 House to Senate” test. The second column shows these classifiers’ prediction accuracies on “2005House.” Three classifiers (“svm-bool”, “svm-ntf”, “nb-bool”) degrade to majority vote by assigning all test examples to the majority class. The “svm-tfidf” and “nb-tf” classifiers are better than the majority baseline, but their accuracies are much lower than that of their counterparts in the last “2005 House to Senate” test.

⁸ “Majority baseline” is a trivial classification method which is often used as a baseline in algorithm performance evaluation. This method predicts the class membership of any test example as the class which contains the majority of the training examples. For example, if a data set consists of 55 positive examples and 45 negative examples, the majority baseline is 55%.

Table 3: 2005 "Senate to House classification accuracies (in percent)

	2005 Senate cross validation	2005 House prediction
majority baseline	55.0	51.5
svm-bool	73.7	51.5
svm-ntf	55.6	51.5
svm-tfidf	69.7	65.8
nb-bool	81.0	51.5
nb-tf	86.0	67.6

The results in Tables 2 and 3 indicate that overall, the “2005 House to Senate” prediction results are better than the “2005 Senate to House” prediction results. This finding supports the hypothesis that the House is more partisan than the Senate. However, in the “2005 Senate to House” experiment, the two naïve Bayes classifiers still achieve over 80% cross validation accuracies on “2005Senate,” which means the “2005Senate” data can be well separated by naïve Bayes methods. The comparatively poor performance of these naïve Bayes classifiers on the “2005House” data is probably due to overfitting of the “2005Senate” training data. In other words, they are more person-dependent. A big difference between the two data sets is that “2005Senate” has only 100 examples, while “2005House” has 377. It would therefore not be surprising if a classifier captures some chamber characteristics which fit the Senate but not the House.

The results of our first experiment demonstrate that the House speeches are better suited than the Senatorial speeches to the task of training person-independent ideology classifiers. We next move on to test whether the 2005House-trained ideology classifiers are time-independent as well. In our second experiment, we test the 2005House-trained ideology classifiers on the Senatorial speeches from 1989 through 2006. Each year’s Senatorial speeches constitute one test set. There are 18 test sets in total, each by about 100 senators. We run the test 18 times, once for

each year. Table 4 shows the classifiers' prediction accuracies on the 18 tests. Figure 3 visualizes the classification accuracy change over time.

Table 4: "2005 House to 1989-2006 Senate" prediction accuracies (in percent)

Year	Rep:Dem	Majority	Svm-bool	Svm-ntf	Svm-tfidf	NB-bool	NB-tf
1989	45:55 (100)	55.0	56.0	50.0	59.0	54.0	60.0
1990	45:55 (100)	55.0	55.0	48.0	56.0	53.0	62.0
1991	43:56 (99)	56.6	61.6	56.6	57.6	56.6	64.7
1992	43:56 (99)	56.6	59.6	48.5	63.6	56.6	68.7
1993	43:57 (100)	57.0	47.0	41.0	44.0	56.0	43.0
1994	43:56 (99)	56.6	39.4	43.4	43.4	54.6	41.4
1995	53:45 (98)	54.1	70.4	50.0	56.1	48.0	64.3
1996	53:46 (99)	53.5	63.6	56.6	70.7	49.5	79.8
1997	55:44 (99)	55.6	73.7	54.6	64.7	46.5	69.7
1998	55:45 (100)	55.0	64.0	52.0	62.0	50.0	63.0
1999	54:45 (99)	54.6	68.7	50.5	61.6	48.5	69.7
2000	54:46 (100)	54.0	72.0	50.0	68.0	49.0	73.0
2001	50:50 (100)	50.0	71.0	53.0	61.0	51.0	74.0
2002	50:50 (100)	50.0	61.0	56.0	63.0	56.0	67.0
2003	49:47 (96)	51.0	81.3	58.3	80.2	51.0	83.0
2004	51:48 (99)	51.5	81.8	62.6	82.8	52.5	82.8
2005	55:45 (100)	55.0	88.0	63.0	81.0	50.0	83.0
2006	55:45 (100)	55.0	87.0	64.0	84.0	58.0	83.0

The accuracy curves in Figure 4 show that the five classifiers form two groups based on their performance. Two classifiers, "svm-ntf" and "nb-bool," are very close to the majority baseline. The other three classifiers, "svm-bool," "svm-tfidf" and "nb-tf," perform similarly to each other. They all exhibit a trend of gradually increasing prediction accuracies from around 60% in 1989 to over 80% in 2006. However, the increase is not steady. There are two "valleys" in the curves, one in 1993-1994 (the 103rd Congress) and the other in the year 2002. There is also an unusual peak in 1995-1997. We notice that the 103rd Congress was the only Congress in our data set in which the Democrats controlled both the House and the presidency. It was also the

last Congress before the Republican take-over. Overall, the three classifiers predict the Senate data of recent years (2003-2006) better than older data.

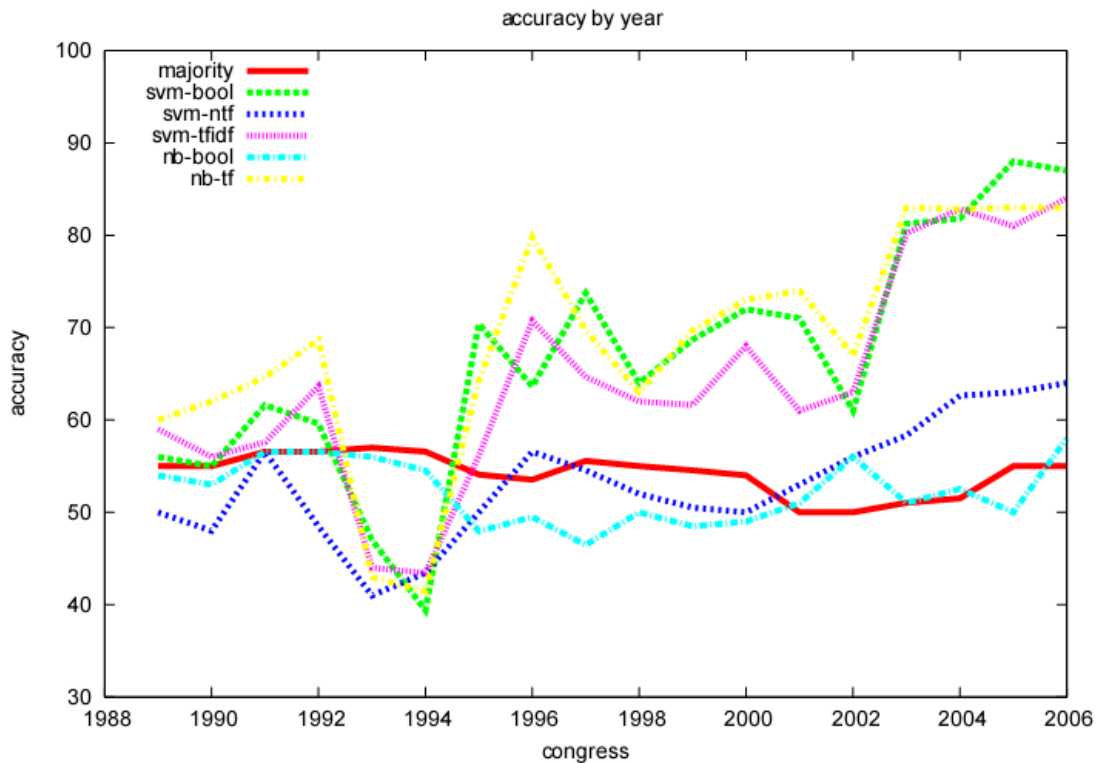


Figure 3: "2005House to 1989-2006 Senate" prediction accuracies

What causes the ideology classifiers' time-dependency? There are two possible explanations. One is that each Congress paid different levels of attention to various issues. For instance, in a specific year the focus may be on the war in Iraq, while in another years it may be on accounting reform or on an appointment to the Supreme Court. Such attention shifts result in vocabulary distribution drift by time. By this reasoning, the time-dependency actually is a consequence of issue-dependency. Changes in the overall agenda can be slow moving, which would explain the gradually increasing differences to the 2005 baseline year. Many issues (e.g. gun control) are re-visited periodically, which would explain the fluctuations in the accuracy

curves. Currently, however, we have only one year of House data. Therefore we cannot yet offer strong evidence for this explanation. If we could repeat the experiment on the House data of different years and still observe the same pattern as shown in Table 4 and Figure 3, we could be more confident in the vocabulary drift explanation. An alternative and more direct approach may be to identify issue drift over time and then compare this to ideological positions.

Another possible explanation is that the ideological orientation of Congress has shifted over time. There may be two reasons for this drift. First, membership in Congress is not constant, and as more partisan members enter the chamber, its overall level of partisanship may slowly change over time. Second, speeches may have become more clearly partisan in recent years, even for incumbent Senators. By this reasoning, ideological orientations in older speeches may have been more moderate and therefore harder to separate. Since we have the Senatorial speeches from 1989 to 2006, we design the third experiment to train ideology classifiers on the Senatorial speeches by year, and then run leave-one-out cross validation to test these classifiers. Because of the low performances of “svm-ntf” and “nb-bool” in the previous two experiments, we do not use them in this experiment.

Table 5 and Figure 4 show the remaining three classifiers’ cross validation accuracies from 1989 to 2006. The “nb-tf” classifier outperforms the majority baseline and the other two SVM classifiers by a large margin. However, this classifier is likely to overfit the Senate data since it did not generalize well to the House data in the “2005 Senate to House” prediction test. The performances of the “svm-bool” and “svm-tfidf” classifiers are similar to each other. In some years prior to 1999 they do not even reach the majority, but they constantly outperform the majority baseline after 1999. Overall, the cross-validation accuracies of all three classifiers between 2003 and 2006 are better than those in previous years. In other words, based on these

classifiers’ performance, the ideologies in recent years are more separable than those in previous years. This result is also consistent with the conventional wisdom in political science that recent Congresses have been more partisan than earlier ones.

However, can we infer based on Figure 4 that the classifiers’ time-dependency is the consequence of changes in the sharpness of ideology contrasts rather than issue changes? If this is true, we should find the curves in Figures 3 and 4 following the same trends. For example, in Figure 3 the accuracies of all three classifiers (“svm-bool”, “svm-tfidf”, and “nb-tf”) are very low in the years 1993, 1994, and 2002. If the same “valleys” can be observed in Figure 4, it is evident that the ideology “classifiability” change over time is the main reason for the time dependence in the “House to Senate” predictions. Otherwise we cannot reject issue changes as a possible explanation.

Table 5: Ideology classification cross-validation accuracies in the 1989-2006 Senate (in percent)

Year	Rep:Dem	Majority	Svm-bool	Svm-tfidf	NB-tf
1989	45:55 (100)	55.0	55.0	57.0	71.0
1990	45:55 (100)	55.0	55.0	59.0	77.0
1991	43:56 (99)	56.6	56.6	64.7	73.7
1992	43:56 (99)	56.6	56.6	57.6	68.7
1993	43:57 (100)	57.0	57.0	60.0	72.0
1994	43:56 (99)	56.6	56.6	69.7	82.8
1995	53:45 (98)	54.1	77.6	57.1	80.6
1996	53:46 (99)	53.5	53.5	50.5	75.8
1997	55:44 (99)	55.6	55.6	56.6	77.8
1998	55:45 (100)	55.0	55.0	67.0	75.0
1999	54:45 (99)	54.6	61.6	68.7	77.8
2000	54:46 (100)	54.0	66.0	65.0	76.0
2001	50:50 (100)	50.0	64.0	70.0	70.0
2002	50:50 (100)	50.0	63.0	77.0	76.0
2003	49:47 (96)	51.0	71.9	75.0	85.4
2004	51:48 (99)	51.5	60.6	71.7	80.8
2005	55:45 (100)	55.0	77.0	75.0	87.0
2006	55:45 (100)	55.0	73.0	67.0	83.0

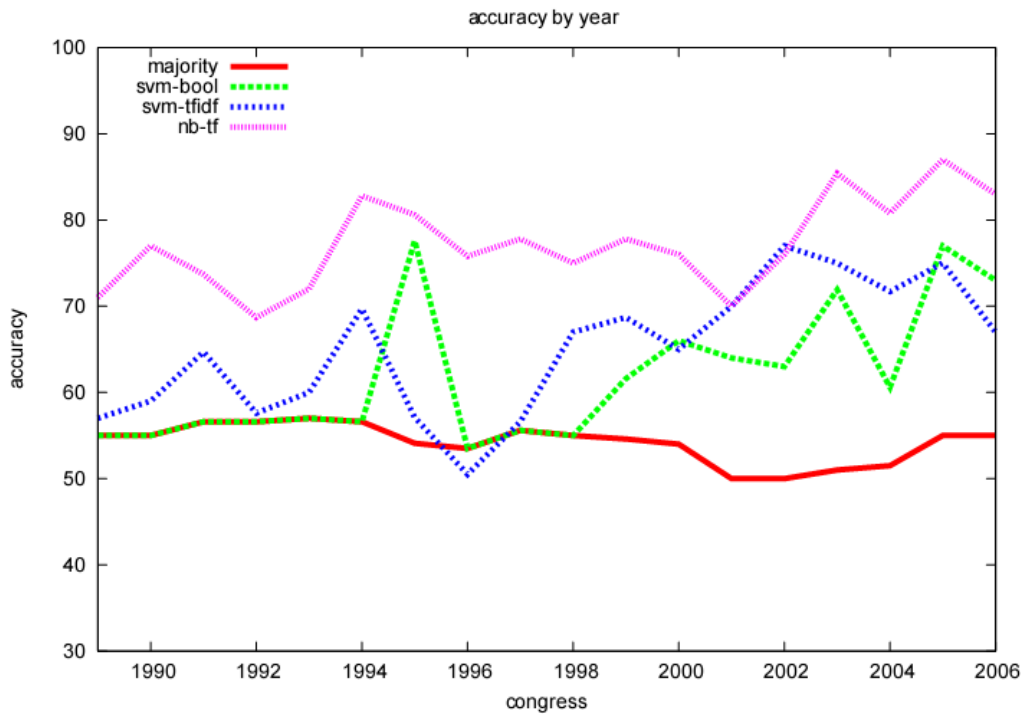


Figure 4: Ideology classification cross-validation accuracies in the 1989-2006 Senate

To compare the curves in Figures 3 and 4 in more details, we pair up each classifier’s accuracy curves in Figure 3 (2005House to Senate prediction by year) and Figure 4 (Senate leave-one-out cross validation by year) and plot them in the new Figures 5, 6, and 7, respectively. In Figure 5 (“svm-bool”) the two curves exhibit the same increase/decrease patterns after the year 1994. However, such patterns are not found in Figures 6 and 7. Therefore we conjecture that both issue changes and changes in the sharpness of ideology contrasts are possible causes of the ideology classifiers’ time-dependency.

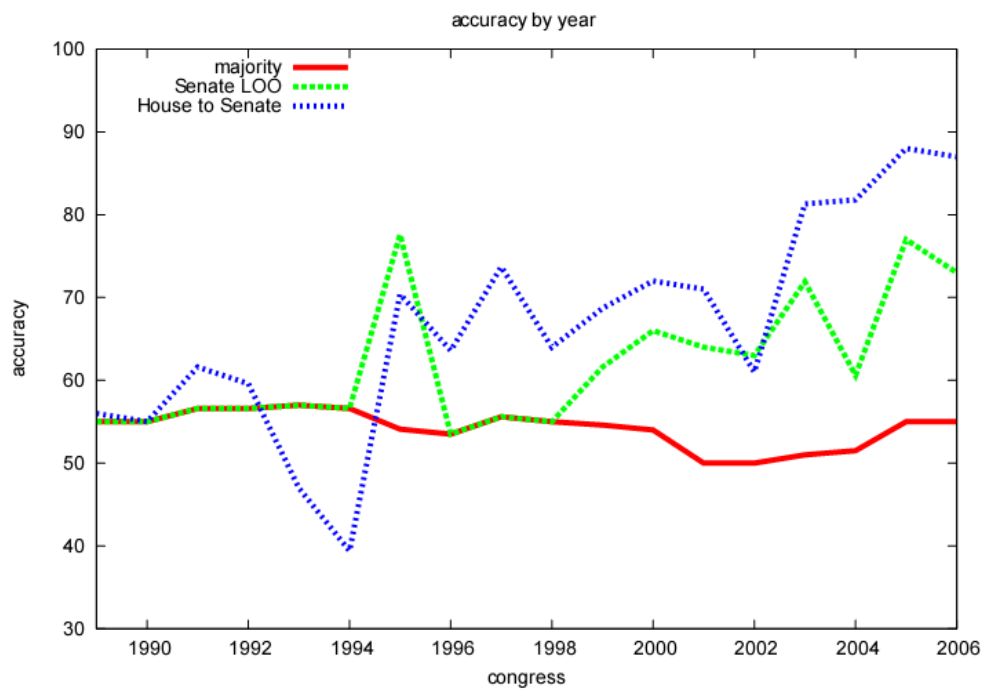


Figure 5: Classification accuracies of the "svm-bool" classifiers

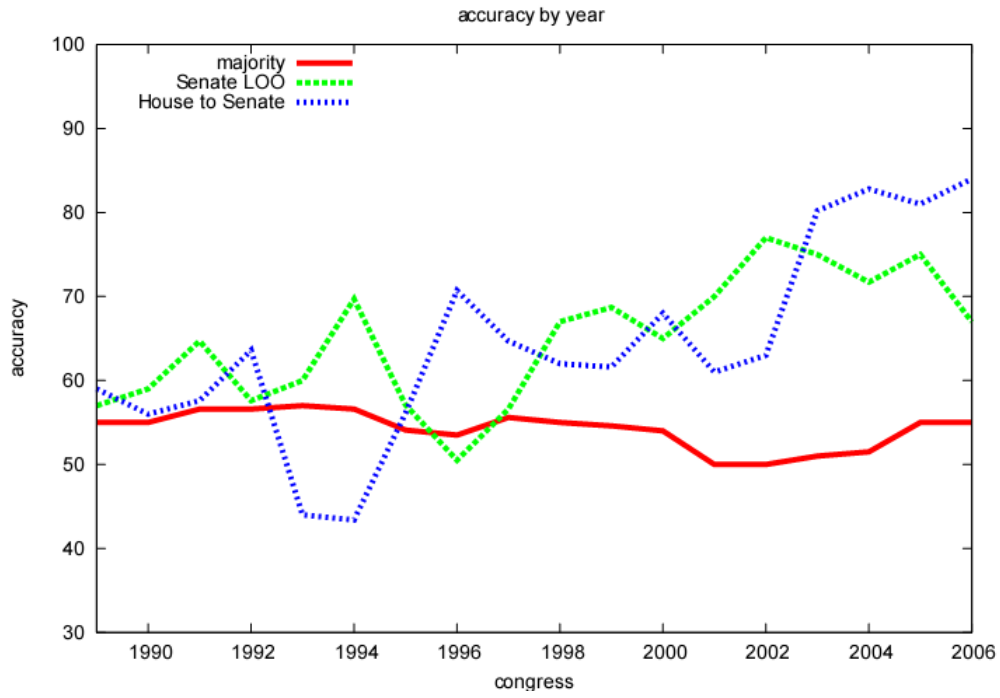


Figure 6: Classification accuracies of the "svm-tfidf" classifiers

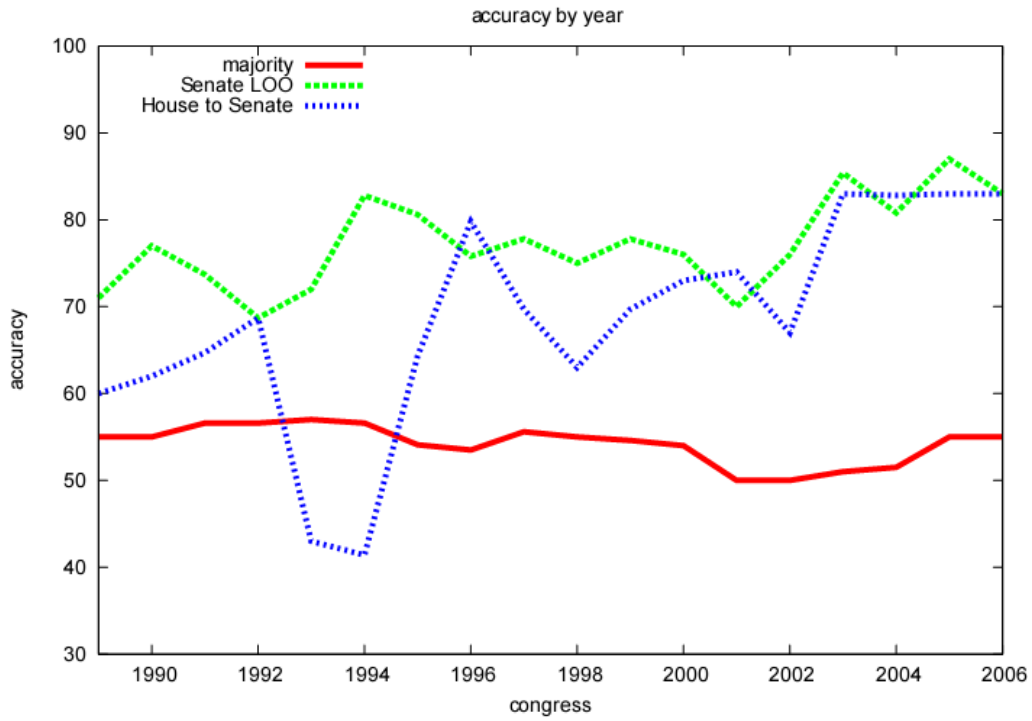


Figure 7: Classification accuracies of the "nb-tf" classifiers

Some general lessons: Data assumption violations and generalizability evaluation

In political text classification studies, it is quite common that both computer scientists and social scientists work together. Computer scientists usually focus on the classification methods and set forth certain assumptions for algorithm research purposes. For example, the class definition should be clear, the class labels should be correct, and, most importantly, the data should be independently and identically distributed and drawn from a fixed distribution. A classifier's performance and generalizability is in question if these assumptions are violated.

However, it is very likely that these assumptions would be violated in real-world applications (Hand 2004). In political text classification, such violations may occur for many reasons. The first problem is the subjectivity of the class definitions. Sometimes even human readers cannot agree with each other on the correct labeling of a given example. The second problem is that class labeling is error-prone. The errors could come from manual annotation mistakes, for example a customer might have written a very positive review while accidentally checked one star rating instead. The third problem is drift in the distribution. The distribution which generates the data might not be fixed. For example, the issue agenda in Congress may change over time. The fourth problem is that data might not be independently and identically distributed. In a debate an individual might adjust what he or she wants to say according to what the previous speakers have said. So the probability of generating one speech could be dependent on the previous speeches. The fifth problem is sample bias. We often pick convenient data sets. Sometimes they are small, so multiple distributions might all fit well. A classifier chooses the best fit according to its own statistical criterion, but the distribution which fits the training data best might not be the one of our interest. For example, if we want to find linguistic patterns

which separate those senators who support the Partial Birth Ban Act from those who oppose it, any pattern that recognizes female speakers would be helpful in prediction if (and since) most female senators oppose it. Actually a male/female classifier might work modestly well on this particular sample set, but it is not the intended opinion classifier.

In the collaboration between computer scientists and political scientists, usually the computer scientists are not deeply familiar with the data characteristics, while the political scientists are not deeply familiar with the classification methods. This gap in mutual understanding makes it difficult to foresee the assumption violations at the beginning of the experiment design. Consequently, the interpretation of the classifiers' generalizability becomes problematic. The sample bias might signify some patterns which fit this particular sample set but are not generalizable to the entire data set of interest. Therefore high classification accuracy might be due to coincidence. On the other hand, low classification accuracy may be attributable to vague class definition, erroneous class labels, or distribution drift.

Generalizability evaluation is especially important for complicated classification models such as ideology classifiers. From the supervised learning perspective, complicated models are more prone to overfitting. The number of Support Vectors (SVs) in a SVM model can be used as a measure of the model's complexity (Luping, 2006). In all our SVM experiments, the numbers of SVs are nearly the same as the numbers of training examples. Simple SVM models with low ratios of SVs to training examples are expected to be more generalizable than the ones with higher ratios. The models generated in our experiments are often on the higher end.

In our previous ideology classification study (Diermeier, Godbout, Yu, and Kaufmann 2007), the speakers in the test set (the 108th Senate) and the training set (the 101st-107th Senates) overlap to a great extent. This experiment design violates the independent and identical

distribution assumption for training and test data. Extra evaluation as reported in this paper is needed to examine the classifiers' generalizability to other sample data sets.

However, it is not easy to identify the potential person, time and issue dependencies which affect the classifiers' generalizability. We did not realize the potential person dependency problem until we found a large number of personal and state names among the top discriminative word features as weighted by the classification algorithms. We then found the time-dependency problem during our effort to evaluate the classifiers' person-dependency (the two dependencies can not be tested separately in the Senate data). Compared to the "black-box" type of classification accuracy evaluation, the weighted feature analysis is a "white-box" type of approach to interpret linear text classifiers. It provides us the opportunity to find "expected" as well as "unexpected" discriminative features. The unexpected features are likely to be the indicators of hidden coincidences which affect a classifier's generalizability. The interpretation of classification models is a research problem in machine learning in its own right (Luping, 2006). Choosing interpretable text classification methods such as linear classifiers is helpful in generalizability evaluation.

Conclusion

In this paper we reported a series of experiments to test the person-dependency and time-dependency of ideology classifiers trained on various Congressional speech subsets. Our experiment results demonstrate that cross-person ideology classifiers can be trained on Congressional speeches. The ideology classifiers trained on the 2005 House speeches are more generalizable than the ones trained on the Senatorial speeches of the same year, consistent with

our expectation that the House is more partisan than the Senate. We also found that the ideology classifiers trained on both House and Senate data are time-dependent. The time-dependency might be caused by changes in issues or vocabulary over time. Another possible explanation is that partisanship in the Senate has increased over time. The increasing classification accuracies in the Senate during the period of 1989 to 2006 support this explanation. This finding is consistent with what has been discovered from voting patterns. Overall, while the use of text classification methods is very promising in political science applications, existing approaches from computer science need to be carefully applied to the new domain.

References:

- Agrawal, R., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. *Proceedings of the 12th international conference on World Wide Web (WWW2003)*, 529-535
- Converse, P. E. (1964). The nature of belief systems in mass publics." In *Ideology and Discontent*, edited by D.E. Apter. New York: Free Press.
- Craig, H. (1999). Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, 14(1):103–113.
- Diermeier, D., Godbout, J-F, Yu, B., & Kaufmann, S. (2007). Language and ideology in Congress. MPSA 2007, Chicago
- Dave, K., Lawrence, S., & Pennock, D.M. (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web (WWW2003)*, 519-528
- Domingos, P. & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103–130
- Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM'98)*, 148-155
- Esuli, A. (2006). A bibliography on sentiment classification.
<http://iinwww.ira.uka.de/bibliography/Misc/Sentiment.html> (last visited: 10/31/2007)
- Evans, M., Wayne M., Cates, C. L., & Lin, J. (2005). Recounting the court? Toward a text-centered computational approach to understanding they dynamics of the judicial system. MPSA 2005, Chicago
- Hand, D.J. (2004). Academic obsessions and classification realities: ignoring practicalities in supervised classification. In *Classification, Clustering and Data Mining Applications*. ed. D.Banks, L.House, F.R.McMorris, P.Arabie, and W.Gaul. Springer.209-232.
- Hu, M. & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD2004)*, 168-177
- Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. *Lecture Notes in Computer Science (ECML'98)*, Issue 1398, 137-142

- Kwon, N., Zhou, L., Hovy, E., & Shulman, S.W. (2006). Identifying and classifying subjective claims. *Proceedings of the 8th Annual International Digital Government Research Conference*, 76-81
- Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2), 311-337
- Luping, S. (2006). Learning interpretable models. Doctoral dissertation, University of Dortmund.
- McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In *AAAI 98 Workshop on Learning for Text Categorization*
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Monroe, B. L. & Maeda, K. (2004). Rhetorical ideal point estimation: mapping legislative speech." Society for Political Methodology, Stanford University, Palo Alto.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumps up?: Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP2002)*, 79-86
- Poole, K. T. and Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespino, M. H., & Radev, D. R. (2006). An automated method of topic-coding legislative speech over time with application to the 105th-108th U.S. Senate. *Unpublished Manuscript*
- Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1-47 □
- Shulman, S. W. (2005). E-Rulemaking: issues in current research and practice. *International Journal of Public Administration* 27(7-8), 621-641
- Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP2006)*, 327-335
- Yang, Y. & Liu, X. (1999). A re-evaluation of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 42-49
- Yu, B., Kaufmann, S., & Diermeier, D. (2008). Exploring the characteristics of opinion expressions for political opinion classification. To appear in the *Proceedings of the 9th Annual International Conference on Digital Government Research (dg.o 2008)*

Appendix A: The preparation of the 101st-108th Senatorial speech data

We describe briefly the preparation process of the 101st-108th Senatorial speech data. Details can be found in our previous work (Diermeier et al., 2007).

We downloaded all Senatorial speeches of the 101st-108th Congresses from the government website www.thomas.gov and converted the original html files to pure text by removing the html tags, headers, tables, lists, and unicode characters. We then segmented the speech files into individual speeches.

An individual speech is a senator's speech given in a continuous time period until he or she stops. However, the Congressional record we downloaded includes not only speeches, but also some non-speech content, such as the officers' actions and documents inserted into the printed record. The beginning of a speech is always "Mr/Ms/Mrs. XXX," but the end of a speech may be the beginning of another senator's speech or a piece of non-speech content. Therefore we created a set of heuristic rules to remove non-speech content before the speeches could be correctly segmented. We removed the content matching any of the following rules:

- a) paragraphs starting with "The PRESIDING OFFICER"
- b) paragraphs starting with "There being no objection, the (\w+\s+)+ ordered to be printed in the RECORD"
- c) paragraphs starting with "The ACTING PRESIDENT pro tempore"
- d) paragraphs in brackets ()
- e) paragraphs starting with "By(\s+\w+):" or "S. number"

We generated the heuristic rules based on an iterative process. At the beginning we manually examined a small amount of the speech data and obtained rules a) and b). We then used these rules to automatically segment the speeches. Subsequently, we examined the longest speech segments, which usually contained non-speech content, and generated more rules to deal with

them. We carried out a few iterations until there were no more suspiciously long speech segments.

Once the speech segmentation was complete, we aggregated all speeches from an individual senator in each Congress into one long document. We then used a simple tokenizer to split the speeches into individual words. The tokenizer recognizes consecutive strings of alphabetical characters as valid words.

Finally we generated the vocabulary and document vectors for classification. The original vocabulary consisted of all word types which occurred in the Senatorial speech data set. To reduce the vocabulary size, we arbitrarily set a minimum term frequency of 50 and document frequency of 10 for a word to be eligible. We assumed the words with frequencies below this requirement are not representative. We also removed the top 50 most frequent words as stopwords (for example “the”, “a”, “of”, etc.). Stopwords are considered useless for classification because they occur frequently in every document. We also removed Senators’ names and state names to prevent the classifiers from picking up the potential correlations between the names and party affiliations. We generated four document vectors in the vocabulary space for each senator in each Congress, each word representing a dimension:

- a) Boolean: the value of each dimension is either word presence or absence (“1” or “0”)
- b) Tf: the value of each dimension is the word frequency in the document
- c) Ntf: the value of each dimension is the word frequency normalized by the document length
- d) Tfidf: the value of each dimension is the word frequency normalized by the inverted document frequency, i.e. the word frequency divided by the document frequency (the number of documents which contain this word in the whole collection)

Appendix B: the top word features (content words only) in the party affiliation classifiers trained on the 2005 House speeches

The following three tables list the most discriminative word features automatically learned by the “svm-bool”, “svm-tfidf”, “nb-bool” party classifiers which were trained on the 2005 House speech data. Each method assigned different weights to the words, but every method grasped core differences between the two parties. For example, the Republicans focus on economy, abortion, tax, and terrorism, etc., while the Democrats focus on social welfare, healthcare, children, and their own minority position in the Congress. Details of the feature analysis method can be found in (Diermeier et al., 2007).

Table 6: Top features of the "svm-bool" classifier

Republican	Democrat
economy	cuts
Commend	republican
reforms	opposition
Bringing	care
thank	new
Understanding	cut
jobs	budget
Gentleman	majority
Worked	programs
assets	iraq
area	debt
hard	middle
times	health
Chairman	substitute
Embryo	children
Urge	oppose
Areas	values
Passage	community
Growing	fails
Dollars	administration
Committee	diabetes
Stop	women

Republican	Democrat
Certainly	benefit
Government	proposed
Terrorists	failed
Growth	medical
Terror	child
Issue	question
Small	bush
Tough	republicans

Table 7: Top features of the "svm-tfidf" classifier

Republican	Democrat
Economy	republican
Embryos	cuts
Embryo	estate
Businesses	iraq
Meth	substitute
Small	majority
Death	republicans
Jobs	debt
Growth	billion
Identification	cut
Spending	CBC
Pension	budget
chinese	health
Human	values
Fence	administration
Commend	social
Proud	coal
Driver	coverage
Gentleman	research
Earmarks	CAFTA
Lawyers	courts
Business	education
Abortion	fails
Embassy	opposition
Gang	instead
Nations	maine
Terrorists	garza
Taxes	gun
Freedom	care
Tough	governor

Table 8: Top features of the "nb-tf" classifier

Republican	Democrat
Meth	CBC
Boutique	richest
Earmarks	garza
Uterus	vela
Democracies	disparities
Contracting	privatize
CNOOC	NCLB
Residential	brownsville
jessica	surpluses
Paragraph	crane
Transport	fleeing
Magnet	fails
wilson	extinction
Mohammed	slash
ATTA	estates
bartlett	ship
Keller	giveaways
Embryo	enron
Physiology	halliburton
Executed	paygo
Prolife	recourse
Liquid	objections
springfield	sample
Blends	pesticides
Continuity	unscrupulous
Blarding	greed
Genertically	refuses
Culmination	ILO
Murderers	trillions
Apple	slashing