CHAPTER 5

Text to ideology or text to party status?*

Graeme Hirst, Yaroslav Riabinin, Jory Graham,
Magali Boizot-Roche, and Colin Morris
Department of Computer Science, University of Toronto

Several recent papers have used support-vector machines with word features to classify political texts – in particular, legislative speech – by ideology. Our own work on this topic led us to hypothesize that such classifiers are sensitive not to expressions of ideology but rather to expressions of attack and defence, opposition and government. We tested this hypothesis by training on one parliament and testing on another in which party roles have been interchanged, and we find that the performance of the classifier completely disintegrates. But removing the factor of government–opposition status, as in the European Parliament, enables a more-ideological classification. Our results suggest that the language of attack and defence, of government and opposition, may dominate and confound any sensitivity to ideology in these kinds of classifiers.

1. Introduction

There have been a number of attempts recently to develop methods to automatically determine the ideological position of a political text. For example, one might wish to take a newspaper editorial or a blog and classify it as socialist, conservative, or Green. In practice, much of the research has taken speeches by members of a legislature (such as the U.S. Congress or the European Parliament) as the text to be classified and indicators such as party membership or legislative voting patterns as a proxy for ideology (indeed, Yu et al. (2008) use the terms *party classifier* and *ideology classifier* almost interchangeably); thus the problem becomes one of predicting one of these indicators from speech. One might expect, a priori, that

---

* This is an extended version of "Party status as a confound in the automatic classification of political text" by Graeme Hirst, Yaroslav Riabinin, and Jory Graham, *Proceedings, 10th International Conference on Statistical Analysis of Textual Data (JADT 2010),* Rome, June 2010. This work is financially supported by the Natural Sciences and Engineering Research Council of Canada.

methods based solely on the vocabulary used in a text would not be effective, because the members of a legislature, regardless of ideology, are all discussing the same topics – e.g., the legislation before them or the issues of the day – and hence would all be using the same topic-derived vocabulary (Mullen and Malouf 2006). The ideology expressed in a text would thus be apparent only at the sentence- and text-meaning levels. Nonetheless, one might hypothesize that different ideological frameworks lead to sufficiently different ways of talking about a topic that vocabulary can be a discriminating feature (Lin et al. 2006). And indeed, several studies have obtained notable results merely from classification by support-vector machines (SVMs) with words as features ('bag-of-words classification').[1]

For example, Thomas et al. (2006) examined speeches made by members of the U.S. House of Representatives to try to determine whether each speaker supported or opposed the proposed legislation under discussion. They combined bag-of-words text classification by SVMs with textual information about each speaker's agreement or disagreement with other speakers, obtaining an accuracy of around 70% (the majority baseline was 58%). Greene (2007) obtained an improved accuracy of over 74% on the same task by annotating each word with its grammatical relation from a dependency parse. Jiang and Argamon (2008), on the related task of classifying political blogs as liberal or conservative, improved results over using word features of the whole text by first trying to identify subjective sentences and the expressions of opinion that they contain, and then limiting the features to those parts of the text.

Diermeier et al. (2007) used SVMs with bag-of-words features to classify members of the U.S. Senate by ideology, labelling each speaker as a liberal or a conservative, and achieved up to 94% accuracy. However, in these experiments, the authors focused on 'extreme' senators – the 25 most conservative and the 25 most liberal members in each Senate. On 'moderate' senators, the results were notably poorer (as low as 52% accuracy). Moreover, there was considerable overlap between the training and testing portions of Diermeier et al.'s dataset, since they extracted content from multiple Senates (101st to 108th) and since members of Congress tend to preserve their beliefs over time. Specifically, 44 of the 50 'extreme' Senators in their test set were also represented in the training data, which means that the classifier was already trained on speeches made by these particular individuals. Thus the classifier might be learning to discern speaking styles rather than ideological perspectives.

1. Observe that this goal differs from that of, e.g., Gryc and Moilanen (this volume) and Dahlberg and Salhgren (this volume), who aim to determine the position expressed in a text with regard to a particular topic, such as Barack Obama or 'outsiders' in Sweden. By contrast, the more general goal here is the ideological position underlying a text, independent of any particular topic.

Later work by the same authors (Yu et al. 2008) made no distinction between moderates and extremes; rather, they tried to classify all members of the 2005 U.S. Congress by party affiliation, achieving an accuracy of 80.1% on the House of Representatives and 86.0% on the Senate. The goal of their study was to examine the person- and time-dependency of the classifier by using speeches from both the Senate and the House and comparing the results. They found that party classifiers trained on House speeches could be generalized to Senate speeches of the same year, but not vice versa. They also observed that classifiers trained on House speeches performed better on Senate speeches from recent years than older ones, which indicates the classifiers' time-dependency.

We began the present work to see whether these kinds of bag-of-words SVM classification methods would hold up in analysis of speech in the Canadian Parliament (Section 3 below). Our results, however, led us to question whether vocabulary differences between parties really reflected ideology or whether they had more to do with each party's role in the Parliament, and we investigate this in Section 4 below.

2. Background: The Canadian party system and Parliament

The Canadian Parliament is a Westminster-style parliament. The party with the most seats in the House of Commons (albeit possibly a minority of them) forms the government; the other parties are the opposition. There may also be a few Independent (unaffiliated) members. In the last 12 years, there have been four or five parties in each Parliament. In broad terms the parties may be classified as conservative (Reform Party, Canadian Reform Conservative Alliance, Progressive Conservative Party, Conservative Party of Canada), liberal or centre (Liberal Party),[2] or left-wing (New Democratic Party and Bloc Québécois); see Collette and Pétry (this volume) for more discussion of the parties' left–right positions.

Both English and French are official languages of Canada. A speaker in Parliament may use either language, and will sometimes even switch between the two within a speech. Everything said in Parliament is professionally translated into the other official language, and the proceedings are published in both languages. Thus the published English text of the debates is a mixture of original English and translations from French, and the French text has the complementary distribution.

2. Thus in our data, all liberals are Liberals, but not all conservatives are Conservatives. Similarly, we distinguish between opposition parties – any party that is not the governing party – and the Opposition party – the opposition party from which the Leader of the Opposition is drawn.

3. First set of experiments: Classifying by party

The present work was intended as a prelude to a larger project on ideological analysis of text. Our first task, intended as a baseline, was to apply bag-of-words support-vector machine classification, as used by Diermeier et al. (2007) and Yu et al. (2008) on U.S. Congressional speech, to speech in the Canadian Parliament, to see whether we could classify the speech by party affiliation (as a proxy for ideology) and obtain similar results, despite the differences in the political systems of the two countries.

In Canadian politics, unlike those of the U.S., party discipline is strong and (with only rare exceptions) all members of a party will vote the same way. The governing party will always vote to support its legislation; an opposition party might oppose it or support it. Thus (in contrast to the tasks described by Diermeier et al. and Yu et al.), there is no meaningful distinction between predicting voting records from parliamentary speech and predicting party affiliation. On one hand, it might be argued that this makes the task easier because parliamentary speech is likely to be highly partisan. On the other hand, it might be argued that it makes the task more difficult, because there is a greater diversity of views with precisely the same voting pattern, and so the classification is less straightforward.

In order to avoid the problems inherent in cross-time analysis, as highlighted by the work of Diermeier et al. (2007), we focus in this section on a single time period, so that there is a one-to-one mapping between members of Parliament (MPs) and documents in our dataset. Each document is a concatenation of all the speeches made by a speaker, and no other document contains text spoken by that person. Thus no speaker appears in both training and test data.

3.1 Data

We used both the English and French *House of Commons Debates* ('Hansard') for the first 350 sitting days of the 36th Parliament (1997-09-22 to 2000-05-10). In the 36th Parliament, a majority government was formed by the Liberal Party, led by Jean Chrétien. This data was available in a convenient plain-text form with sentence breaks identified (Germann 2001), as it has been widely used for research in machine translation.

We considered two sections of the proceedings: the debates on legislation and other statements by members ('Government Orders') and the oral question period. And we focused on the governing Liberal Party and the opposition conservative

parties,[3] in order to do a binary discrimination, liberal versus conservative; the left-wing parties had relatively few members in this Parliament and were excluded from the analysis.

For each MP who was a member of one of the liberal or conservative parties, and for each language, we formed a 'document' by concatenating all their utterances in debates, question period, or both, throughout the Parliament. (For simplicity, we will refer to all utterances as 'speeches', regardless of their length, including questions and answers in the oral question period.) We experimented with a variety of pre-processing methods, including stemming the words or leaving them whole, removing or retaining stopwords (defined as the 500 most frequent words in the text), and removing or retaining rare words (defined as those occurring in fewer than five documents). (Details of these and other pre-processing matters are given by Riabinin 2009.) In some of our experiments, we discarded the data for members who said very little, or nothing at all, in question period or in debates, using 200 documents representing 121 liberals and 79 conservatives; in other experiments, we considered all 156 liberal and 79 conservative members who spoke at all.[4] In all, depending on our choices in pre-processing, we had about 4 million words in each language for liberals (of which approximately 900,000 were from the question periods) and 2.7 million for conservatives (of which approximately 500,000 were from the question periods).

Generally, these variations in pre-processing made little difference to the results. In this paper we report results for experiments on the texts for all speakers, with words left unstemmed and with rare words removed, which usually, though not invariably, gave the best results.

3.2 Method

Taking word-types as the features for classification – that is, regarding the document for each speaker as a bag of words – for each language we trained an SVM classifier for ideology as indicated by party membership, liberal or conservative.

3. At the time of this Parliament, the conservative parties were in disarray. The Opposition was the conservative Reform Party (which became the Canadian Reform Conservative Alliance in March 2000), but the conservative Progressive Conservative Party also held a number of seats.

4. Several members of the conservative parties either defected to the Liberal party or became independents during this Parliament; and one member of the left-wing NDP defected to a conservative party. We treated all these members as conservatives in our experiments; for details and rationale, see Riabinin (2009).

In training and testing, we used five-fold cross-validation. We experimented with four weighting schemes: *boolean* (presence of feature), *tf* (term frequency), *tf-norm* (term frequency normalized by document length), and *tf-idf* (term frequency by inverse document frequency). The best results were obtained with *tf-norm* and *tf-idf*; the results we present below all use the latter.

3.3 Results

Table 1 shows the accuracy of classification of party membership by the SVM for each language on the documents of each data set: oral question period (OQP), debates (GOV), and the two combined (OQP + GOV). In all cases, retaining the 500 most frequent features led to higher accuracy than removing them. The baseline method of choosing the larger class (liberal) for all members would give an accuracy of 65.5%. All our results are well above this baseline, and in fact reach almost 97% for oral question period in English when frequent words are retained. The reason for the discrepancy between this result and the 89.5% obtained for the same data in French is unclear, as the two texts are mutual translations and no such effect was seen with the debates texts.[5] We also observe that in three cases out of four, combining debates and question period in a single classifier is deleterious to accuracy compared to classifying each separately. Generally speaking, our results are similar to, or better than, those of Yu et al. (2008) on the U.S. Congress.

Table 1. Accuracy (%) of classification by ideology on speech in the oral question period (OQP) and debates (GOV) by liberal and conservative members of the 36th Parliament, with and without removal of the 500 most frequent features (majority baseline = 65.5%).

	OQP + GOV	**OQP**	**GOV**
With 500 most frequent features retained			
English	83.8	96.9	83.3
French	83.2	89.5	86.0
With 500 most frequent features removed			
English	78.7	92.9	79.6
French	80.8	84.8	83.5

5. Compare the results of Collette and Pétry (this volume) on the differences that they found in locating English and French political manifestos on a left–right spectrum, and the differences between languages that they adduce in explanation. In our results, however, while the accuracy obtained for each language sometimes varies quite noticeably within each condition, there is no apparent system in the differences; sometimes the English results are more accurate and sometimes the French results are; sometimes the difference is marginal.

3.4 Discussion

The higher accuracy of classification for question period than for debates suggests that the language of question period is in some way more partisan than that of debates. However, our examination of the most discriminative words suggests that this partisanship is not so much ideological as a matter of attack and defence. In particular, in the Canadian Parliament, the oral question period consists largely of hostile questions from members of the opposition parties to ministers of the government, with only occasional friendly questions from government backbenchers, which themselves often serve primarily to set up an attack on the opposition.[6] It's possible, therefore, that our classifier may be learning – at least in part – not to distinguish ideologies but to distinguish questions from answers or attack from defence, which is not the goal of our research. Table 2 shows the ten most discriminative English words for each side in question period. For the governing liberals, the top words are *hon* and *member*, as in *the hon. member for Halifax West*, which is how a minister from the governing party typically addresses a member who has asked a question. Also, the word *we* might be used by a minister to speak on behalf of the entire party or government when responding to questions. For the opposition conservatives, the word *why* serves the obvious purpose of posing a question, and the words *he* and *her* are likely used to refer to government ministers who are the targets of the questioning. Also, observe the use of words such as *bloc*, *reform*, and *opposite* by the liberals, and *prime* (as in *Prime Minister*) and *liberal* and *liberals* by the conservatives.[7] This lends further support to the hypothesis that the classifier is partially learning to distinguish government members from opposition members.

When frequent words are removed we see this effect less, with a corresponding drop in accuracy (see the second part of Table 1), but it does not disappear entirely. In this condition, we certainly see reflections of ideology in vocabulary. The liberal lexicon is characterized by words related to Québec (*French*, *Francophonie*, *MAI* [*Montréal Arts Interculturels*], *PQ* [*Parti Québécois*]) and various social issues (*housing, violence, humanitarian, youth, society, technology*), while the conservatives tend to focus on monetary concerns (*APEC, taxpayer, dollar, millions, paying, premiums*), aboriginal affairs (*native, Indian, chief*), and, to a lesser degree, national defence (*military, marshall*). Nonetheless, the governing liberals use language that

6. This contrasts with the practice in similar parliaments, such as those of Australia and the U.K., in which questions are more evenly balanced between those of the opposition and those of government backbenchers.

7. Interestingly, this tendency for the names of opponents to be discriminating features is the converse of what Lin et al. (2006) found in their analysis of an Israeli–Palestinian debate, in which naming one's own side was discriminating; but see Section 6.4.3 below.

Table 2. The top 10 English words characterizing each class in the oral question period.

Rank	liberal (government)	conservative (opposition)
1	hon	prime
2	member	why
3	we	liberal
4	opposite	solicitor
5	quebec	farmers
6	housing	finance
7	bloc	he
8	reform	liberals
9	québécois	hrdc[a]
10	women	banks

[a] HRDC = Human Resources Development Canada, a federal government department.

is generally positive (*congratulate, excellent, progress*) and is intended to create the appearance of a government at work (*established, inform, improve, assist, developing, promote*). In contrast, the opposition conservatives use negative words that are meant to call the government's competence into question (*justify, resign, failed, admit, refusing, mismanage*). So again, it seems that many of the features relate not to ideology but to attack and defence – not to the party's beliefs but to its status as government or opposition.

4. Second set of experiments: Classifying by party status

Even if a classifier for political speech were truly using features related to ideology, we would expect that at least some of these features would specifically pertain to views of current events and therefore, if it is trained on one Parliament, it will not perform as well on a different Parliament in which different events are current, as in the results of Diermeier et al. discussed in Section 1 above. Nonetheless, we would expect that many of the features will be invariant over time and that such a classifier will still perform much better than a baseline.

On the other hand, if the 'ideological' classifier is in reality using (solely or primarily) features related to government and opposition status, then training on one Parliament would carry over only to other Parliaments in which the parties hold the same status; if they swap roles, then the classifier will fail. Indeed, in such a case it might (or should!) perform *worse* than the majority baseline, tending to classify liberals as conservatives and vice versa. In our second set of experiments, we tested the hypothesis that the latter is the case – that an SVM bag-of-words

classifier for Canadian parliamentary speech is primarily sensitive to party status, not ideology. We also looked at the in-between case: training an 'ideological' classifier on data in which all combinations of ideology and party status are present.

4.1 Data

To test our hypothesis, we needed a Parliament in which, in contrast to the 36th Parliament, a conservative party was in government. We chose the recent 39th Parliament (2006-04-03 to 2008-09-07), with a minority Conservative Party[8,9] government led by Stephen Harper; the Liberal Party was in opposition, along with the New Democratic Party and the Bloc Québécois. The proceedings were downloaded from the Parliament of Canada website in HTML-formatted documents and processed into a format similar to that of the 36th Parliament data.

4.2 Method and results

4.2.1 *Replication of the first experiments on the new data*

We first replicated the experiments of Section 3 on the new data, discriminating liberal members from conservative members (there was sufficient data for 104 liberals and 130 conservatives) within the same Parliament. Training and testing with five-fold cross-validation on the 39th Parliament, we achieved results similar to those of the 36th Parliament, albeit with slightly lower accuracy, especially for the English OQP documents; see Table 3 and compare Table 1. In particular, the accuracy of the classification on French text of speakers in Government Orders is anomalously low (baseline level) compared to all our other results including those for the English translation of the same text; we have no explanation for this. We also observe that for this data, unlike the 36th Parliament, the strategy of removing the 500 most frequent words is sometimes superior to that of retaining them.

Examining the primary features used in the classification for oral question periods, we observed that several words 'swapped sides': four of the top 10 English words that characterized the liberals in the 36th Parliament characterized conservatives in the 39th Parliament, and the primary word that characterized conservatives in the 36th Parliament was the second word that characterized liberals in the 39th; see Table 4. This is evidence for our hypothesis that the classifier is really picking up features related to government and opposition status.

8. So in this Parliament, unlike the 36th, all conservatives are Conservatives.

9. http://www2.parl.gc.ca/housechamberbusiness/ChamberSittings.aspx

Table 3. Accuracy (%) of classification by ideology on the 39th Parliament, with and without the 500 most frequent words retained (majority baseline = 55.8%).

	OQP + GOV	OQP	GOV
With 500 most frequent features retained			
English	83.8	88.3	72.3
French	75.5	88.8	56.8
With 500 most frequent features removed			
English	79.9	83.5	73.2
French	79.0	88.2	57.2

Table 4. The top 10 English words characterizing each class in oral question periods in each Parliament (extending Table 2). Boldface indicates words that 'swap sides' between the two Parliaments. Boldface italic words characterize the governing side; the boldface roman word characterizes the opposition.

Rank	36th Parliament		39th Parliament	
	liberal (government)	**conservative (opposition)**	**liberal (opposition)**	**conservative (government)**
1	hon	**prime**	conservatives	***bloc***
2	***member***	why	**prime**	liberals
3	***we***	liberal	conservative	senate
4	opposite	solicitor	immigration	violent
5	quebec	farmers	mulroney	***we***
6	housing	finance	kyoto	***québécois***
7	***bloc***	he	admit	greenhouse
8	reform	liberals	minority	ndp
9	***québécois***	hrdc	promise	corruption
10	women	banks	her	***member***

4.2.2 *Classifying across Parliaments*

Again we used the proceedings of the 36th and 39th Parliaments, both English and French, but in each language we took the classifiers trained on one Parliament and tested them on the other. (In these experiments, we have the deprecated situation that some individual speakers, being members of both parliaments, occur in both the training data and the test data and thereby might give the classifier an unfair boost.) The results, shown in Table 5, are in all cases well below the majority baseline scores, just as we hypothesized; when party status changes, there are no constant ideological features to save the classifier.

We also tried training classifiers on the data of the two Parliaments combined. This dataset includes all combinations of ideology and party status – that is liberals in government, liberals in opposition, conservatives in government,

Table 5. Accuracy (%) of classification by ideology when training on one Parliament (36th or 39th) and testing on the other.

Training → Testing	**OQP + GOV**	**OQP**	**GOV**
36 → 39 *(Majority baseline = 55.8%)*			
With 500 most frequent features retained			
English	44.9	43.3	44.6
French	45.7	46.1	47.0
With 500 most frequent features removed			
English	46.2	44.6	44.1
French	43.5	49.6	43.5
39 → 36 *(Majority baseline = 65.5%)*			
With 500 most frequent features retained			
English	36.8	34.5	36.2
French	35.2	51.1	33.5
With 500 most frequent features removed			
English	35.0	49.6	42.7
French	36.4	51.1	33.5

and conservatives in opposition. Some speakers, those who were members of both Parliaments, appear with each possible party status, whereas others, those who were members of only one of the two Parliaments, appear in only one of these four conditions. A classifier trained on the former group performs at around the level of the majority baseline (Table 6); one trained on the latter does better (Table 7), but the results are overall below the level of the original experiments (Tables 1 and 3), especially for OQP data. (The exception is that the anomalously low results for French GOV data are not seen when frequent features are retained.)

Table 6. Accuracy (%) of classification by ideology on speakers who were members of both the 36th (liberal government) and 39th Parliament (conservative government), with and without the 500 most frequent words retained (majority baseline = 64.0%).

	OQP + GOV	**OQP**	**GOV**
With 500 most frequent features retained			
English	62.0	66.9	61.1
French	63.0	63.0	63.0
With 500 most frequent features removed			
English	64.0	66.9	59.4
French	64.0	64.0	64.0

Table 7. Accuracy (%) of classification by ideology on speakers who were members of either the 36th (liberal government) or 39th Parliament (conservative government), but not both, with and without the 500 most frequent words retained (majority baseline = 51.9%).

	OQP+GOV	OQP	GOV
With 500 most frequent features retained			
English	78.5	81.7	72.6
French	76.6	78.3	71.2
With 500 most frequent features removed			
English	76.3	73.5	71.9
French	75.0	76.1	61.9

4.2.3 *Including the other opposition parties*

Another way to see whether the classifier is more sensitive to party status than to ideology is to muddy the ideological waters by including the left-wing parties, which were in opposition in both Parliaments, in the analysis. If the classification were truly ideological, lumping these parties in with the other conservative (36th Parliament) or liberal (39th Parliament) opposition parties would markedly degrade the performance of the classifier. On the other hand, if party status is what matters, there should be little effect in doing so as the opposition parties will be more or less indistinguishable. We carried out this experiment on the English data with frequent words retained.

The results are shown in Table 8. They should be compared with the liberal/conservative results for the same Parliament and same processing method, shown in the first lines of Table 1 (96.9%, 83.3%) and Table 3 (88.3%, 72.3%). There is almost no degradation of performance on the 36th Parliament; for the 39th Parliament, there is a noticeable drop (10.12 percentage points) for the question period, but little for the debates.

Table 8. Accuracy (%) of classification of government and opposition (all parties) on English text of the 36th and 39th Parliaments with the 500 most frequent words retained (majority baselines = 51.5% and 59.4% respectively).

	OQP	GOV
36th	95.6	82.6
39th	78.2	70.9

4.3 Discussion

The results seen in Sections 4.2.1–3 are consistent with the hypothesis that the SVM bag-of-words classifier is sensitive not to expressions of ideology for which party membership is a reasonable proxy, but rather to expressions of attack and defence, opposition and government. When we train on one parliament and test on another in which party roles have been interchanged, the performance of the classifier completely disintegrates; the degradation is far worse than can be explained merely by the difference between the two parliaments in the vocabulary of the current topics of discussion. Some features that are indicative of each party 'swap sides' with the change of government. And combining ideologically inconsistent opposition parties in the classifier does not in most cases seriously degrade its performance.

5. Classification based on the emotional content of speeches

Recall that our feature analysis of the 36th Parliament showed that liberal members tended to use words that convey a more positive sentiment than those used by conservatives. This suggests that it might be possible to distinguish parties or ideologies (solely) by the emotional content of their speeches. Indeed, researchers such as James Pennebaker have made something of an industry of interpreting politicians from a statistical analysis of their use of a single category of words. For example, during the 2008 U.S. presidential election, Pennebaker (2008) wrote:

> Over the last few years, some have argued that the use of negations (e.g., *no, not, never*) indicate [*sic*] a sign of inhibition or constraint. Low use of negations may be linked to impulsiveness. … Across the election cycle, Obama has consistently been the highest user of negations – suggesting a restrained approach – where as [*sic*] McCain has been the lowest – a more impulsive way of dealing with the world.

Similarly, Pennebaker concluded that McCain's greater use than Obama of the first-person singular (*I, me, my*) signalled a likely greater openness and honesty.[10]

In the context of our results above, the questions we ask are not just whether liberals can be distinguished from conservatives in the Canadian Parliament

10. The validity of this kind of analysis is discussed and defended by Pennebaker et al. (2007a). But Pennebaker (2008) also writes: "No one should take any text analysis expert's opinions too seriously. The art of computer-based language analysis is in its infancy. We are better than tea-leaf readers but probably not much."

merely by the emotional content of their speeches, but also, if so, whether the feature actually discriminates ideology (in line with the stereotype of happy liberals, dour conservatives) or is again confounded by the parties' status in the Parliament.

5.1 Method and data

To test these questions, we used Pennebaker et al.'s (2007b) software *Linguistic Inquiry and Word Count (LIWC2007)*. LIWC counts the proportion in a text of particular words and word stems in over 60 categories, including linguistic properties (pronouns, adverbs, prepositions, etc), psychological denotation (positive emotion, negative emotion, etc), and various topics (work, money, religion, etc); it does not, itself, provide any interpretation of the counts.

For these experiments, we used the English speeches of the oral question periods and debates of the 36th and 39th Parliaments, excluding MPs who spoke very little. This gave us a dataset of documents for 200 MPs (121 liberals, 79 conservatives) in the 36th Parliament and 220 MPs (125 conservatives and 95 liberals) in the 39th Parliament. First, we ran LIWC on this data, which gave us a 64-component vector for each document, each component being the value that LIWC computed for the document for one of its categories. We then performed classification experiments on the data (with five-fold cross-validation) using this 64-component representation of the documents, in order to see whether positive and negative emotion were among the top discriminating features for liberals and conservatives, respectively. Then we repeated the classification, using *only* positive emotion and negative emotion (referred to as POSEMO and NEGEMO) as features. Finally, we performed a third experiment, in which affect was reduced to a single feature, the amount by which the positive emotion in the text exceeded the negative (i.e., POSEMO minus NEGEMO); this representation does not distinguish a completely unemotional text from one that contains emotion of each polarity in equal amounts.

5.2 Results

Table 9 shows the results of these experiments. In the first experiment, with 64 features, the accuracy for both datasets was equal to the majority baseline, because all MPs were classified as members of the majority party! In contrast, using only POSEMO and NEGEMO, either as two features or as a single feature, yielded a substantial improvement of up to 20.5 percentage points over the baseline (a relative

Table 9. Accuracy (%) of classification by party using LIWC features for English text of the 36th and 39th Parliaments' oral question period (OQP) and debates (GOV) (majority baseline = 60.5% and 56.8% respectively).

	36th		39th	
	OQP	**GOV**	**OQP**	**GOV**
64 features	60.5	60.5	56.8	56.8
POSEMO and NEGEMO	80.5	79.5	73.1	55.0
POSEMO minus NEGEMO	81.0	78.5	72.2	59.1

error reduction of 51.9%) for the 36th Parliament and 16.3 points for the oral question periods of the 39th. However, performance remained around baseline for the debates of the 39th Parliament.

Nonetheless, a feature analysis confirmed that in the 36th Parliament, positive emotion was among the top five liberal features and negative emotion was among the top ten conservative features, whereas in the 39th Parliament, positive emotion was the fourth feature for conservatives in oral question periods and sixth in debates, whereas negative emotion was eighth and tenth respectively for liberals. Hence, we can see that positive emotion is a characteristic of members of the governing party, and negative emotion is a characteristic of members of an opposition party; again, party status confounds ideological classification. The result of the classifier on all 64 features may be explained by the fact that no LIWC category had a *significant* impact on the classification. In other words, even though some LIWC categories were discriminating features for liberals and others were discriminating features for conservatives, the overall difference between the two groups was so slight that without feature selection the resulting classifier simply labelled all test instances as belonging to the majority class. This seems to be the case also for POSEMO and NEGEMO by themselves in debates in the 39th Parliament.

6. Third set of experiments: European Parliamentary data

If our 'ideological classifier' is in reality sensitive to government and opposition, then this effect should disappear when it is applied to data in which there is no government or opposition per se, but merely position-based debate with a more or less equal amount of attack and defence on both sides. Such a situation may be found in the European Parliament, in which a left–right ideological division dominates government–opposition divisions (Hix et al. 2007).

Our goal here is thus very similar to one of the tasks of the 2009 DEFT text mining challenge[11] (DEFT 09): classification by political group of speeches by Members of the European Parliament (MEPs). The DEFT corpus consisted of speeches from 1999 to 2004 by MEPs belonging to the five largest groups. Three teams attempted this task, but two declined to share their results. The remaining team, from the University of Montreal (Forest et al. 2009), reported an *F*-measure of about 0.33 on multiclass classification, which the organizers described as "mediocre" as the random baseline accuracy for the corpus was about 28% (Grouin et al. 2009: 49). We attempted both binary classification of left-wing versus right-wing MEPs, and multiclass classification of MEPs from the five largest groups, as in the DEFT task.

6.1 Data

We used English data from the proceedings of the European Parliament as our corpus.[12] Ranging from 2000 to early 2010, it was almost a strict superset of that used in the DEFT task. However, the data used in the DEFT task had been stripped of any explicit references to groups. Thus, tokens such as *PPE, Christian-Democrat,* and *United Left,* were all replaced with an anonymous tag. We understand that this was because, in the DEFT task, the organizers had human judges attempt a classification on the same data for comparison, and phrases such as *As vice-chairman of the PPE-DE group, I…* were presumably considered too much of a giveaway to a human reader. By contrast, we left all group names in place in our data. In Section 6.4.3 below, we will discuss the effect that anonymization has on classification.

6.2 Method

The choice of how to organize the raw text into vectors proved to be a key one. Our first approach was, for each MEP, to concatenate all of their utterances and consider that to be one document, as we did for the Canadian Parliament. This

11. DEFT (DÉfi Fouille de Textes) is an annual challenge and evaluation conference for researchers in text mining and classification. Each year, one or more tasks related to text mining are set, and training and test corpora are provided; research teams compete to get the best results. Results and methods are then discussed at the conference.

12. The data was collected and marked-up in XML by Dr Maarten Marx of the University of Amsterdam, who kindly made it available to us; see Marx and Schuth (2010) for details of the XML markup.

contrasted with the approach taken in the DEFT challenge, in which each individual speech remained a separate document. With our concatenation policy, however, we achieved accuracy only slightly above a random-guessing baseline. However, we observed that the amount of text we had per MEP varied widely, from a few hundred words to tens of thousands of words. Yet each MEP's document was being turned into a vector that affected the classifier equally, contradicting the natural notion that a document should have an affect on the classifier commensurate with its size. We rectified this by dividing each MEP's concatenated utterances into a number of equal-sized documents. We experimented with different document sizes, and found that it had a marked effect on accuracy (as shown in our results below).[13] The sizes that we used begin with 267 words, which was the average document length in the DEFT challenge.

As features, we used log *tf-idf*–weighted word types with words appearing in fewer than five documents removed (though we experimented with a variety of pre-processing methods, none of which had a profound effect on our results). We used SVMs for binary classification and SVM-multiclass for multiclass classification. All of the results presented below are the averaged results of five-fold cross-validation.

Binary classification: In performing binary classification, we were first faced with the task of meaningfully splitting the groups involved into left-wing and right-wing, a task that was further complicated by changes in groups and their names over the ten-year study period and by inconsistencies in identification of the groups in the data (e.g., *Greens, Verts*).[14] From descriptions of the groups, we classified as either broadly left or right 15 of the 18 affiliations observed in the data,[15] which we then grouped into the ten bins shown in Table 10.

Multiclass classification: We followed the example of the DEFT task in using only the five largest groups for multiclass classification (see Table 10), excluding the smaller right-wing groups. In multiclass classification, we found that tuning the error cost C on a logarithmic range of values was especially important, and that our best results were achieved with C on the order of 10^9.

---

13. If an MEP spoke significantly less than the document size, they were discarded from the data. Even with the highest value for document size (6666 words), this depleted the data by only 2%. In our earlier Canadian experiments described above, we discarded small documents but we did not subdivide large ones.

14. Group abbreviations usually appeared in French – e.g., *PSE* rather than *PES* for the Party of European Socialists – even in the English data. Here, we use the predominant label.

15. Omitted were the *non-inscrits* (independents), the Technical Group of Independents (a group described as politically heterogeneous), and the Alliance of Democrats and Liberals (ALDE) (described variously as conservative liberals, or as centrist).

Table 10. European political groups as clustered, ordered from left-wing to right-wing. For the purposes of binary classification, groups above the centrist group ALDE are considered left-wing (L), and all groups below are considered right-wing (R). Asterisks mark the five largest groups, which were used in the multiclass classification experiments.

Group	Speakers in corpus	Description	L/R
*NGL	104	Communist / far-left	L
*PSE	446	Social democrats	L
*Greens	114	Green	L
*ALDE	195	Liberal / centrist	–
*PPE	571	Conservative / Christian democrat	R
ECR	41	Conservative	R
EDD	75	Eurosceptic	R
UEN	75	National conservatism	R
EFD	22	Eurosceptic, national conservatism	R
ITS	18	Far-right nationalist	R

6.3 Results

Table 11 shows the accuracy of binary left–right classification with varying document sizes. The ten words most characteristic of each class are shown in Table 12.

Table 13 shows the accuracy of multiclass classification for varying document sizes. The confusion matrix for multiclass classification is shown in Table 14; it reflects a limited subset of the data, chosen so that each group was equally represented.[16] Table 15 shows the ten words best characterizing each of the five classes.

Table 11. Precision, recall, and accuracy (%) of left–right classification on speech in the European Parliament, with varying document sizes. Baseline accuracy (more frequent class) is 50–51%, varying slightly with document size.

Document size (words)	Precision	Recall	Accuracy
267	62.6	65.2	62.3
833	67.6	70.1	67.4
1667	69.9	71.9	69.8
3333	72.9	77.5	73.9
6666	77.6	81.3	78.5

16. Note that the confusion matrix reflects a sample of 1350 documents from the almost 3000 that we considered. We chose this sample so that each group had an approximately equal number of documents. (Classification of the full set of documents tended to favour the groups which were heavily represented, thus obscuring the measure of ideological similarity we were looking for.)

Table 12. The top 10 English words characterizing each class in left–right classification of speech in the European Parliament.

Rank	Left-wing	Right-wing
1	socialist(s)	subsidiarity
2	unions	christian
3	pse	strasbourg
4	employees	competitiveness
5	greens	healthy
6	scotland	prosperity
7	gender	democrats
8	equality	competitive
9	supports	communist
10	myself	truth

Table 13. Accuracy (%) of five-way multiclass classification of speech in the European Parliament by political group, with varying document sizes. Baseline accuracy (most frequent group) is 38–39%, varying slightly with document size.

Document size (words)	Accuracy
267	44.0
833	48.0
1667	52.7
3333	56.2
6666	61.8

Table 14. Confusion matrix for multiclass classification of speech in the European Parliament by political group. Column headings are our classifications, rows are true affiliations. Boldface indicates correct classifications; italics indicates incorrect classification of a group as an ideologically adjacent group. Shaded cells show confusion between the PPE and the PSE.

	NGL	PSE	Greens	ALDE	PPE	Total
NGL	**204**	*17*	36	9	10	276
PSE	*16*	**136**	*20*	34	71	*277*
Greens	20	*25*	**153**	*30*	16	244
ALDE	3	39	*14*	**170**	*50*	276
PPE	3	65	9	*41*	**159**	*277*
Total	246	282	232	284	306	1350
Accuracy (%)	73.9	49.0	62.7	61.5	57.4	61.8

Table 15. The top 10 English words characterizing each group in multiclass classification of speech in the European Parliament.

Rank	NGL	PSE	Greens	ALDE	PPE
1	confederal	socialist	greens	liberal	christian
2	nordic	socialists	alliance	liberals	subsidiarity
3	military	pse	nuclear	eldr	conservatives
4	unemployment	institution	ngos	democrat	morning
5	profits	eplp	basque	alde	wrong
6	occupation	balanced	scotland	alliance	competitiveness
7	nato	interinstitutional	comments	rapporteurs	healthy
8	liberalization	millennium	planes	obvious	competitive
9	yugoslavia	repeatedly	ale	china	communist
10	militarisation	portuguese	conflict	7	phenomenon

6.4 Discussion

6.4.1 *Comparison with DEFT results*

With equal average document length, our multiclass accuracy was only marginally better than the DEFT results of Forest et al. (2009) (about 5 percentage points over baseline, rather than 2 points). However, as document size was raised to a maximum of 6666 words, accuracy increased steadily, up to 61.8%, about 23 points over baseline. This suggests that the average DEFT size of 267 words is simply an insufficient size for bag-of-words-based methods over such a noisy corpus.

6.4.2 *Relative difficulty of classification tasks*

The accuracy of multiclass compared to binary classification suggests that associating a speech with a specific group is not much harder than just classifying it as left or right. This may, more than anything else, speak to the composition of the European Parliament. Hix, Noury, and Roland (2007) suggest that, rather than falling at some point on a line from left to right, European Parliament groups can be placed in a space where the primary dimension "is the traditional left–right axis and the second dimension is a mixture of attitudes towards European integration (in favour and against) and government–opposition status in the EU" (p. 217). In addition, 'green' sentiment, while often lumped in with liberalism, is not quite a strict subset thereof (and implies a completely different vocabulary). These multiple dimensions complicate the task of binary ideological classification.

The confusion matrix for multiclass results (Table 14) may shed some light on the relative ideological distances between groups. As we would wish, confusion is for the most part clustered around ideologically similar groups. Because

groups are arranged from left to right in order of ideology, this fact is reflected by the tendency of confusion to cluster around the diagonal. The most surprising result is the high amount of confusion between PSE (a socialist group) and PPE (Christian democrats, the most conservative group we considered) (shaded cells). This may be because these two groups had perhaps the least coherent feature lists (see Section 6.4.3 below). It may be significant that the two most accurately classified groups, NGL and the Greens, also had the most subjectively coherent feature lists.

6.4.3 *Discriminative features*

A few trends emerge from the lists of the top features for each group. The most obvious is that MEPs tend to talk about their own groups. Hence, the top feature for the Greens is *greens*, the top two features for the PSE are *socialists* and *socialist*, and so on. This contrasts with Canadian MPs who we found (Section 3.5 above) tend to talk about their opponents more than themselves. This striking difference demonstrates the domain-specificity of the features learned by the classifier. We do however find some instances of MEPs talking about their opponents, most notably the appearance of *communist* in the top 10 features of the PPE, the most right-wing group we looked at. As might be expected, the contexts in which PPE MEPs actually used the word were highly negative, phrases such as *communist tyranny*. But clearly, whether MEPs talk about themselves or their opponents, the names that each group tends to utter are important discriminators. Thus we see a second reason why the results of Forest et al. (2009) were so poor; anonymization of the groups removes crucial discriminators.

Some of the top feature lists are highly coherent with respect to the issues of concern to the group. For example, among the top 50 features for the Greens we find *nuclear*, *organic*, *contaminated*, *ecological*, *toxic*, *culling*, and *depleted*. The top 50 features for NGL, the most left-wing of the groups, included *wages*, *unemployment*, *capitalist*, *wage*, *inequality*, and *poverty*. The top features for right-wing PPE are less coherent, though as Diermeier et al. (2007) found of right-wing U.S. senators, there tended to be a focus on cultural and moral issues: *christian*, *moral*, *conscience*, *faith*, and *euthanasia* all appear in their top 100 features.

Some trends that the classifier seems to pick up on aren't overtly ideological, and indeed hint at the language of attack and defence. In the case of centre-left group PSE, the classifier seems tuned to the language of felicitation, with words like *wholehearted*, *congratulations*, *congratulating*, *impressed*, *proud* and *achievement* all in their top 50, whereas the centrist group ALDE seems to be associated with censorious language: *accountability*, *needless*, *shameful*, *shame*, *breaches*.

7. Conclusion

Our results cast doubt on the generality of the results of research that uses words as features in classifying the ideology of speech in legislative settings – and possibly in political speech more generally. Rather, the language of attack and defence, of government and opposition, seems to dominate and confound any sensitivity to ideology. Such research therefore reduces in effect to the classification of support or opposition, much as in the linguistic component of the work of Thomas et al. (2006) described in Section 1 above. However, even if our classifiers are construed as distinguishing support from opposition, our results are much more accurate than those of Thomas et al., even though we did not use any explicit component for detecting agreement or disagreement between individual speakers. This may be partly attributed to one of the differences between Canadian and U.S. politics: Canadian parties have strong party discipline, and agreement between speakers may be reliably inferred from shared party membership.

Our results contrast with the conclusions of Diermeier et al. (2007), who argue from their own results that speakers' words in debates in the U.S. Congress are "expressions or representations of an underlying belief system". Again, political differences might be a partial explanation of the difference. Perhaps the weak party discipline of the U.S. and the separation of the Congress from the Executive branch motivates greater attention to ideological substance in debates than does the Canadian (Westminster-style) system in which an explicit governing party, including the head of government and all cabinet ministers, is represented as such in the legislature. This possibility is supported by our results from European Parliamentary data. But this is speculation; our results have demonstrated a confound that must be taken into account in research on ideological classification of speech in any context.

References

DEFT (2009). Actes de l'atelier de clôture de la cinquième édition du DÉfi Fouille de Textes, Paris, 22 June 2009.

Diermeier, D., J.-F. Godbout, B. Yu and S. Kaufmann. 2007. Language and ideology in Congress. *Annual Meeting of the Midwest Political Science Association*.

Forest, D., A. van Hoeydonck, D. Létourneau and M. Bélanger. 2009. Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents. *Actes de l'atelier de clôture de la cinquième* édition *du DÉfi Fouille de Textes,* Paris, 22 June 2009, pp. 75–88.

Germann, U. 2001. Aligned Hansards of the 36th Parliament of Canada. Available at www.isi.edu/natural-language/download/hansard/

Greene, S. 2007. Spin: Lexical semantics, transitivity, and the identification of implicit sentiment. PhD thesis, University of Maryland, College Park.

Grouin, C., B. Arnulphy, J.-B. Berthelin, S. El Ayari, A. Garcia-Fernandez, A. Grappy, M. Hurault-Plantet, P. Paroubek, I. Robba and P. Zweigenbaum. 2009. Présentation de l'édition du D'Efi Fouille de Textes (DEFT'09). *Actes de l'atelier de clôture de la cinquième* édition *du DÉfi Fouille de Textes,* Paris, 22 June 2009, pp. 35–50.

Hix, S., A.G. Noury and G. Roland. 2007. *Democratic Politics in the European Parliament*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511491955

Jiang, M. and S. Argamon. 2008. Political leaning categorization by exploring subjectivities in political blogs. *Proceedings, 4th International Conference on Data Mining (DMIN 2008)*, pp. 647–653.

Lin, W., T. Wilson, J. Wiebe and A. Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. *Proceedings of the 10th Conference on Natural Language Learning (CoNLL-X)*, pp. 109–116.

Marx, M. and A. Schuth. 2010. DutchParl: The parliamentary documents in Dutch. *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

Mullen, T. and R. Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. *Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159–162.

Pennebaker, J.W. 2008. The meaning of words: Obama versus McCain. Weblog, 12 October 2008. Available at wordwatchers.wordpress.com/2008/10/12/

Pennebaker, J.W., C.K. Chung, M. Ireland, A. Gonzales and R.J. Booth. 2007a. *The Development and Psychometric Properties of LIWC2007*. Available at www.liwc.net/LIWC2007 Language-Manual.pdf

Pennebaker, J.W., C.K. Chung, M. Ireland, A. Gonzales and R.J. Booth. 2007b. Linguistic inquiry and word count (LIWC2007). Available at www.liwc.net

Riabinin, Y. 2009. Computational identification of ideology in text: A study of Canadian parliamentary debates. MSc paper, Department of Computer Science, University of Toronto, January 2009. Available at www.cs.toronto.edu/compling/publications.html

Thomas, M., B. Pang and L. Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pp. 327–335. DOI: 10.3115/1610075.1610122

Yu, B., S. Kaufmann and D. Diermeier. 2008. Classifying party affiliation from political speech. *Journal of Information Technology in Politics* 5(1), pp. 33–48. DOI: 10.1080/19331680802149608