Ideology Classifiers for Political Speech

Bei Yu
Stefan Kaufmann
Daniel Diermeier

Abstract:

In this paper we discuss the design of ideology classifiers for Congressional speech data. We then examine the ideology classifiers' person-dependency and time-dependency. We found that ideology classifiers trained on 2005 House speeches can be generalized to the Senate speeches of the same year, but not vice versa. The ideology classifiers trained on 2005 House speeches predict recent year Senate speeches better than older speeches, 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 by 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 political texts. A common goal, especially among computer scientists, has been the construction of general-purpose political opinion classifiers because of their potential applications in e-government development and mass media analysis (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), 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 deliberative congressional speech for comparison. Below are a few opening sentences from sample movie reviews[1].

> "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 similar expressive language terms can be found in the following comment on Partial Birth Ban Act[2]. Nevertheless, an educated reader can easily infer that this speaker is opposing the bill. The message conveyed is on of annoyance and "waster of time" while more important issues do not get tackled.

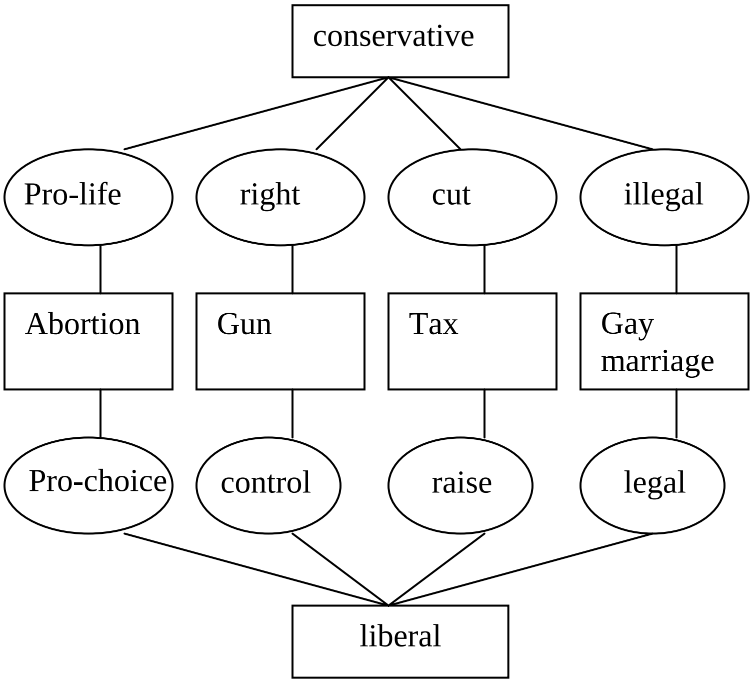
> "Mrs. MURRAY. Madam President, here we are, once again debating this issue. Since we began debating how to criminalize women's health choices yesterday, the Dow Jones has dropped 170 points; we are 1 day closer to a war in Iraq; we have done nothing to stimulate the economy or create any new jobs or 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)

Another important property of political speech is the importance of political ideology. In political setting, opinions on a given issue can be expected to depend on the person's underlying ideology rather than common standards as may be more typical of commercial speech (see figure 1). In other words, ideology will shape each individual's views on given issues and these influences will be identifiably different for Liberals and Conservatives.

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



For our purposes, the importance of political ideology suggests a different research orientation. Rather than classifying isolated opinions this approach would focus on classifying the underlying ideology of the person who holds the opinion. What makes this approach promising is the fact that ideologies give coherence to a person's opinions and attitudes which means 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 an individual.

> “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. context 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 or 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

[3] Understanding why certain ideologies resonate is an interesting research question in itself. For some recent approach from cognitive linguists see Lakoff (2002).

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.

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 set 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 been publicly accessible through the Thomas database[4], which provides ideal data for ideology analysis in speech. The goal is to use text classification as an analytical tool to probe whether the abstract concept of ideology constrains political speech as well.

The use of text classification as an analytical tool is not unique to the political science domain. Humanist scholars have been working on it for many years, most importantly in the context of identifying literary style. Craig (1999) once explained the connection between authorship attribution and stylistic analysis as two sides of a coin - you must have learned something about the authors' stylistic differences if you can tell them apart. Similarly if we observe high accuracy in the ideology classification result, we are confident that the classifier has learned some patterns to infer what texts look more like conservative or liberal. We could then extract and interpret these patterns and see if they make sense in the political science context. Currently the text data explored in related studies are mostly formal discourse, such as the Senatorial speech (XXX 2007), the Supreme Court briefs (Evans et al. 2005), and the party manifestos (Laver, Benoit, and Gary 2003). These studies all observe high classification

[4] The url for the database is http://thomas.loc/gov/ (last accessed 10/30/2007).

accuracy on their data sets, which indicate the existence of an ideological orientation at least in various formal political discourses.

As an example, in our previous study (XXX, 2007) we used the signs of Senators' D-nominate scores to label ideology categories (liberal or conservative) of Senatorial speeches from the 101st-108th Congresses. 25 most conservative and 25 most liberal Senators in each of the 101st-107th Congresses were selected as the training examples. Similarly, 50 "extreme" Senators in the 108th Congress were selected as the test examples. We used an SVM algorithm to train an ideology classifier and observed high classification accuracy on both the training set (through 5-fold cross validation) and the test set. The purpose of using the 108th Senatorial speech as the test set is to examine whether the 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 Senators correctly, our approach also allowed us to explore why this persistence across different Congresses occurs and whether it indeed reflects a coherent belief system. 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 most separating adjectives for Democrats we find the word *gay*, for the Republicans we find the word *homosexual*.

While these results are encouraging, we need to verify whether they truly 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, ideologies need to be fairly stable across issues and over time. Empirically, this means that an estimated 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. In other words, 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 will make it more likely to predict the position of another conservative Senator rather than a Liberal one.

A limitation of our existing results is that it was difficult to evaluate these characteristics within the Senatorial speech data alone because it was impossible to control all three sources of variation (person, issue, and time) in the same data set. For example, most of the 108th Senators were also Senators in previous Senates. While our estimates do a good job on the new Senators (4 out 5 are correctly classified) that sample is too small to draw reliable inferences. On the other hand, removing the speeches given by the 108th Senators in previous Congresses from the training data resulted in the lack of recent year speeches in the training data. Hence the person and time factors can not 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 over time and (if the Quinn et al. results are correct) over issues it does not constitute a direct test.

In this paper we try to control the person and time factors respectively by using the speeches in both House and Senate. Obtaining the 2005 House speech data from Thomas et al. (2006), we firstly test ideology classifiers' generalizability across House representatives and Senators of the same year (2005). We run a cross evaluation which consists of two tests. In the first test, we train ideology classifiers on speeches of 2005 House representatives and then use the classifiers to predict speeches in the 2005 Senate. In the second test we switch the training data and the test data, and then redo the classification. 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 another group.

We test the cross-time generalizability of our approach by using different-year speeches in the House and the Senate for training and testing. For example, we train ideology classifiers on 2005 House data and test these classifiers on the Senate data in 2004 and the years before and after. Stable prediction accuracies over time will provide evidence that the ideology classifiers can be generalized to speech data at different periods, otherwise the classifiers are time-dependent.

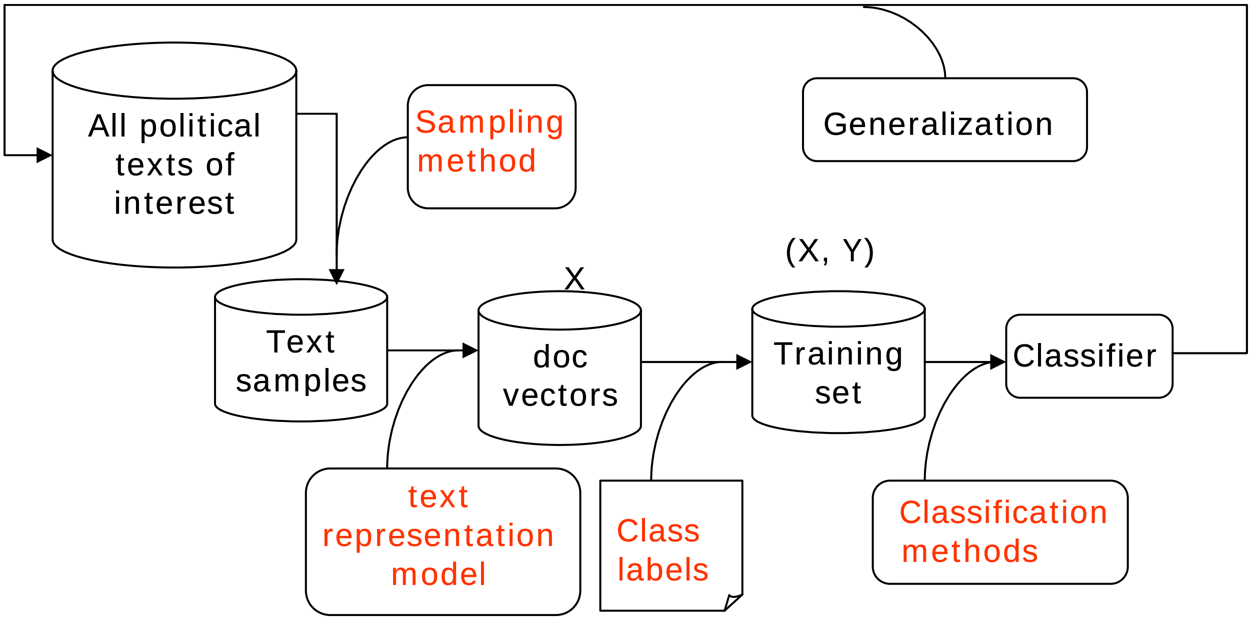
The paper is outlined as follows. We firstly introduce the text classification process, the text classification methods and evaluation measures used in this study. Then we report a series of generalizability evaluation experiments and results. Before concluding we discuss the difficulty in evaluating classifier generalizability and its relationship to data assumption violations in text classification experiment design.

The text classification process

As in the case of 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 can choose the 108th Senatorial speeches as a sample set of the whole Congressional speech collection. Then each text document in the sample set is converted into a numerical document vector, which is usually a vector of counts of linguistic patterns such as words and phrases. Then we have to obtain the correct labels for the sample data. Some labels are objective, such as a person's party affiliation. Some labels are subjective, such as the opinions of speeches as interpreted by coders. Sometimes human coders might not agree with each other whether a document is positive, negative or neutral. For these cases, inter-coder reliability test should be taken before applying automatic classification methods.

After attaching the labels to the corresponding examples, we can designate a classification method (e.g. SVM and naïve Bayes) to train a classifier on the labeled examples. 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 from.

Figure 2: Text classification process



Ideology classification experiment design

Figure 2 also shows that there are many choices to make in the design of text classification experiment, such as the sampling method, the text representation model, the label acquisition, the classification methods, and the evaluation measure. Without any prior knowledge regarding the particular classification problem, we start with the simplest text representation, the Bag-of-Words (BOW) approach, which converts each document into a vector of word occurrences 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.

For classification applications, some classes are easy to separate for most algorithms. But in many cases the data sets have some characteristics which favor some methods over the others. Therefore it is common to try multiple algorithms on a new data set. In our case we choose Support Vector Machines (SVM) 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 SVM is one of 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 use the SVM-light package[5] and its default parameter settings as the implementation of SVM algorithm in this study. SVM allows for the use of various kinds of word frequency measures as feature values, which results in multiple variations. We combine SVM with three different kinds of feature values. The first one is “svm-bool”, which uses word presence or absence in a document example as feature value. The second one is “svm-ntf”, which uses normalized word (term) frequency as feature value. The third one is “svm-tfidf”, which uses term frequency weighted by inverse document frequency as feature value.

We implement two variations of naïve Bayes algorithms according to (Mitchell 1997). The first one uses word presence and absence as feature value (“nb-bool”). The second one uses word frequency as feature value (“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 one training data set, each method will generate a different classifier. We evaluate the five ideology classifiers’ person-dependencies and time-dependencies in parallel.

Table 1: variations of SVM and naive Bayes classification methods

	Feature values			
Algorithms	word presence/absence	term frequency	normalized term frequency	idf-weighted term frequency
SVM	svm-bool	n/a	svm-ntf	svm-tfidf
naive 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 splits a data set into N folds and runs classification experiment N times. Each time one fold of data is used as test set and the classifier is trained on

[5] This software can be downloaded from http://svmlight.joachims.org/.

the other N-1 folds of data. The classification accuracy is averaged over the results of N runs. Hold-out test divides a data set into a training subset and a test subset. A classifier is trained on the training subset and tested on the test subset. Leave-one-out test is a special case of N-fold cross validation, when N equals the number of examples in the whole data set. For data sets with a small number of examples, an arbitrary train/test split would 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 use both leave-one-out cross validation and hold-out test in our study.

Evaluation of ideology classifiers' time and person dependencies

In the introduction section we have briefly discussed the ideology classification results in our previous study, in which we demonstrated that SVM-based ideology classifiers trained on the 101st-107th Senatorial speeches can effectively predict the ideologies of the 108th speeches as measured by D-NOMINATE scores. In this section we use a series of experiment to evaluate the ideology classifiers' person-dependency and time-dependency.

Our first experiment is intended to test whether our infer ideology classifiers exhibit too much person-dependency, i.e. that they are essentially person classifiers. Recall that in the Congressional context the notion of ideology presupposes as shared belief system. Our approach is to design an experiment that (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 and use one chamber as the training, 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 do set their own agenda- due to the fact that both chambers have to agree on each proposed bill to become law we can expect substantial overlap between the two agenda. Rather than using D-NOMINATE derived categories we use party affiliation to label the legislators' ideology classes. This is necessitated by the fact that D-NOMINATE score cannot necessarily be compared across chambers. However, as we showed in XXX (2007) for Senate D-NOMINATE and party based classifications are highly correlated.

We use the 2005 Congressional speeches in the House[6] and the Senate, here labeled as two data sets "2005House" and "2005Senate". In addition to within-chamber validation tests we also run a cross evaluation which consists of two tests: 1) train classifiers on the "2005House" data and test them on the "2005Senate" data; and 2) train classifiers on the "2005Senate" data and test them on the "2005House" data. By this design we make sure the training and test examples are two groups of people without overlap, yet that the issues under discussion are highly similar because the speeches happened in the same Congress in the same year.

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 on 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 XXX (2007). In that analysis we found that using ideological extreme

[6] 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.

Senators allowed us to classify moderate Senators well, but not vice versa. We interpreted this as evidence that the ideology of extremist Senators is more well defined compared to the more blurry views held by 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 finding (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 achieve over 80% prediction accuracies, which demonstrate that they are not likely person-dependent. The “nb-bool” classifier performs worse than the majority baseline. The svm-ntf classifier is better than the majority baseline[7] 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
svm-bool	75.1	88
svm-ntf	69.8	63
svm-tfidf	80.1	81
nb-bool	77.9	50
nb-tf	78.7	83

[7] “Majority baseline” is a trivial classification method which assigns all test examples to the category where the majority of the training examples belong to. For example, if a data set have 55 positive examples and 45 negative examples, the majority baseline is 55%.

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”. This time svm-ntf is still the worst among the five classifiers. It’s 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. “Svm-tfidf” and “nb-tf” classifiers are better than the majority baseline, but their performances are much lower than their counterparts in the last “2005 House to Senate” experiment.

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

	2005 Senate cross validation	2005 House prediction
Majority baseline	55	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 result is better than the “2005 Senate to House” prediction result. 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 fact that these naïve Bayes classifiers do not predict the “2005House” data well can be explained as that the classifiers trained on “2005Senate” are simply overfitting the 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 within the period of 1989-2006. Each year's Senatorial speeches consist of one test set. There are 18 test sets in total, each has about 100 examples (Senators). We run the test 18 times, once for each year. Table 4 shows the classifiers' prediction accuracies in the 18 tests. Figure 3 visualizes the classification accuracy change over time.

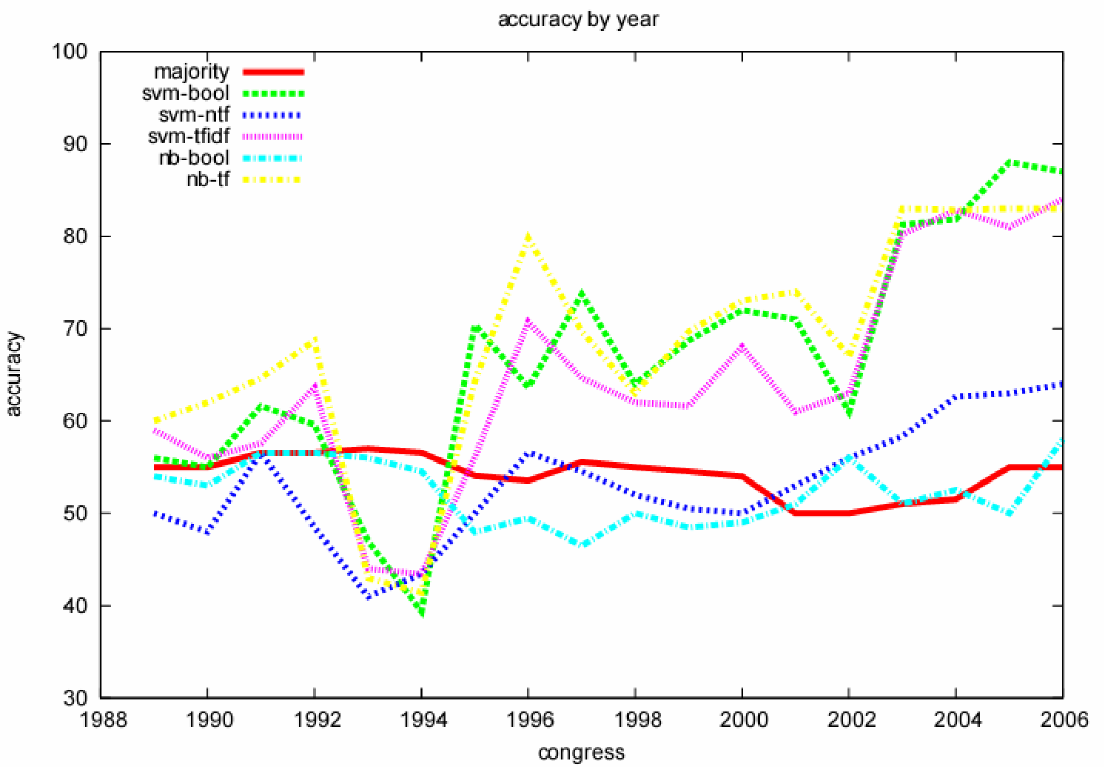
Table 4: "2005 House to 1989-2006 Senate" prediction accuracies (in percent)

Year	Republicans vs. Democrats	Majority	Svm-bool	Svm-ntf	Svm-tfidf	NB-bool	NB-tf
1989	45:55 (100)	55	56	50	59	54	60
1990	45:55 (100)	55	55	48	56	53	62
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	47	41	44	56	43
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	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	64	52	62	50	63
1999	54:45 (99)	54.6	68.7	50.5	61.6	48.5	69.7
2000	54:46 (100)	54	72	50	68	49	73
2001	50:50 (100)	50	71	53	61	51	74
2002	50:50 (100)	50	61	56	63	56	67
2003	49:47 (96)	51.0	81.3	58.3	80.2	51.0	83

2004	51:48 (99)	51.5	81.8	62.6	82.8	52.5	82.8
2005	55:45 (100)	55	88	63	81	50	83
2006	55:45 (100)	55	87	64	84	58	83

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. 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 (by year)



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. In other words, over a specific year, the focus may be on the war in Iraq. In another, it may be on accounting reform, or on an appointment to the Supreme Court. Such attention shifts result in vocabulary distribution drift by time. By this reasoning, the time-dependency actually is a consequence of the 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 House data. So we still can not provide 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 would be more confident in the vocabulary drift explanation. A more direct approach may also try to directly 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 vague 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.

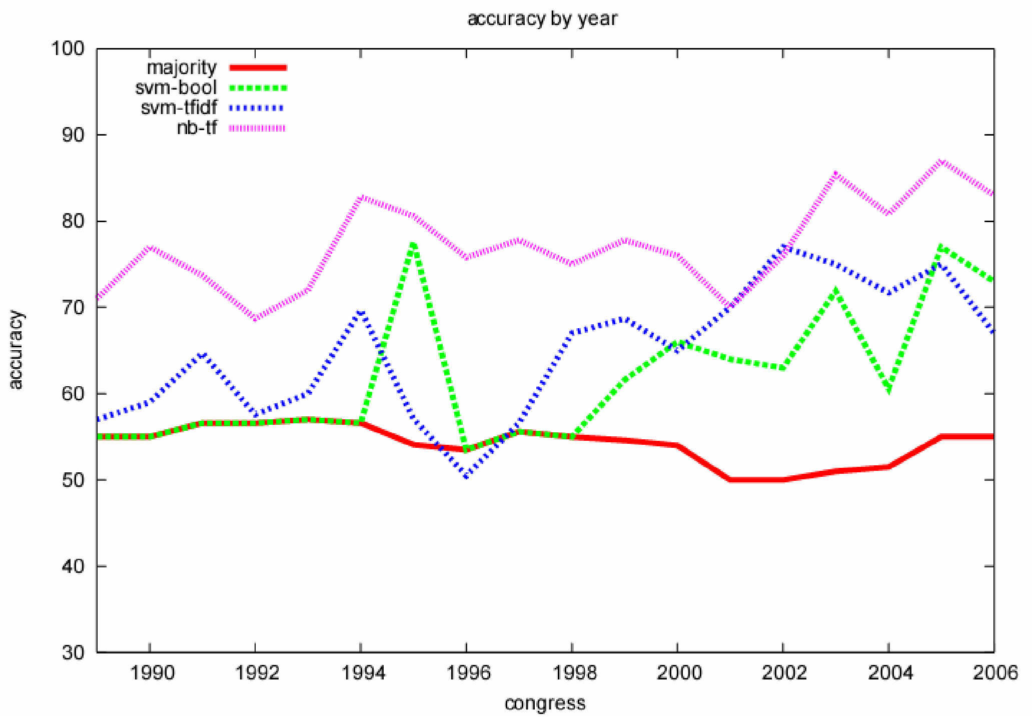
Table 5 and Figure 4 show the remaining three classifiers’ cross validation accuracies from 1989 to 2006. The “nb-tf” classifier outperforms the majority baseline and the other two SVM classifiers by a large margin. However, this classifier is likely to overfit the Senate data in that it cannot be well generalized 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. Sometimes they can not even beat the majority baseline before the year 1999, but they constantly outperform the majority baseline since 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’ criteria, the ideologies in recent years are more separable than those in older time. This result is also consistent with the common knowledge in political science that recent Senates are more partisan than in previous years.

However, can we infer based on Figure 4 that the classifiers’ time-dependency is the consequence of the changes in the sharpness of the ideology concept rather than the 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 classification 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 dependency in the “House to Senate” predictions. Otherwise we can not reject issue changes as a possible explanation.

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

Year	Republicans vs. Democrats	Majority	Svm-bool	Svm-tfidf	NB-tf
1989	45:55 (100)	55	55	57	71
1990	45:55 (100)	55	55	59	77
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	57	60	72
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	55	67	75
1999	54:45 (99)	54.6	61.6	68.7	77.8
2000	54:46 (100)	54	66	65	76
2001	50:50 (100)	50	64	70	70
2002	50:50 (100)	50	63	77	76
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	77	75	87
2006	55:45 (100)	55	73	67	83

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 corresponding 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 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 the ideology concept sharpness changes are possible causes of the ideology classifiers' time-dependency.

Figure 5: classification accuracies of “svm-bool” classifiers

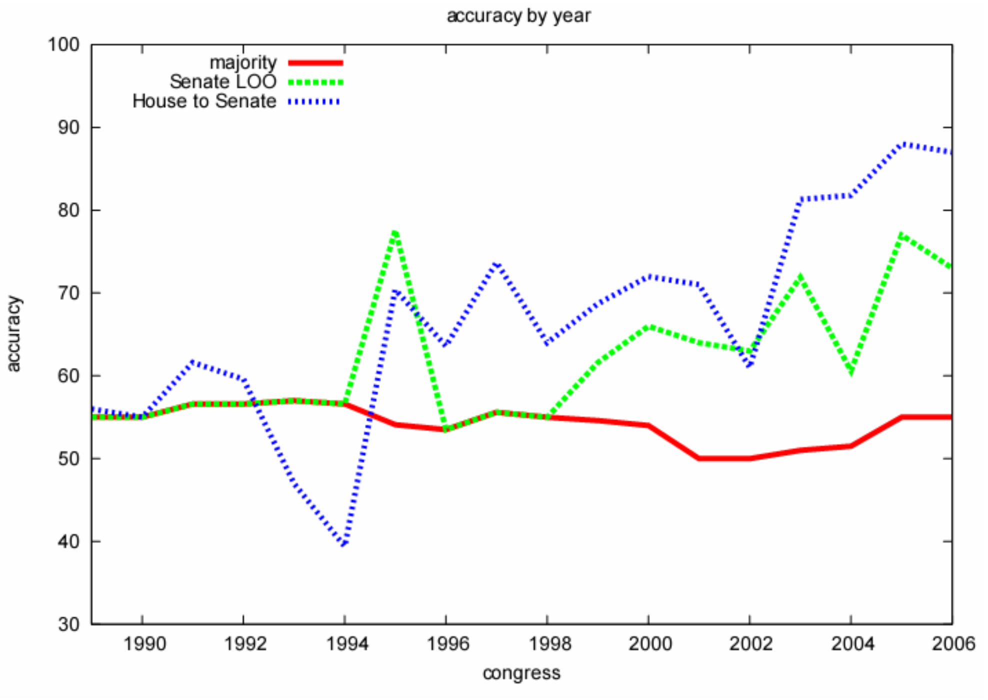


Figure 6: classification accuracies of “svm-tfidf” classifiers

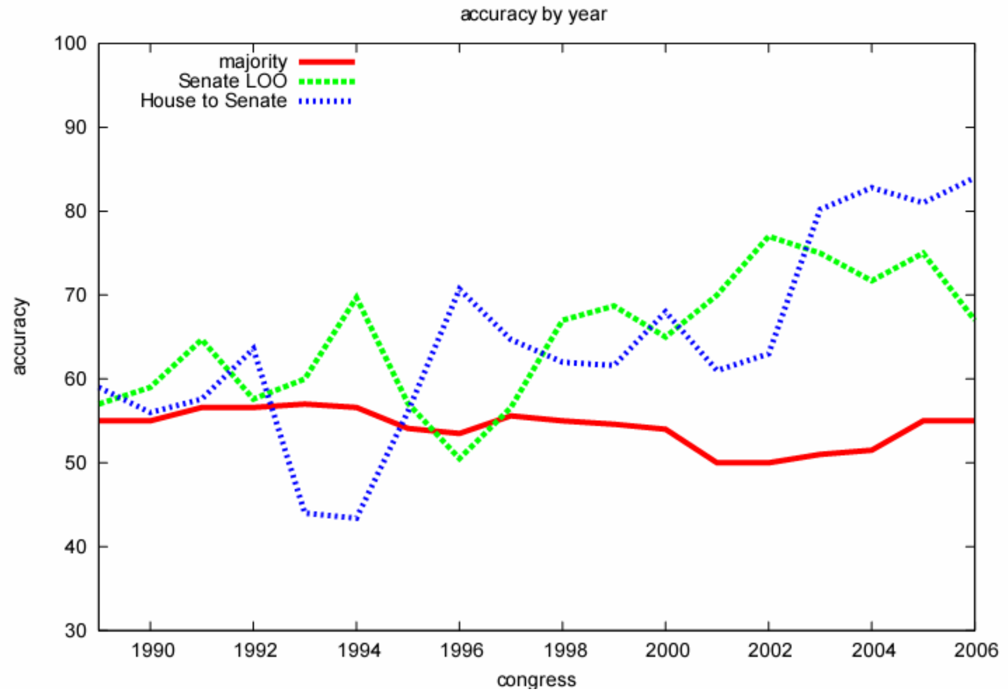
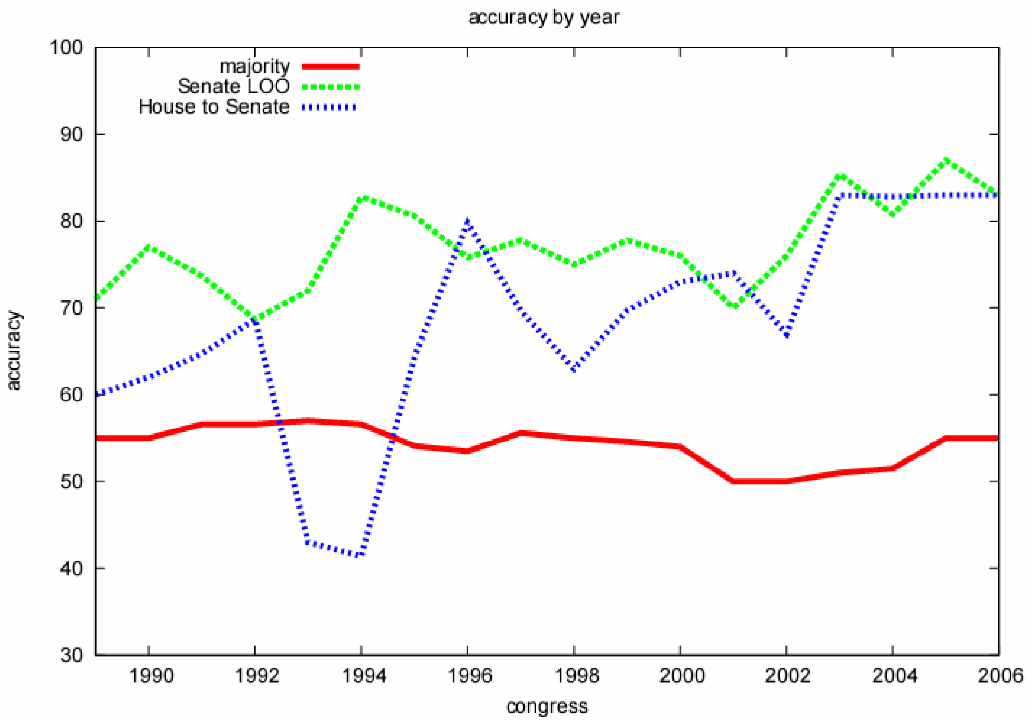


Figure 7: classification accuracies of “nb-tf” classifiers



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 in the exploration. Computer scientists usually focus on the classification methods. They set up some assumptions for algorithm research purpose. For example, the class definition should be clear, the class labels should be correct, and the most important one is the assumption of independently and identically distributed data from a fixed distribution. A classifier’s performance and generalizability is in question if the assumptions are violated.

However, it is very likely that these assumptions would be violated in real applications (Hand 2004). In the setting of political text classification, many reasons could result in the assumption violation. The first problem is the subjective class definitions. Sometimes even human readers cannot agree with each other which is the correct label for an example. The second is the erroneous class labels. The errors could come from manual annotation mistakes, or convenient labels which are not equivalent to the real labels. The third problem is the drifting distribution. The distribution to generate 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 probability of generating the previous speeches. The fifth problem is the sample bias. We often pick a convenient data set. 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, we want to find linguistic patterns to separate the senators who support or oppose the Partial Birth Ban Act. But because most female senators oppose it, any pattern that recognizes female speakers is helpful in prediction. Actually a male/female classifier might work modestly well on this particular sample set, but it is not the real opinion classifier we expected.

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 experiment design. In many cases the trained classifiers are never tested in another independent sample set because the purpose of classification is to use the accuracy as a confidence measure of the “classifiability” of the given data set. This makes the examination of assumption violation even harder. 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 driven by some coincidences. On the other hand, low classification accuracy may be attributed to vague class definition, erroneous class labels or distribution drift.

The generalizability evaluation is especially important for complicated classification models such as the 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 always nearly 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. But the models generated in our experiments are always on the higher end.

In our initial ideology classification (XXX 2007), the speakers in the test set (the 108th Senate) and the training set (the 101st-107th Senates) overlap to 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 large number of person and state names among the top discriminative word features 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 the linear classifiers are helpful for generalizability evaluation.

Conclusion

In this paper we use 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 classifier can be trained on the 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. We also found that the ideology classifiers trained on both House and Senate data are time-dependent. The time-dependency might be caused by the issue and vocabulary changes over time. Another possible explanation is the fact that the Senates are more partisan than before. 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 the 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://liinwww.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., Crespin, 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□

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