Note: This is a slightly corrected version of the paper by the same title published in the *Proceedings of the 9th Annual International Digital Government Conference (dg.0 2008)*, pages 82-91. Aside from a couple of typographical corrections, the substantive changes concern Tables 17 and 19, each of which was printed in the proceedings with two rows switched.

Exploring the Characteristics of Opinion Expressions for Political Opinion Classification

Bei Yu
Kellogg School of Management
Northwestern University
1-847-491-8791
bei-yu@northwestern.edu

Stefan Kaufmann
Department of Linguistics
Northwestern University
1-847-491-5779
kaufmann@northwestern.edu

Daniel Diermeier
Kellogg School of Management
Northwestern University
1-847-491-5177
d-diermeier@northwestern.edu

ABSTRACT
Recently there has been increasing interest in constructing general-purpose political opinion classifiers for applications in e-Rulemaking. This problem is generally modeled as a sentiment classification task in a new domain. However, the classification accuracy is not as good as that in other domains such as customer reviews. In this paper, we report the results of a series of experiments designed to explore the characteristics of political opinion expression which might affect the sentiment classification performance. We found that the average sentiment level of Congressional debate is higher than that of neutral news articles, but lower than that of movie reviews. Also unlike the adjective-centered sentiment expression in movie reviews, the choice of topics, as reflected in nouns, serves as an important mode of political opinion expression. Manual annotation results demonstrate that a significant number of political opinions are expressed in neutral tones. These characteristics suggest that recognizing the sentiment is not enough for political opinion classification. Instead, what seems to be needed is a more fine-grained model of individuals' ideological positions and the different ways in which those positions manifest themselves in political discourse.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic processing*

General Terms
Algorithms, Measurement, Design, Experimentation.

Keywords
Machine learning, feature selection, sentiment classification, text categorization, e-Rulemaking.

1. INTRODUCTION
Recently there has been increasing interest in constructing general-purpose political opinion classifiers because of their potential applications in e-Rulemaking and public opinion analysis [1][17][29]. The goal of political opinion classification is to correctly sort political texts depending on whether they support or oppose a given political issue or proposal under discussion. Previous studies have tried to categorize opinions in various genres of political text, such as Congressional debates [29], online newsgroup discussions [1], and email feedback on government policy by the general public [17].

Previous work has in general assumed that this task is closely related to sentiment classification, which has been studied for more than ten years [9]. Sentiment classifiers have achieved good classification accuracies (>80%) in predicting the positive or negative orientation of texts with strong expressive content, such as movie and customer reviews.

There are two main approaches to sentiment classification. The knowledge based approach predefines an affect dictionary and then searches the input documents for occurrences of words in that dictionary [30]. An affect dictionary is a list of words with positive or negative sentiment orientations. Initially, these affect dictionaries were manually constructed for cognitive linguistics research purposes. They were then borrowed to analyze emotions in other text domains. The General Inquirer [28], the Dictionary of Affect in Language [32] and LIWC [24] are a few commonly used affect dictionaries. More recently, automatic methods have been developed to construct or expand affect dictionaries [12][31][33][26]. The second approach to sentiment classification is supervised learning, which trains a statistical classifier on a set of pre-labeled examples and then uses the classifier to predict the sentiment orientations of new texts [6][14][21]. In the TREC-2006 blog track opinion retrieval task, half of the teams used the knowledge based approach and half used the learning based approach [20]. Both approaches recognize the importance of affective vocabulary to sentiment classification. However, in the dictionary based approach it is given *ex ante* while in the classification approach it can be inferred from texts through feature selection.

In the political context, this line of research is trying to apply the same methodology to political opinion classification. However, the published results are not as good as those achieved in review classification. The politics newsgroup debate classifiers in [1] were not better than a simple majority vote. The EPA public email classifiers in [17] only slightly outperformed the baseline. The Congressional debate classification accuracy in [29] was around 70%, modestly above the majority baseline (58%), but still not comparable to the sentiment classifiers in the customer review domain.

To understand the performance gap between the political domain and the review domain, we designed a series of experiments to explore some characteristics of political opinion expression which might affect the sentiment classification performance.

Specifically, we chose the floor debates in the United States Congress as the focus of our study. Congressional speech has long

been considered a paradigmatic example of political deliberation. The speeches are usually well prepared, and they are accurately recorded and preserved in the Thomas database (http://thomas.gov).

In general in sentiment classification, the frequency of affective words indicates the sentiment intensity level of a document. Given an affect dictionary, we define the sentiment level of a piece of text as the proportion of affect words in the total number of words. Intuitively, political debates and customer reviews belong to two different text domains/genres, and different domains/genres are characterized by unique linguistic properties [2]. Thus the question arises whether sentiment level is a characteristic that is stable and unique for different opinion domains/genres. The purpose of our first experiment is to measure the sentiment level of Congressional debates and compare it against reference points in two other domains/genres. One is business news, characterized by a high prevalence of neutral tones. The other is movie reviews, in which individuals usually express their opinions in strong emotional terms. Legislators are expected to express their opinion clearly to the Congress and the constituents. At the same time, they are bound by implicit norms and conventions of legislative debate that, in most cases, would limit highly emotional language in favor of giving reasons for a particular political position. An important question therefore is whether the Congressional debate is more similar to business news or movie reviews in terms of sentiment level.

Previous studies also indicated that different parts-of-speech contribute differently toward sentiment classification. Adjectives are considered the most informative sentiment indicators [12][30]. The purpose of our second experiment is to investigate whether this is also true for political opinion classification. We used statistical classification and feature selection methods to select the most discriminant words for classifying political opinions and movie reviews, respectively, and then compared the distributions of nouns, verbs, adjectives and adverbs in the top feature lists.

Most sentiment classification studies deal with "positive vs. negative" binary classification [23][16]. For example, [22] demonstrated that removing neutral sentences does not affect the accuracy of sentiment classification of movie reviews. In contrast, [17] found that some political opinions (support or opposition) in email messages from the public to the EPA were fully expressed in neutral tones, but that three-class sentiment classification seemed much more difficult than binary classification. The authors were also concerned with the impact of poorly written texts on the classification result. The Congressional debates provide us with good-quality data for an investigation of the consistency between political opinions and sentiment tones. The purpose of our third experiment therefore is to explore the distribution of positive, negative and neutral political opinions expressed in speeches in the Senate and the House, and its correspondence to the distribution of positive, negative and neutral tones used in those speeches. Objective political opinion labels (derived from voting records) and subjective labels (obtained by manual annotations) were both investigated in this study. The sentiment tones of the debate speeches were manually annotated by three annotators.

This paper is organized as follows. Section 2 discusses the experiment preparation. Section 3.1 describes the first experiment on sentiment level measurement; Section 3.2 the second experiment on the parts-of-speech distribution of informative features; and Sections 3.3 and 3.4 the third experiment on consistency between political opinions and sentiment tones. Section 4 concludes with discussions.

2. EXPERIMENT SETUP

2.1 Data Preparation

For the political domain, we downloaded all Senatorial speeches during the period of 1989-2006 from the government website Thomas.gov. To compare the two chambers, we also used the 2005 House speeches used in [29]. For the business news domain, we collected 130,000 news articles on Wal-Mart in the year 2006, including contents from both traditional media and various internet news sources. For the customer review domain, we used the 2000 movie reviews used in [21].

2.2 Computational Software

LIWC (Linguistic Inquiry and Word Count) is a text analysis program designed by psychologists to gauge the linguistic expression of emotions in texts of a wide range of genres. It has been widely used in psychology and linguistics [24][25]. LIWC counts the occurrences of 2300 words or word stems in 70 categories, including overall affect (sentiment) and positive and negative feelings. The proportions of words in these categories indicate the emotion levels along the corresponding dimensions.

SVMs (Support Vector Machines) are among the best methods for text classification [7][15][37] and feature selection [10][11]. SVMs select discriminative features with broad document coverage, and therefore reduce the risk of over-fitting and increase the feature reduction rate [37][38]. Studies have shown that SVM classifiers can reach >90% feature reduction rate, and the 10% most discriminative features work as well as, or even better than the entire original feature set [11][38]. External feature selection does not improve SVM classification accuracy [10]. In this study we used the SVM-light package [15] with default parameter settings.

3. EXPERIMENTS AND RESULTS

3.1 Sentiment Level Comparison

In this experiment we used LIWC to measure the sentiment levels in sample texts from three domains: Congressional debates, movie reviews, and business news articles.

For the Congressional debate domain, we measured the sentiment levels of Senatorial speeches in 18 years (1989-2006). The speeches in each year were concatenated into one large text file. Table 1 lists the overall sentiment, positive emotion and negative emotion levels. We also measured the 2005 House floor debates used in [29] for comparison between the two chambers. As Table 2 shows, the sentiment levels of the 2005 Senate debates and the 2005 House debates are very similar to each other.

In the movie review domain, we measured the sentiment levels of 2000 movie reviews (1000 positive and 1000 negative) used in [21]. Since this data set does not contain temporal information, we split the reviews randomly into ten subsets, each containing 100 positive reviews and 100 negative ones. Table 3 lists the sentiment levels of all 10 subsets.

For the business news domain, we measured the sentiment levels of 130,000 news articles on Wal-Mart published in 2006. The

articles were grouped into 12 subsets by month, and we measured the sentiment level for each month. Table 4 shows the results.

Table 1: Sentiment level of Senate speeches (%)

Year	Sentiment	Positive	Negative
1989	3.41	2.36	1.00
1990	3.33	2.29	0.99
1991	3.51	2.40	1.06
1992	3.38	2.29	1.05
1993	3.33	2.23	1.05
1994	3.32	2.27	1.02
1995	3.28	2.21	1.04
1996	3.27	2.21	1.02
1997	3.33	2.35	0.95
1998	3.31	2.31	0.96
1999	3.46	2.37	1.05
2000	3.39	2.35	1.00
2001	3.49	2.41	1.04
2002	3.63	2.48	1.12
2003	3.54	2.42	1.08
2004	3.71	2.49	1.19
2005	3.61	2.43	1.13
2006	3.47	2.36	1.07
Avg.	*3.43*	*2.35*	*1.05*
SD	*0.13*	*0.09*	*0.06*

Table 2: Sentiment level of 2005 Senate and House speeches (%)

Chamber	Sentiment	Positive	Negative
Senate	3.61	2.43	1.13
House	3.71	2.48	1.20

Table 3: Sentiment level of movie reviews (%)

Subset	Sentiment	Positive	Negative
1	5.08	3.14	1.94
2	4.73	2.85	1.87
3	5.11	3.11	2.00
4	5.04	2.99	2.04
5	5.04	3.08	1.96
6	4.98	3.03	1.93
7	4.87	3.03	1.83
8	5.06	3.10	1.96
9	5.01	3.11	1.89
10	4.89	2.95	1.94
Avg.	*4.98*	*3.04*	*1.94*
SD	*0.12*	*0.09*	*0.06*

Table 4: Sentiment level of business news (%)

Month	Sentiment	Positive	Negative
1	2.85	1.94	0.89
2	2.91	2.00	0.89
3	2.81	1.89	0.89
4	2.69	1.86	0.81
5	2.86	2.01	0.82
6	2.72	1.90	0.79
7	2.77	1.84	0.90
8	2.78	1.84	0.91
9	2.81	1.89	0.90
10	2.79	1.91	0.85
11	2.70	1.94	0.74
12	2.67	1.91	0.73
Avg.	*2.78*	*1.91*	*0.84*
SD	*0.07*	*0.06*	*0.07*

Table 5: Sentiment levels in different domains (%)

Data set	Sentiment	Positive	Negative
News articles	2.78	1.91	0.84
Senate debate	3.43	2.35	1.05
Movie reviews	4.98	3.04	1.94

The summary in Table 5 shows that the average sentiment level of Congressional debates (3.43%) is higher than that of news articles (2.78%) but lower than that of movie reviews (4.98%). The difference is shown graphically in Figure 1. The small standard deviations in all three domains indicate that the sentiment levels in these domains are stable across different samples. We used SPSS to test the significance of two sentiment level differences: 1) Congressional debate vs. business news; and 2) Congressional debate vs. movie reviews. None of the Shapiro-Wilk tests in the three domains rejected the null hypotheses of normal distribution. Neither of the Levene's tests rejected the null hypotheses of variance equality in the two comparisons. Without violation of the normal distribution and variance equality assumptions, both independent samples test results were significant ($p<0.001$). In other words, the Congressional debate domain is significantly different from both the business news and the movie review domains in terms of sentiment level.

The intermediate sentiment level of Senatorial speeches poses a challenge for knowledge based approaches to sentiment classification. Statistical classifiers, in contrast, presumably learn sentiment indicators automatically and therefore should perform well in this situation. Our next experiment examines whether this is true.

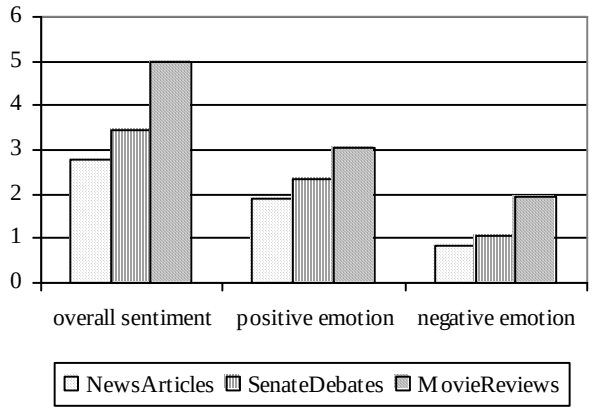

Figure 1: Sentiment levels in different domains

3.2 Parts-of-speech Distribution of Informative Word Features

In this experiment, we compared the contributions of different parts-of-speech (POS) to sentiment classification in the Congressional debate and movie review domains. We first used the Brill tagger [3] to select content words (nouns, verbs, adjectives, and adverbs). The presence or absence of these content words in the documents was then used in training an SVM for classification and feature selection. We compared the proportions of the word groups in the top-ranked 100 features in both domains.

Table 6 shows the classification results before and after feature selection in the movie review domain. Only one movie review did not contain any of the 100 selected features and therefore was removed from the data set. For better comparison with previous work, we used the same 3-fold cross validation evaluation method as in [21]. The classification accuracy is 85.7% with 12426 content word features and 85.5% with the top-ranked 100 features. Thus the small feature set works as well as the entire content word feature set.

Table 7 lists the top-ranked 100 features selected by the SVM classifier for movie reviews. Adjectives form the largest group, and almost all of them have obvious sentiment orientations. The other three word groups include both affective words and neutral words, for example affective nouns like "flaws" and "fun", verbs like "fails" and "save", and adverbs like "wonderfully" and "badly." This result confirmed previous studies which had shown that adjectives are important indicators for sentiment classification in the review domain.

Table 6: SVM classification of moview reviews

	Feature selection	
	Before	After
Feature set size	12426	100
Empty docs	0	1
3-fold CV acc	85.7%	85.5%

Table 7: POS distribution of movie review features

Nouns (28)	Verbs (16)	Adjectives (35)	Adverbs (21)
nothing	have	***bad***	***unfortunately***
mess	work	***worst***	only
script	supposed	***boring***	maybe
others	show	***awful***	***perfectly***
plot	tries	***hilarious***	also
performances	falls	***ridiculous***	especially
flaws	've	***memorable***	sometimes
life	makes	***terrific***	***wonderfully***
reason	fails	***poor***	then
fun	looks	***excellent***	***definitely***
none	wasted	***lame***	most-
waste	deserves	overall	very
people	seen	***stupid***	***well***
works	allows	***great***	anyway
today	save	flat	yet
town	played	***terrible***	***poorly***
history		***dull***	***badly***
tv		***better***	extremely
attempt		***best***	quite
class		***enjoyable***	truly
point		***perfect***	easily
material		least	
bill		***worse***	
pace		***pointless***	
filmmakers		***true***	
little		different	
director		many	
laughs		potential	
everything		only	
		normal	
		entertaining	
		realistic	
		annoying	
		solid	

The result of our first experiment demonstrated that the sentiment levels are consistent across large enough sample texts in each domain. On the other hand, the sentiment levels of individual short documents may vary. For example, some reviews or news articles are more emotional than others. Similarly, the political debates on some bills are more heated than others. Here we chose one of the most controversial debates, the Senate debate on the Partial Birth Ban Act, for comparison with the movie reviews. The debate occurred on March 11 and 12, 2003. Over 30 senators joined the debate. An individual speech is defined as a speech given in a continuous time period until the speaker stops [29]. The debate includes 193 speeches, with 96 from Senators who voted "yea" and 97 from Senators who voted "nay" on the bill.

The SVM classification accuracy is 80.3% with 1252 content word features and 92.2% with the top-ranked 100 features (Table 8). Because of the small size of the data set, we used leave-one-out cross validation in this experiment. The top 100 features include 45 nouns, 27 verbs, 17 adjectives and 11 adverbs (Table 8). Unlike in movie review classification, most of the top 100 features in the partial birth ban debate are neutral words, especially nouns, with

the exception of a few emotional adjectives such as "brutal" and "horrible" used by the pro-ban side. In fact, the two sides also tended to use different, seemingly neutral nouns to refer to the same concept, thereby imbuing these neutral nouns with emotional connotations for this particular debate. For example, the pro-ban side presumably used words like "partial-birth" and "abortion" to express (and perhaps trigger) sadness and anger about loss of lives, while the anti-ban side used the medical term "D and X" to create a sense of emotional detachment.

The important role of emotionally neutral but topic-specific nouns in this case suggests a closer investigation of the role of topics in political opinion classification. Opinions are often considered orthogonal to topics, with opinions often expressed by adjectives and topics expressed by nouns. However, related content analysis results on the same debate illustrate the connection between topics and opinions. Schonhardt-Bailey [27] found four main topic clusters in the debate we used: 1) women's right to choose and morality of abortion; 2) constitutionality and legal standing; 3) personal experience; and 4) legislative procedure. The words in Table 9 demonstrate the resonance of these topics in the top-ranked features.

Table 8: SVM classification of the partial-birth abortion debate

	Feature selection	
	Before	After
Feature set size	1252	100
Empty docs	0	0
Leave-one-out acc	80.3%	92.2%

Table 9: POS distribution of partial-birth abortion debate features

Nouns (45)	Verbs (27)	Adjectives (17)	Adverbs (11)
Boxer	been	unconstitutional	much
women	say	partial-birth	obviously
health	done	medical	forward
abortion	have-VB	right	then
people	have-VBP	legislative	simply
legislation	has	pregnant	***unfortunately***
doctor	asking	***terrible***	necessarily
v	told	multiple	roughly
opportunity	means	***better***	also
family	think	***true***	really
life	address	***sure***	repeatedly
someone	pass	own	
reason	speak	such	
woman	try	***brutal***	
pregnancy	tell	***safest***	
families	performed	identical	
friend	facing	***horrible***	
society	let		
issue	work		
ban	does		
case	was		
time	taught		
exception	apply		

Nouns	Verbs	Adjectives	Adverbs
doctors	underlying		
approach	stop		
picture	went		
stake	goes		
lot			
way			
percent			
pregnancies			
anything			
abortionist			
level			
brain			
debate			
hospitals			
thing			
courts			
problems			
clerk			
today			
couple			
term			
colleague			

The above experiment examined the debate on one controversial bill. How does this generalize to the parts-of-speech distribution for general Congressional debates? The 2005 House debate speeches have been labeled by voting records of the speakers in [29]. We repeated the parts-of-speech distribution experiment on the entire 2005 House debate data[1]. The classification result (Table 10) shows that the accuracy with 100 top-ranked features is only slightly lower than that with all content word features. The top 100 features include 48 nouns, 27 verbs, 17 adjectives and 8 adverbs. The distribution is almost the same as that in the partial birth ban debate classification. Again, nouns constitute the largest word group, and few affective words were included in the top features for Congressional debate classification (Table 10).

Although the parts-of-speech distribution is similar for single bill debates and general Congressional debates, the general debate classification accuracy is much lower than that of the partial-birth abortion debate. This phenomenon can be explained by the characteristics of low sentiment level and topic/opinion connection in Congressional debates. SVM classifiers tend to choose discriminative words with broad document coverage, but affective words do not widely occur in Congressional debates because of the low sentiment level. Instead, many neutral words are important opinion indicators. However, these words are so specific to individual bills that they do not have broad enough coverage to be picked by the classifier as good indicators for general debate classification. For example, "unconstitutional" was a strong indicator in partial-birth abortion debate, but it was not involved in many bill debates. The SVM classifier successfully captured some obvious opinion indicators like "support" and "oppose," but these words will be missing if an opinion is not expressed directly. Our next experiment examines the directly and indirectly expressed political opinions through human annotation.

[1] This data set contains debates on 53 controversial bills. Bill debates with nearly unanimous votes were excluded.

Table 10: SVM classification of the 2005 House debates

	Content words	
	Before	After
Feature set size	5951	100
Empty docs (train/test)	0/0	24/2
3-fold CV acc on training set	66.8%	70.9%
Acc on test set	65.5%	64.9%

Table 11: POS distribution of House debate features

Nouns (48)	Verbs (27)	Adjectives (17)	Adverbs (8)
i	***oppose***	republican	n't
demand	***appreciate***	parliamentary	very
leader	is	present	not
vote	recorded	small	here
majority	support	important	instead
support	vote	corporate	forward
cuts	reclaiming	***safe***	finally
inquiry	give	***right***	already
nays	ask	many	
yeas	ask	open	
work	live	human	
opposition	committed	fair	
reserve	help	few	
health	working	environmental	
call	distinguished	ranking	
chairman	funded	general	
objection	according	***controversial***	
time	***understand***		
businesses	***cut***		
gentlewoman	resulting		
mr	seen		
communities	asking		
bill	look		
folks	***oppose***		
alternative	consume		
minutes	have		
regard	passed		
terri			
quorum			
look			
balance			
gentleman			
budget			
ms			
question			
reforms			
congress			
florida			
wisconsin			
fact			
percent			
interest			
people			
nothing			
provisions			
substitute			
debt			
fashion			

3.3 Consistency between Political Opinions and Voting Records

In the customer review domain, the customers also give ratings when they write reviews. These star ratings are transformed into class labels for sentiment classification experiments. High classification accuracy is evidence that the labels basically match the sentiment of the reviews. Based on the same idea, voting records, as used in [29] and our second experiment, may be used as objective and convenient class labels for Congressional debate classification. Although it is reasonable to assume that a legislator's speech should be consistent with the voting decision, a legislator is not required to defend his or her voting decision in every speech. For example, he or she might just explain the advantages and disadvantages of a certain bill without stating support or opposition toward the bill. Similarly, legislators may base their voting decisions on procedural or strategic considerations that may not be directly expressed in speech, because of their complexity or because a direct expression may be disadvantageous to the legislator. Therefore, it is possible that the voting records and the political opinions are not consistent in the speeches. Such inconsistency would negatively affect the classification result.

We firstly examined human readers' ability to recognize opinions in Congressional speeches. We hired three undergraduate students to read the speeches in the development set of the 2005 House debate corpus. The development set includes 5 bill debates. Some speeches are very short, for example "Mr. Speaker, I reserve the balance of my time." To minimize the number of such irrelevant speeches, we concatenated the speeches of each speaker in each debate, hoping each speaker would make his or her opinion clear at least once. We obtained 113 speeches after the concatenation. The three annotators were asked to annotate the opinion in each speech as "S" (support), "O" (oppose), "N" (neutral), or "I" (irrelevant).

Table 12: Inter-coder agreement in political opinion recognition

	Kappa agreement
A1 and A2	0.80
A1 and A3	0.61
A2 and A3	0.64

A1 vs. A2	S	O	N	I
S	**38**	1	5	2
O	0	**30**	0	0
N	2	2	**15**	2
I	0	2	0	**14**

A1 vs. A3	S	O	N	I
S	**38**	0	5	3
O	0	**27**	3	0
N	3	5	**9**	4
I	5	3	0	**8**

A2 vs. A3	S	O	N	I
S	**35**	0	3	2
O	1	**30**	4	0
N	5	3	**9**	3
I	5	2	1	**10**

The annotations from the three annotators, labeled A1, A2 and A3, are compared in Table 12. The three annotators agreed on 77 of the 113 annotated speeches. This represents an agreement of 68.14%. The annotators reached substantial kappa agreements (kappa>0.6) among each other. The annotators have almost unanimous agreement on the "support vs. opposition" decision (S/O, only two exceptions). However, each annotator classified over 1/3 of the speeches as either neutral or irrelevant (N/I), and most of the disagreements are about the "N/I vs. S/O" decisions.

We reviewed the 36 speeches with annotation disagreements and found the following five recurring reasons for disparate answers. Some disagreements have multiple reasons.

1) Mixed speeches containing both supporting and opposing statements (7 speeches)

Annotators cannot easily determine the opinion of such speech, but must make a judgment as to which statements contain more emphasis or are better representative of the speaker's opinion. For example:

"Mr. chairman, I want to thank the gentleman from North Carolina (Mr. Taylor), our distinguished chairman, for offering to work with me and the committee to resolve this through the conference process. I believe that this is an important and critical step toward addressing what has been a very real injustice." (source: [199_400077])

2) Implicit opinions in speech (9 speeches)

Some discrepancies between annotators occur if the speaker does not express an opinion explicitly. For example, the speaker may discuss other actors who support or oppose a bill, or he or she may discuss the implications of a bill. A reader may infer an opinion from speech where another reader sees no opinion. The level of inference may depend on the reader's background knowledge of the issue, bill, and even speaker.

3) Implicit bills in speech (13 speeches)

Speeches with expressed opinion do not always specify the target or object of the opinion. Annotators expect the opinion to be of the discussed bill, and may mark irrelevant or no clear opinion if the speaker discusses more than one bill, or uses multiple names or acronyms to discuss a bill, or expresses an opinion regarding something other than a bill. For example:

"I urge we defeat CAFTA as negotiated and return to the table, which we can do, and refinish this agreement in about a month." (source: [421_400238])

"I rise in support of that motion, and I thank him for his hard work and support on this issue." (source: [199_400420])

4) Opinions on part vs. whole bill (8 speeches)

Annotator disagreement also occurs when the speaker expresses an opinion on a part of the bill which may not correspond with their opinion on the whole bill. For example:

"Mr. chairman, I raise a point of order against section 413 of H.R. 2361, on the grounds that this provision changes existing law in violation of clause 2(b) of house rule XXI, and therefore is legislation included in a general appropriation bill." (source: [199_400098])

"Mr. chairman, I yield myself such time as I may consume. Mr. chairman, kill head start? Supporting religious discrimination which was added by the majority to this otherwise very good bill is exactly what would kill head start." (source: [493_400436])

5) Procedural speech (8 speeches)

Lastly, disagreement among annotators may be due to confusion over what constitutes procedural statements in Congress. Determining if a statement is procedural, and therefore irrelevant, requires background knowledge of the House. For example:

"Mr. speaker, on May 19, 2005, I was unable to be present for rollcall vote No. 190, on ordering the previous question to provide for consideration of H.R. 2361, making appropriations for the department of the interior, environment, and related agencies for the fiscal year ending September 39, 2006 and for other purposes. Had I been present I would have voted `` yea '' on rollcall vote No. 190." (source: [199_400293])

This discussion suggests that Congressional bills are considerably more intricate subjects than consumer products like digital cameras and movies. The multiple rounds of revisions and debates inherent in the legislative process further complicate the opinion analysis.

The opinion annotations resulting after the resolution of the annotation differences are listed along with the corresponding voting records in Table 13. Only in 68 (60.2%) speeches did the readers identify expressions of support or opposition. The significant number of "neutral" or "irrelevant" speeches suggests that voting records do not necessarily correspond to the political opinions expressed in speeches. The mismatch between voting records and real opinions adds difficulty to voting record based political opinion classification. The inter-coder agreement analysis also demonstrates that while manually annotating political opinion labels is promising, the background knowledge required may make annotator recruitment difficult.

Table 13: Voting records and opinion annotation in speeches

	Support	Oppose	Neutral	Irrelevant
Yea (56)	38	0	9	9
Nay (57)	0	30	21	6
Total (113)	38	30	23	16

3.4 Consistency between Political Opinions and Sentiment Tones

The practice of sentiment-analysis based political opinion classification is driven by the idea that individuals use positive

tones to express support and negative tones for opposition. However, a legislator might praise life to oppose abortion, or criticize opponents in support of a bill. The low sentiment level and the importance of neutral-word opinion indicators also point to the prevalence of neutral tones in political opinion expression. In our third experiment, we investigated the consistency between political opinions and sentiment tones in Congressional debate. Previous studies have shown that subjectivity can be recognized with high accuracy at the sentence level [34][35][36]. In this experiment we annotated the sentiment tones of Congressional speeches at the sentence level and compared those ratings with the overall opinions of the speeches. We used the debates on the Stem Cell Research Act in both House and Senate. In our earlier annotation task (see previous section), one of the three annotators had demonstrated the strongest political science background knowledge. She was asked to annotate the opinions and tones in each speech in the stem cell research debate as positive, negative, or neutral. Irrelevant speeches were excluded. The House and Senate debates include 126 and 63 speeches, respectively.

The results of this round of annotation suggest that the sentiment tones change frequently within single speeches in both House and Senate. A correlation test also shows that the number of tone changes is highly correlated with the speech length for both House and Senate. For example:

Mr. Speaker, I rise in strong support of H.R. 810. <POS> Our research policies should be decided by scientists and doctors at the National Institutes of Health and not by Karl Rove and self-appointed religious gurus. <NEG>

To investigate the relationship between opinions and tones, we computed the number of words covered by each tone and used the tone which covers the largest number of words as the fundamental tone of each speech. We then examined the relationship between opinions and the fundamental tones.

Table 14: Opinions and fundamental tones in the House

	Positive	Negative	Neutral	*Total*
Support	43	**6**	31	*80*
Oppose	**4**	14	19	*37*
Neutral	1	0	8	*9*
Total	*48*	*20*	*58*	*126*

Table 15: Opinions and fundamental tones in the Senate

	Positive	Negative	Neutral	*Total*
Support	21	**3**	19	*43*
Oppose	**5**	0	11	*16*
Neutral	1	0	3	*4*
Total	*27*	*3*	*33*	*63*

Table 16: Opinions and beginning tones in the House

	Positive	Negative	Neutral	*Total*
Support	63	**2**	15	*80*
Oppose	**13**	7	17	*37*
Neutral	0	1	8	*9*
Total	*76*	*10*	*40*	*126*

Table 17: Opinions and beginning tones in the Senate

	Positive	Negative	Neutral	*Total*
Support	19	**1**	23	*43*
Oppose	3	1	12	*16*
Neutral	0	0	4	*4*
Total	*22*	*2*	*39*	*63*

Table 18: Opinions and ending tones in the House

	Positive	Negative	Neutral	*Total*
Support	64	**1**	15	*80*
Oppose	**2**	20	15	*37*
Neutral	2	0	7	*9*
Total	*68*	*21*	*37*	*126*

Table 19: Opinions and ending tones in the Senate

	Positive	Negative	Neutral	*Total*
Support	7	**0**	36	*43*
Oppose	**0**	1	15	*16*
Not clear	1	0	3	*4*
Total	*8*	*1*	*54*	*63*

The results in Tables 14 and 15 suggest that even if one could perfectly recognize the fundamental tones, the overall opinion prediction accuracy would not exceed (43+14+8)/126=51.6% in the House debate and (21+0+3)/63=38.1% in the Senate. A major reason for the low accuracy is that nearly half of the support/opposition opinions are expressed in neutral tones. Negative tones are the least used in the debates. Overall support with fundamental negative tones and opposition with fundamental positive tones occur in about 10% of the speeches, suggesting that it is not a common approach for speakers to deliver their opinion in opposite tones.

To further investigate the relationship between political opinions and sentiment tones, we specifically looked at the tones in the beginning and ending sentences in speeches (Tables 16-19). For the House debate, perfect recognition of the beginning tones would result in (63+7+8)/126=61.9% opinion classification accuracy. The accuracy is (64+20+7)/126=72.2% if ending tones can be recognized perfectly. For the Senate debate, perfect recognition of the beginning tones would result in (19+0+12)/63=49.2% opinion classification accuracy. The accuracy decreases to (7+0+15)/126=34.9% if ending tones are used. Apparently neutral

tones are more popular in Senate debates. A possible reason is that the House debates are more constrained and the representatives have shorter speaking time. In consequence, they have to make their points sooner and more succinctly. Alternatively, the House may be more ideologically polarized than the Senate. Regardless of whether we look at House or Senate debate, however, the above level of accuracy is not better than the word-based classification results in [29]. This means that even if one could perfectly classify sentiment tones down to sentence level, one would still not be able to correctly classify political opinion.

4. CONCLUSIONS AND DISCUSSIONS

General-purpose political opinion classification tools are important for e-Rulemaking applications. We have investigated some characteristics of political opinion expression which might affect sentiment-analysis based political opinion classification. We found that the average sentiment level of Congressional debate is rather low. It is higher than that of neutral news articles, but much lower than that of movie reviews. Furthermore, affective adjectives are not the most informative political opinion indicators. Instead, the choice of topics, as reflected in neutral nouns, is an important mode of political opinion expression by itself. Our manual annotation results demonstrate that a significant number of political opinions are expressed in neutral tones. These characteristics suggest that political opinion classification is not equivalent to sentiment classification. Identifying the sentiment is not sufficient for general-purpose political opinion classification.

Although we have not found a satisfying general-purpose political opinion classification method, the characteristics of political opinions expression that we have discovered suggest the exploration of approaches alternative to sentiment-based classification. Recently, another type of political text classification task also attracted much attention. This task aims to classify speakers' ideology positions [8][18][19][5]. Converse [4] viewed ideologies as "belief systems" that constrain the opinions and attitudes of individuals. In other words, ideology will shape each individual's views on given issues and these influences will be identifiably different for Liberals and Conservatives. Once we have properly identified a person's ideology, we may be able to predict his or her opinions on various political issues. It is our goal for future work to explore viable approaches for ideology based political opinion classification.

5. ACKNOWLEDGMENTS

We wish to thank Marcella Wagner, Cheng Xun Chua, and Christopher Wickman for their careful annotations.

6. REFERENCES

[1] 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

[2] Biber, D. (1988). Variation across speech and writing. Cambridge University Press.

[3] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics, 21(4), 543–566

[4] 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.

[5] Diermeier, D., Godbout, J-F, Kaufmann, S., and Yu, B. (2007). Language and ideology in Congress. MPSA 2007, Chicago

[6] 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

[7] Dumais, S., Platt, J., Heckerman, D., and M. Sahami. (1998). Inductive learning algorithms and representations for text categorization. Proceedings of the 7th International Conference on Information and Knowledge Management, 148-155

[8] Durant, K. T. & Smith M. D. (2006). Mining sentiment classification from political web logs. Proceedings of workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (WebKDD2006)

[9] Esuli, A. (2006). A bibliography on sentiment classification. http://liinwww.ira.uka.de/bibliography/Misc/Sentiment.html (last visited: 10/31/2007)

[10] Forman, G. (2003). An extensive empirical study of feature selection metrics for text categorization. Journal of Machine Learning Research, 3:1289–1305

[11] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machiness. Machine Learning, 46(1-3), 389–422

[12] Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of the 35th ACL / 8th EACL conference, 174-181

[13] Hatzivassiloglou, V. & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of COLING

[14] 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

[15] 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

[16] Koppel, M. and Schler, J. (2006). The importance of neutral examples for learning sentiment. Computational Intelligence 22(2), 100-109

[17] 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

[18] 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

[19] Mullen, T. and Malouf R. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In

Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs (). 159–162. DOI=

[20] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. and Soboroff, I. (2007). Overview of the TREC-2006 blog track. Proceedings of the 15th Text REtrieval Conference. NIST

[21] 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

[22] Pang, B. and Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of ACL 2004

[23] Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of the 43rd Annual Meeting of Association for Computational Linguistics, 115–124, Ann Arbor, MI.

[24] Pennebaker, J. W. & Francis, M. E. (1999). Linguistic Inquiry and Word Count (LIWC). Mahwah, NJ: LEA Software and Alternative Media/Erlbaum

[25] Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. Annu. Rev. Psychol. 54, 547-77.

[26] Riloff, E. Wiebe, J., and Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. Proceedings of the 7th Conference on Natural Language Learning (CoNLL-03), 25-32

[27] Schonhardt-Bailey, C. (2008). The Congressional debate on partial-birth abortion: constitutional gravitas and moral passion. British Journal of Political Science. Forthcoming

[28] Stone, P. J. (1966). The General Inquirer: A computer approach to content analysis. MIT Press

[29] 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

[30] Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL'02), 417-424

[31] Turney, P. and Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4), 315-346

[32] Whissell, C. M. (1989). The dictionary of affect in language. Journal of Emotion: Theory, Research and Experience, vol 4, 113-131

[33] Wiebe, J. (2000). Learning subjective adjectives from corpora. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas, July 2000

[34] Wilson, T., Wiebe, J. and Hwa, R. (2006). Recognizing strong and weak opinion clauses. Computational Intelligence, 22(2), 73-99

[35] Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M. (2004). Learning subjective language. Computational Linguistics, 30(3):277–308.

[36] Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from un-annotated corpus. Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)

[37] 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

[38] Yu, B. (forthcoming). An evaluation of text classification methods for literary study. Journal of Literary and Linguistic Computing