Agenda-setting by the European Commission: Using computer-assisted classification methods to classify policy documents

Frank M. Häge
University of Limerick

Abstract

The European Commission is often considered to be the main agenda-setter in EU policy-making. The Commission produces hundreds of policy documents every year. However, reliable quantitative accounts of the agenda-setting activity of the Commission over extended periods of time and across several policy areas are not available. We need such accounts if we want to be able to adjudicate between competing theories of the role of the Commission in the European integration process. While information on the Commission's agenda-setting activity is now in principle available in online databases such as PreLex, a major problem is the absence of a comprehensive policy classification scheme with mutually exclusive categories. For a considerable number of documents in PreLex, policy labels are either missing or several labels are assigned to a single document, making the classification ambiguous. In this study, I conduct classification experiments to investigate the performance of computer-assisted document classification methods for generating correct policy labels for Commission documents solely based on the words in their titles. The findings indicate that the support vector machine classifier is able to classify about 75 percent of the documents into the correct policy category, regardless of which combination of pre-processing options are chosen to generate the input matrix for the analysis.

Mapping the Commission's agenda-setting activity

The European Union's (EU) policy competences have grown extensively since its foundation in 1957. In the early years of European integration, the EU's predecessor organizations were mainly concerned with regulating its members' coal and steel industries, establishing a customs union, and setting up the common agricultural policy. Over time, the EU gained jurisdiction in a number of newly emerging as well as already existing policy fields. Currently, much legislation and regulation in the fields of environment, industry, transport and telecommunications, energy, consumer protection, monetary policy, and justice and home affairs originates from the European level. Even in areas that are still largely the domain of member states, such as health policy, taxation, social policy, and foreign and defence policy, the EU is playing a significant role. How can we explain this remarkable development? Some researchers argue that European integration is the result of conscious decisions by government leaders (e.g. Moravcsik 1998), while others see it as an unintentional consequence of the empowerment of supranational institutions like the European Court of Justice and the European Commission (e.g. Sweet & Brunell 1998). A third perspective takes the middle ground and argues that the discretion and influence of supranational institutions depends on the preferences of member states and the EU's institutional decision-making rules (e.g. Tsebelis & Garrett 2001).

In order to examine these competing theoretical claims about the influence of supranational institutions empirically, we need reliable and valid measures of what these institutions actually do. In this paper, I focus on the agenda-setting activity of the European Commission. Expanding the reach of EU policy implies more power and resources for the Commission and its office holders. Thus, the Commission has sufficient incentives to promote further integration. In addition, the Commission does not only have the motive, it also has the means to promote further integration. In most policy areas, the Commission has the exclusive right to introduce legislative proposals. As the first mover in the legislative game, it can exploit differences in opinion amongst member states and select the winning coalition that yields the outcome most favourable to its own ideas (e.g. Steunenberg 1994; Crombez 1996). Another advantage of being the first mover is the possibility of setting the broad parameters of the subsequent discussions. By framing the political debate, the Commission is able to exclude some issues and highlight others, thereby steering the debate into a direction that suits its interests (for a review, see Daviter 2007). Besides exploiting institutional (e.g. Shepsle & Bonchek 1997: 102) and rhetorical constraints (List 2004), the

Commission can also exert influence by changing member states' preferences. The Commission is often the only actor that has a Europe-wide overview of existing regulation in a certain policy field and it can draw on an extensive expert network to develop policy ideas and proposals (e.g. Gornitzka & Sverdrup 2008). Thus, the Commission's informational advantages put it in a good position to persuade member states of the merits of its proposals.

Much of the Commission's agenda-setting activity manifests itself in formal policy documents, ranging from proposals for generally binding legislative acts to study reports produced by Commission staff or external consultants. The Commission transmits these policy documents to the legislative institutions, the Council and the European Parliament (EP), for further discussions and, where appropriate, a final decision. For an examination of competing claims about the agenda-setting power of the Commission, we can consider the co-variation between the, possibly importance-weighted, number of Commission documents and the increase in EU competences and legislative activity in specific policy areas. However, in order to measure the Commission's agenda-setting activity in a certain policy area in a certain period of time, we need a policy classification scheme with an exhaustive and mutually exclusive set of categories. While the Commission's own database of policy documents (PreLex) tags each entry with policy labels, several labels are often attached to a single document, making an unambiguous classification of such documents impossible. The goal of this study is to explore the usefulness of applying computer-assisted document classification techniques to the classification of Commission policy documents. More precisely, the study examines the prediction accuracy of computer-assisted document classification techniques in classifying Commission documents into pre-existing policy categories.

Methods for computer-assisted text analysis have received growing attention in Political Science in recent years. In the next section, I discuss existing works employing these techniques and how they relate to the current study. Subsequently, I describe the dataset and its collection. The study relies on a dataset of 27,000 Commission documents transmitted to the EU legislative institutions between 1975 and 2007. The text analysis focuses on the titles of those documents, which have been extracted from the Commission's online database PreLex. After describing how the text elements were pre-processed, I conduct several experiments examining the validity of support vector machine classification results using different input data. The results are promising, with the experiments generally indicating a correct prediction rate just below 75 percent. After using the support vector machine to assign policy labels to documents for which the policy descriptor is missing or ambiguous, I briefly

present some descriptive statistics about the Commission's agenda-setting activity over time and across policy areas. In the concluding section, I summarize the analysis and its findings and point to possible improvements in future research.

Computer-assisted document classification in Political Science

Recent years have seen considerable growth in the use of automated content analysis methods in Political Science. We can divide the existing literature in roughly two groups according to their research goals. The first group consists of studies that use political text as data to generate measures of actors' positions on policy issues or ideological dimensions. The goal of these studies is to locate each document in a continuous policy or ideology space. Popular data sources in this line of research are party manifestos (Pennings & Keman 2002; Laver *et al.* 2003; Slapin & Proksch 2008) and parliamentary speeches (Proksch & Slapin 2009). Some of the authors active in this area have also developed valuable new tools for this type of text analysis (Laver *et al.* 2003; Slapin & Proksch 2008). The second group of studies is concerned with classifying documents into distinct categories. Document classification techniques have been used to classify policy documents into policy categories (Purpura & Hillard 2006; Hillard *et al.* 2008; Quinny *et al.* 2008) and law-makers' speeches into ideological groups or parties (Diermeier *et al.* 2006; Yu *et al.* 2008; Hoyland & Godbout 2009).

As I am concerned with classifying documents into distinct nominal categories, it is useful to briefly discuss the approaches and findings of this second type of research in more detail. On methodological grounds, applications of automated topic classification techniques can be subdivided into projects that apply supervised learning algorithms and projects that apply unsupervised learning algorithms. In supervised learning, the classifier is first trained on a subset of the data for which categories have been assigned by human coders. Subsequently, the classifier is used to predict category membership of documents in a dataset with unknown category membership. In contrast, unsupervised learning techniques are a form of cluster analysis. They rely purely on patterns in the data to inductively determine category membership.

The unsupervised learning approach is favoured by Quinn and colleagues (Quinny *et al.* 2008), who use it to develop a policy classification scheme based on 118,000 speeches given by US Senators between 1997 and 2004. While the authors are correct in noting that unsupervised learning has benefits in that it does not require possibly unwarranted assumptions about the category structure and comes with much less costs in terms of research

resources, the approach is not useful for my purposes. About two-thirds of the documents in my dataset are already assigned into a unique policy category. Only the remaining one-third of documents that lack any policy category assignment or are assigned into several categories need to be classified. Not relying on the already classified documents to develop a classifier for the un-classified documents would be a waste of valuable information. Supervised learning techniques are more appropriate in this respect.

The supervised learning approach has been popular for classifying legislators into political groups. Diermeier and collaborators (2006) trained a support vector machine on 350 speeches given by the 25 most liberal and the 25 most conservative US Senators between 1989 and 2001 (101st to 107th Congress). The Senators were identified on the basis of ideal point estimates generated by roll-call analysis. The resulting classifier assigned 47 out of the 50 (94 percent) most extreme Senators in the 108th Congress correctly into the two ideological groups. Yu, Kaufmann, and Diermeier (2008) followed up on this study by investigating the source- and time-dependency of different party classifiers. Next to support vector machines, they also applied so-called Naïve Bayes classification methods. To examine source dependency, classifiers trained on speeches given by US Representatives in the House in 2005 were used to predict group membership of Senators in the same year and vice versa. To examine time dependency, the 2005 House classifiers were used to predict group membership of Senators in individual years from 1989 to 2006. The results indicate considerable variation in the prediction accuracy of classifiers, depending on the type of classifier and the pre-processing of the input data. Also, while classifiers based on 2005 House speeches correctly predicted group membership of up to 88 percent of the 2005 Senators, the maximum accuracy of classifiers trained on 2005 Senate speeches and used to predict group membership in the 2005 House was only 68 percent. Furthermore, the prediction accuracy of the 2005 House classifiers decreased with Senate speeches further back in time. In general, the results indicate substantial dependency of classifiers on time and document source.

Høyland and Godbout (2009) rely on legislative speeches to classify Members of the European Parliament (MEPs) into seven different party groups. They train a support vector machine on the speeches of 615 MEPs in the 5th European Parliament (EP) and use it to predict party group membership of the 623 MEPs in the 6th EP. Using actual party group membership in the 6th EP as an evaluation standard, their classifier reaches a prediction accuracy of 52 percent. Keeping in mind that the classification task includes seven categories,

this result is still a considerable improvement over classification results generated by chance alone.

Purpura and Hillard's (2006) work is most related to the current study in that they apply supervised learning methods to classify policy documents into a multi-category policy scheme. They base their analysis on a dataset of 108,000 Congressional bills introduced since 1947, which were manually classified into 21 major policy categories and 226 subcategories. The dataset was split randomly into training and test sets of equal size. The generated support vector machine classifier reached a prediction accuracy of 82 percent with respect to the major categories and 71 percent with respect to the subcategories. Hillard, Purpura, and Wilkersion (2008) repeat this analysis relying on a larger dataset of 379,000 Congressional bills. In addition, they compare the performance of the support vector machine algorithm to three other classification techniques and to a so-called ensemble prediction. The ensemble prediction is generated by aggregating the results of the three best-predicting classifiers. The comparison shows that the support vector machine performs better than other classification methods and just as well as the ensemble approach. The prediction accuracy for the major categories lies at 89 percent and the prediction accuracy for the minor categories at 81 percent. Hillard et al.'s (2008) findings promise great value for applying support vector machines to classify policy documents. However, as Yu et al.'s (2008) research shows, the results of automated document classification can be strongly influenced by the peculiarities of a specific data source. Thus, whether or not support vector machines can also be successfully applied to classify Commission policy documents needs to be established independently. Examining this question is the goal of the following analysis.

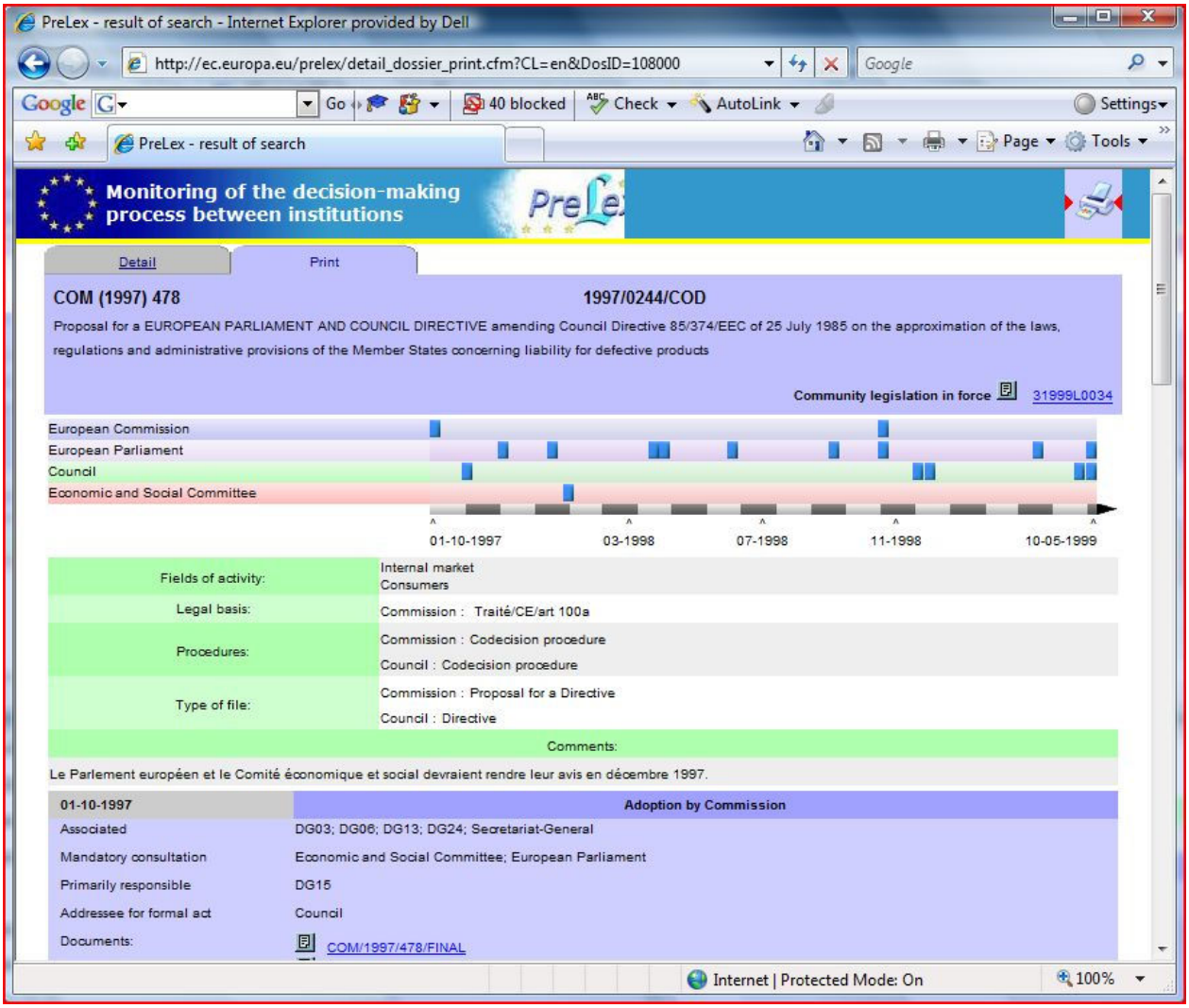
Data collection

The text data used in the classification analysis consists of the titles of policy documents transmitted by the European Commission to the legislative institutions between 1975 and 2007. The document titles were collected as part of a larger research project on the Commission's agenda-setting power, which involved the extraction of the information contained in the Commission's online database PreLex. The extraction was automated through a computer script[1]. In a first step, the script collected the unique identification numbers for each policy document contained in the database. The basic search function of the database returns a list of hyperlinks to the documents retrieved as a result of the search query. Each hyperlink contains a unique identification number for the respective document,

[1] The script was written in Python 2.6, which can be downloaded from www.python.org.

otherwise the links are identical. The computer script extracted these identification numbers and used them to download the HTML source code of individual webpages and to save them as text files on the local hard drive. The screenshot in Figure 1 gives an example of a PreLex webpage.

Figure 1: Example webpage from PreLex database



Source: PreLex (http://ec.europa.eu/prelex/apcnet.cfm?CL=en)

The address field in the internet browser shows the address information of the webpage with the unique identification number, in this case the number is '108000'. The title of the policy document is given at the top of the page. Here, the title indicates that the document is a proposal for a directive to be adopted by the EP and the Council on harmonising member states' rules concerning the liability of defective products. The relevant policy categories are mentioned below the timelines as 'fields of activity'. The example illustrates the main problem of the policy classification scheme as it exists in PreLex: a single document is

categorized into several policy categories. In this case, the proposed directive has been assigned to the 'Internal Market' as well as to the 'Consumer' policy category.

Table 1: Types of policy documents included in the dataset

Type of policy document	Frequency	Proportion	Cumulative
Regulation	8,812	32.5	32.5
Decision	5,356	19.8	52.3
Communication	3,741	13.8	66.1
Report	3,183	11.7	77.8
Directive	1,915	7.1	84.9
Transfer of appropriations	1,491	5.5	90.4
Working paper	1,178	4.4	94.7
Assent	303	1.1	95.9
Opinion	269	1.0	96.8
Recommendation	181	0.7	97.5
Resolution	103	0.4	97.9
Preliminary draft supplementary budget	85	0.3	98.2
Letter of amendment	78	0.3	98.5
Green paper	76	0.3	98.8
Preliminary draft budget	65	0.2	99.0
Memorandum	39	0.1	99.2
Programme	26	0.1	99.3
Periodic report	25	0.1	99.3
White paper	25	0.1	99.4
Letter	23	0.1	99.5
Missing	22	0.1	99.6
Other documents	108	0.4	100.00
Total	**27,104**	**100**	

Source: Own data generated from information in PreLex (http://ec.europa.eu/prelex/apcnet.cfm?CL=en).

In a second step, the script navigated through the HTML structure of each of these webpages to identify and copy the relevant information into a database[2]. The result of this data extraction procedure is a dataset containing information on 27,104 policy documents. Most important for the purposes of the current study, it includes the title of each document and information on how Commission staff classified the document in terms of policy content. Table 1 gives an overview of the types of policy documents contained in the dataset. The dataset includes many formal proposals for binding legislation, such as decisions, regulations,

[2] The Python module BeautifulSoup was instrumental in this respect. BeautifulSoup is a HTML parser that reads the source code of webpages and memorizes their HTML structure. Subsequently, BeautifulSoup methods can be used for navigating through the HTML structure. The module is not part of the core release of Python, but can be downloaded from the internet (www.crummy.com/software/BeautifulSoup [accessed on 13 September 2008]).

and directives, as well as for non-binding recommendations, resolutions, and opinions. The dataset also includes Commission communications, working papers, reports, and so-called green and white papers. These documents are of a purely informative nature and elaborate on the policy views and plans of the Commission. In addition, the dataset includes various policy instruments related to the budget, particularly transfers of appropriations, and a variety of less common types of policy documents.

Table 2: Distribution of documents across policy fields

Policy field	Frequency	Proportion
Agriculture	4,159	15.3
Budget	2,226	8.2
Customs Union	1,641	6.1
External Relations	1,475	5.4
Commercial Policy	1,407	5.2
Internal Market	1,134	4.2
Fisheries	1,072	4.0
Transport and Telecommunications	855	3.2
Development Policy	810	3.0
Social Policy	727	2.7
Environment	585	2.2
General Affairs	555	2.1
Economic and Monetary Policy	540	2.0
Energy	462	1.7
Research	429	1.6
Financial Affairs	417	1.5
Education and Culture	229	0.8
Justice and Home Affairs	212	0.8
Regional Policy	154	0.6
Consumer Policy	117	0.4
Health	113	0.4
Total with known policy category	**19,319**	**71.2**
Multiple policy areas	4,784	17.7
Missing	3,001	11.1
Total with unknown policy category	**7,785**	**28.8**
Total	**27,104**	**100.0**

Source: Own data generated from information in PreLex (http://ec.europa.eu/prelex/apcnet.cfm?CL=en).

Table 2 presents the distribution of documents across policy fields after aggregating the 43 'field of activities' labels of the PreLex database into a more manageable number of 21 policy categories. Some of the PreLex policy field labels overlap considerably or stand in a hierarchical relationship with each other. For example, the labels 'Justice and Home Affairs'

and ‘Justice, Freedom, and Security’ are likely to refer to the same types of policies and the label ‘Economic and Monetary Policy’ logically includes ‘Economic Policy’ as well as ‘Monetary Policy’. However, the decision to aggregate original PreLex descriptors into more encompassing policy labels was not always so clear-cut. Thus, the full correspondence table between original and new policy descriptors is given in Table A in the appendix. Besides the distribution across different categories, Table 2 also shows that information on the policy field is completely missing for about 11 percent of the documents. In addition, another 18 percent of the documents are assigned to two or more policy fields, making a clear-cut classification of those documents impossible. The ultimate goal of the analysis is to generate a classifier that correctly groups the documents with missing values and those assigned to multiple policy areas into the correct policy category.

Pre-processing of the data

The titles of the policy documents represent the raw data for the analysis. For conceptual as well as practical reasons, the raw data should not be used directly in the classification analysis. The raw titles are likely to include many uninformative words and additional symbols, such as punctuations, that increase the computational requirements of the analysis without improving the predictive ability of the classifier. As is common in computer-assisted text analysis, several pre-processing procedures were applied to prepare the data for the analysis.[3] First, all words in the titles were changed to lower-case letters and the titles were stripped of all punctuation, numbers, and additional whitespaces. In addition, a list of English stop words was used to exclude common but uninformative words such as ‘every’, ‘already’, ‘about’, or ‘he’, ‘she’, ‘it’. A case can also be made for reducing words to their common base form. This transformation is achieved through stemming. The commonly used Porter (1980) algorithm does this by removing suffixes from words. For example, applying the Porter stemmer to the words ‘connection’, ‘connections’, ‘connected’, and ‘connecting’ would reduce all of them to their common word stem ‘connect’. Whether stemming improves the classification results is not clear a priori. Thus, I experimented with both the raw words and a stemmed version of the data.

The classification analysis is an instance of the ‘bag of words’ approach, which ignores the ordering of words in the document and considers only their frequency (Manning *et al.* 2008: 107). The classification analysis relies on a word frequency matrix as input data, with

[3] Most of the pre-processing tasks were implemented through the ‘tm’ package (Feinerer 2009) in R (R Development Core Team 2009).

documents as rows and terms as columns. After removing all numbers, punctuation, and stop words, the dataset included 8002 stemmed and 10597 raw terms, resulting in a 27,104 x 8,002 word frequency matrix of stemmed terms and a 27,104 x 10,597 word frequency matrix of original terms. From these matrices, I removed 21 of the most frequent and uninformative words, such as 'council', 'proposal', and 'eec'.[4] In addition, I discarded all terms that occurred in less than 0.4 percent of the documents. Thus, a term had to be mentioned in at least 109 out of 27,104 document titles to be included in the analysis. While excluding infrequent terms is a common practice in automated text analysis, the precise threshold for excluding a term in the analysis was determined on practical grounds. Including more than 500 terms in the classification analysis was simply not feasible due to computer memory limitations. As a result of the removal of the infrequent terms, the size of the word frequency matrices decreased considerably. The frequency matrix of raw words shrank to a manageable 439 columns and the frequency matrix of stemmed words to 423 columns.

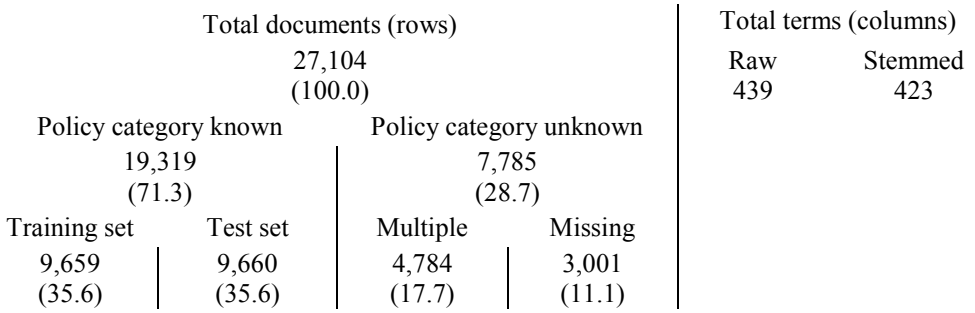
Another pre-processing choice regards the way individual terms are represented in the word frequency matrix. Besides using the actual frequencies of words, binary and weighted representations are also possible. The weighting of terms is used to distinguish between words with low and high discriminating power. Words that occur too frequently are not helpful in distinguishing the policy content of documents. Thus, a popular method multiplies term frequencies with their inverse document frequency. This so-called tf-idf weighting scheme results in high weights for terms that occur many times in a small number of documents and low weights for terms that occur few times in most documents (Manning *et al.* 2008: 109). In contrast, the binary representation only indicates the presence or absence of a word in a certain document. If, as in our case, a dataset includes many documents but few terms in each of them, not much information is lost by a binary representation. In fact, the absence or presence of a certain term might be more indicative of the policy content of a document than the absolute or weighted frequency of its occurrence in a relatively short document title. Again, no commonly agreed guidelines exist on which representation is more useful under which circumstances. Thus, I experiment with all three word representations.

[4] The following words were removed from the matrix: 'proposal', 'draft', 'commission', 'council', 'parliament', 'court', 'conseil', 'regulation', 'directive', 'decision', 'recommendation', 'communication', 'report', 'european', 'europe', 'community', 'communities', 'union', 'cee', 'eec', 'amending'.

Experimental setup and support vector machine classification

The analysis of this study consists of two steps. In a first step, I conduct experiments on the part of the dataset for which the policy category is known. The goal of this step in the analysis is to identify the term representation that yields the best prediction results and to quantify the overall prediction accuracy of the support vector machine classifier. Following the identification of the best-performing term representation and the training of the relevant support vector machine, I use the trained support vector machine in a second step to assign policy labels to documents in the part of the dataset for which the policy category is unknown or ambiguous. Figure 2 gives an overview of the characteristics of the data used in the two-step analysis.

Figure 2: Summary of data characteristics



Source: Own data generated from information in PreLex (http://ec.europa.eu/prelex/apcnet.cfm?CL=en).

The relevant policy category is known for about 71 percent of all documents contained in the dataset. This part of the dataset is used for the experiments. For that purpose, it is randomly divided into a training set and a test set of equal size. As the name indicates, the training set is used to train the support vector machine. Subsequently, the accuracy of the support vector machine classifier is examined by comparing its predictions to the actual known policy categories of the documents in the test set. In order to investigate which term representation performs best, I train the support vector machine on binary, frequency, and tf-idf weighted term frequency matrices of raw as well as stemmed words. The best-performing classifier identified through the experiments is then used to predict the policy category for the 29 percent of documents for which the policy category is unknown. Note that, in this context, 'unknown' refers to both documents that do not have any policy label as well as documents that have more than one policy label attached to them.

Many different methods have been proposed in the text analysis literature to classify documents. However, support vector machines have earned a strong reputation for comparatively high prediction accuracy (see e.g. Manning *et al.* 2008: 261, 307). Two-class

support vector machines separate the data by maximizing the distance between the classification boundary and adjacent data points (Manning *et al.* 2008: 294). In other words, support vector machines generate two classes by maximizing the margin between the data points in the two groups and the decision boundary. The data points at the edge of each class that define the position of the decision boundary are called support vectors, hence the name support vector machine. Once the classification function has been estimated by maximizing the margin between the two classes in the training set, it can be applied to predict class membership of new documents not part of the original training data.

The basic two-class support vector machine classification method can be extended to multi-class classification. Two common approaches are 'one-versus-all' and 'one-versus-one' classification (see Karatzoglou *et al.* 2006: 4; Manning *et al.* 2008: 303). According to the 'one-versus-all' method, one classifier is trained for each class to separate it from the rest of the data. The different classifiers are then compared and the one that produces the class with the greatest margin is used to label the data point. In contrast, according to the 'one-versus-one' method, one classifier is trained for each possible combination of pairs of classes. The prediction of class membership occurs then through voting, where the class that is selected by most classifiers wins. Although the 'one-versus-one' method requires the estimation of more support vector machines, it is often computationally more efficient than the 'one-versus-all' method as the training set for each individual support vector machine is considerably smaller. In the following analysis, I use the support vector machine command of the 'kernlab' package (Karatzoglou *et al.* 2004) in R (R Development Core Team 2009), which implements the 'one-versus-one' method for multi-class classification (Karatzoglou *et al.* 2006: 9).[5]

Results of the document classification experiments

To evaluate the effect of different term representations on the predictive accuracy of the support vector machine predictions, I varied the input data along two dimensions. First, words were either used in their raw form or were reduced to their word stems. Second, each word or word stem was represented by either a binary value indicating its presence or absence in the document, the frequency of its occurrence in the document, or its frequency weighted by its inverse document frequency. These combinations of pre-processing options

[5] More precisely, I estimate support vector machines with the 'ksvm' command and all options set to their default values. Thus, the estimation employs a Gaussian radial basis kernel function and the cost parameter is set to C = 1.

produced six different input matrices that formed the basis for the experimental analysis. The support vector machine was trained on the training set and then applied to predict the policy categories of the documents in the test set. The results of this procedure are presented in Table 3. The cell entries of the table provide the prediction accuracy of the support vector machine based on different combinations of data pre-processing options. The numbers represent the proportion of documents in the test set that were correctly classified.

Table 3: Classification accuracy with different term representations

	Stemming		
Representation	**No**	**Yes**	**Mean**
Binary	0.734	0.750	0.742
Frequency	0.730	0.744	0.737
tf-idf	0.730	0.745	0.737
Mean	0.732	0.746	

Notes: Table entries give the proportion of documents in the test set correctly predicted by the support vector machine, using different term representations as input data.

Interestingly, the results do not vary much as a consequence of different pre-processing decisions. The prediction results are rather robust, with all analyses pointing to a prediction accuracy of just below 75 percent. With a policy category scheme consisting of 21 categories, the possibility that a document is correctly classified by chance alone is extremely small.[6] Thus, predicting the policy category of about 75 percent of all documents correctly is a formidable achievement. Although the differences between the prediction success rates are generally quite small, two patterns are nevertheless observable when the individual and mean prediction rates across rows and columns of the table are compared. First, the binary term representation performs always better than the frequency and the weighted frequency representation. Second, analyses based on stemmed words consistently outperform those based on words in their raw form. The combination of word stems with binary value representation produces the best prediction results. Therefore, the remainder of the analysis focuses and relies on the support vector machine classifier based on this configuration of the input data.

[6] Assuming a uniform distribution over all policy categories, the expected chance agreement would be 21 x $(1/21)^2 = 0.05$.

Table 4: Classification accuracy for individual policy categories

Policy category	Total number of documents in category	Number of documents correctly classified	Proportion of documents correctly classified
Agriculture	2038	1900	0.93
Budget	1106	1042	0.94
Customs Union	852	702	0.82
External Relations	725	505	0.70
Commercial Policy	708	488	0.69
Internal Market	578	358	0.62
Fisheries	536	480	0.90
Transport and Telecommunications	438	242	0.55
Development Policy	389	251	0.65
Social Policy	345	221	0.64
Environment	301	193	0.64
General Affairs	289	149	0.52
Economic and Monetary Policy	274	146	0.53
Energy	248	163	0.66
Research	211	144	0.68
Financial Affairs	210	126	0.60
Education and Culture	114	46	0.40
Justice and Home Affairs	105	17	0.16
Regional Policy	81	30	0.37
Consumer Policy	65	38	0.58
Health	47	5	0.11
Total	9660	7246	0.75

Notes: The table shows the prediction results based on a binary document-term input matrix of stemmed words.

However, the overall proportion of documents correctly classified illustrates only part of the picture. Table 4 takes a more disaggregated view on the prediction accuracy of the best-performing classifier.[7] The table reveals a high variability in the prediction success rates across different policy fields. Although significant exceptions to this rule are observable, the classifier generally reaches higher prediction success rates for policy fields with a large number of documents and lower prediction success rates for policy categories with a small number of documents. For example, the prediction success rates for the two fields with the largest number of documents, Agriculture and Budget, are 93 and 94 percent, respectively. In contrast, the prediction accuracy for the policy field with the fewest number of documents,

[7] Table B in the appendix presents the entire contingency table. Besides reporting the proportion of documents of a certain category correctly predicted, it also takes the opposite perspective by looking at what proportion of documents predicted to be part of a certain category is actually part of that category.

Health, is a mere 11 percent. This bias of the classifier is problematic and needs to be taken into account when interpreting the prediction results.

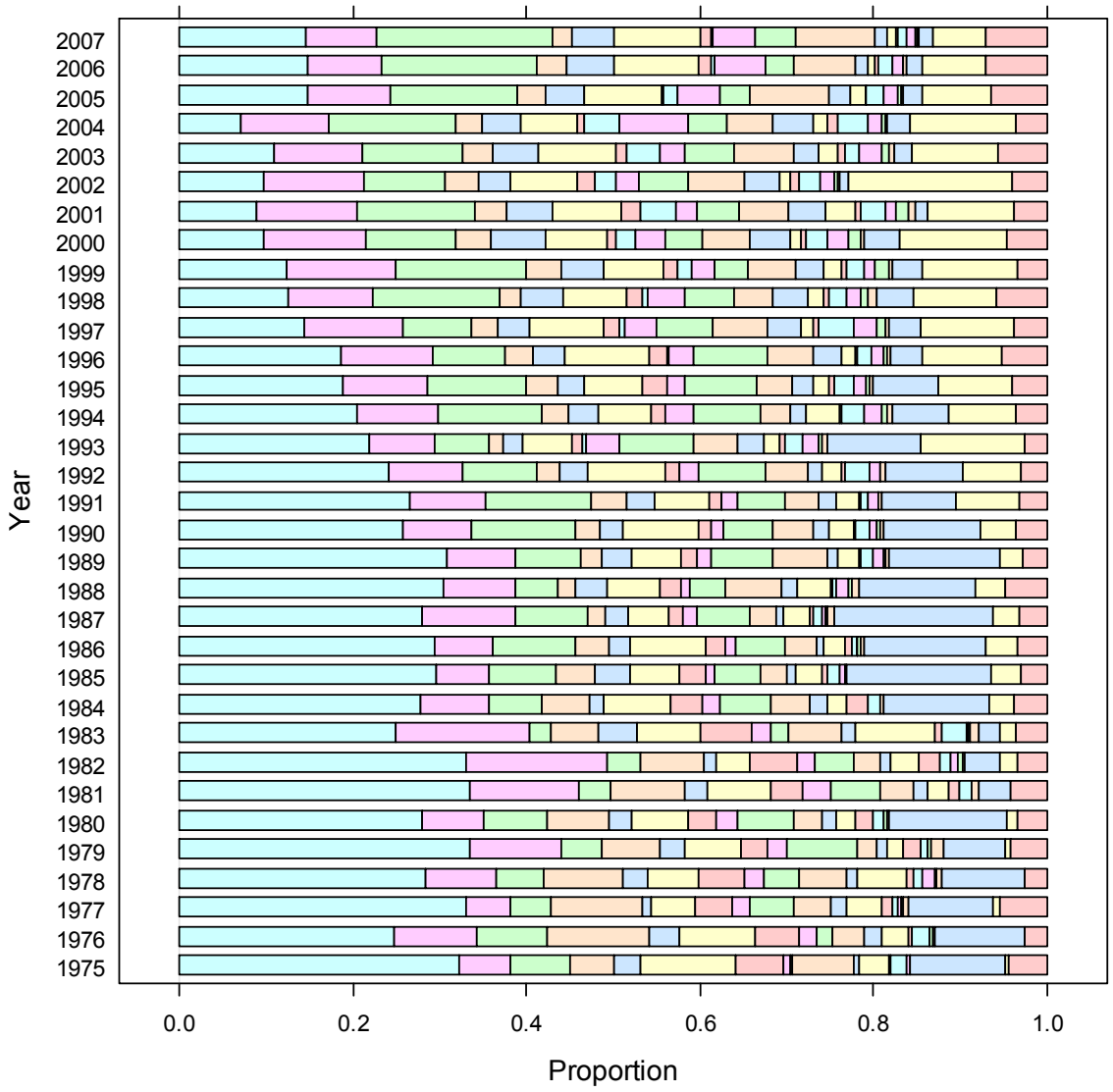
The Commission's agenda-setting activity over time and across policy areas

An overall success rate of predicting 75 percent of all documents correctly is substantially significant, but is the prediction accuracy high enough to justify the application of the support vector machine to generate policy labels for the documents for which labels are ambiguous or missing? If the entire dataset would need to be classified, the answer would probably be negative. However, more than 70 percent of all documents in the dataset are already classified. Only the remaining 30 percent of the documents need to be classified by the support vector machine. If the classifier's performance is similar to its performance on the test data and if the already existing classifications of 70 percent of the documents are correct, then applying the support vector machine to the remaining 30 percent will result in an overall misclassification rate of about 30 - (30 x 0.75) = 7.5 percent. While this error rate is still of considerable size, relying on the support vector machine classifier as a measurement instrument is arguably preferable to the alternative of not mapping the Commission's agenda-setting activity at all.

Figure 3 presents the proportion of Commission documents in different policy areas over time after the support vector machine has been used to classify documents with missing or ambiguous policy field membership.[8] The figure illustrates the strong variability of the Commission's agenda-setting activity not only across different policy fields but also over time. For example, in 1975, the first year covered by the dataset, the Commission's agenda-setting activity was still dominated by issues in Agriculture. Almost one third of all Commission documents were concerned with Agriculture, while no documents at all were published in areas such as Justice and Home Affairs, Consumer Protection, or Health. Over time, the Commission's agenda-setting activity became less concentrated. In 2007, Agriculture accounted only for about 15 percent of the Commission's documents and was surpassed by External Relations with 20 percent of the documents. The proportion of documents increased in other areas as well. The fields of Environment, Economic and Monetary Policy, Transport and Telecommunications, and Social Policy all registered growth in the proportion of Commission documents. These changes over time are easier to identify in Figure 4, which provides the same information in a different format.

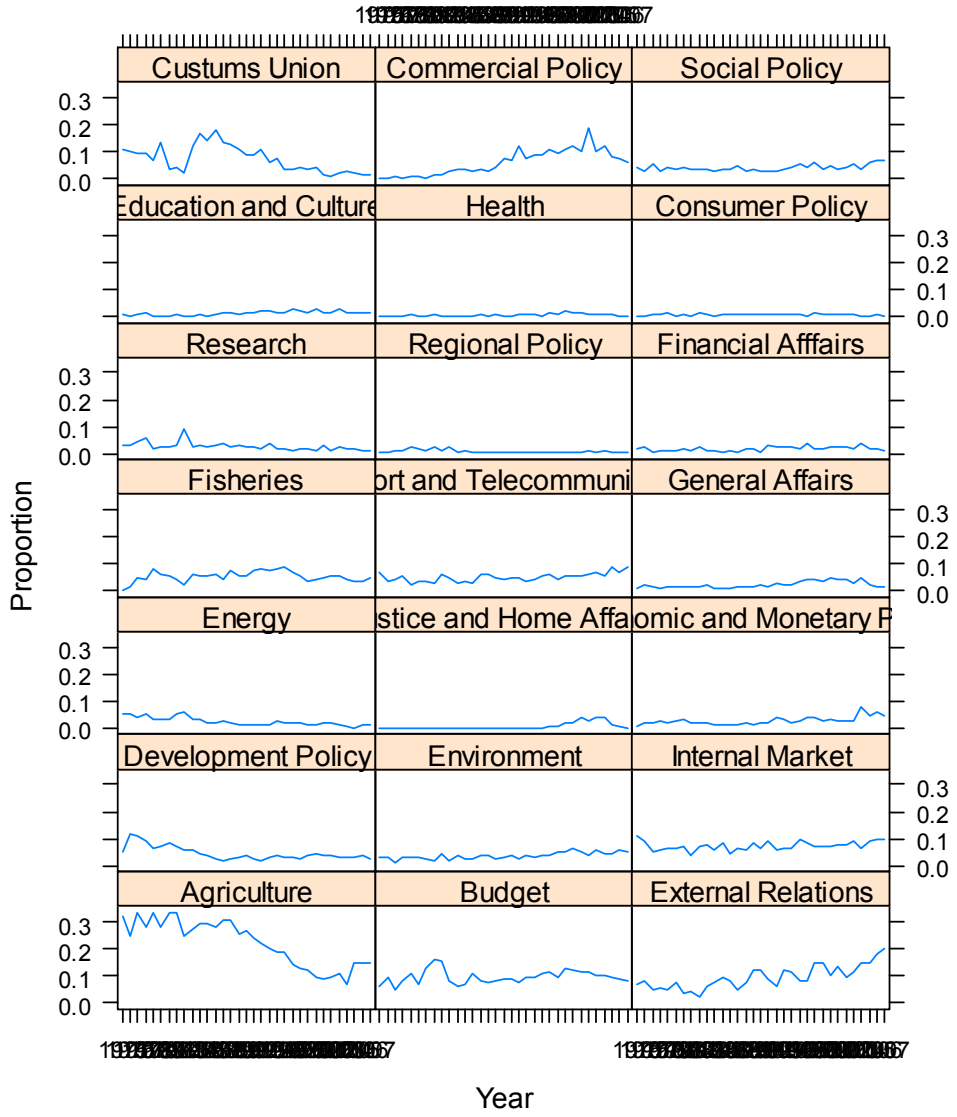
[8] See also Tables C and D in the appendix, which present the relative and absolute document frequencies in different policy areas over time.

Figure 3: Composition of Commission agenda-setting activity, 1975-2007



Notes: Policy areas included (from left to right): Agriculture, Budget, External Relations, Development Policy, Environment, Internal Market, Energy, Justice and Home Affairs, Economic and Monetary Policy, Fisheries, Transport and Telecommunications, General Affairs, Research, Regional Policy, Financial Affairs, Education and Culture, Health, Consumer Policy, Customs Union, Commercial Policy, Social Policy.

Figure 4: Proportion of Commission documents in different policy areas, 1975-2007



A full and detailed descriptive analysis of the Commission's agenda-setting activity is beyond the scope of this paper. However, it is worth noting that the preliminary overview given here is consistent with generally held views about the historical development of the EU's policy competences in general and the Commission's activities in particular: The role of Agriculture in the EU and within the Commission itself has continuously declined since the first reforms were started in that sector in the early 1990s. In contrast, the EU's growing ambitions and actual activities as an actor on the global stage are reflected in a steadily increasing document rate in the area of External Relations. Similarly, the growth in the Commission's agenda-setting activity in areas such as Environment or Economic and Monetary Policy is an indication of the commonly acknowledged extension of the EU's competences to new policy

areas over time (see e.g. Alesina *et al.* 2005). In other words, these observations indicate an acceptable level of face validity of the policy indicator generated with the help of the support vector machine classifier and thereby strengthen our trust in the overall quality of the measure.

Summary and conclusions

To be able to judge competing claims about the role of the Commission in the process of European integration, we need time- and policy-specific measures of the Commission's agenda-setting activities. One major way in which the Commission's agenda-setting activity manifests itself is through written text in policy documents. Thus, such documents form an invaluable source for tracking the Commission's efforts to advance policy in a certain area. The online database PreLex includes information on policy documents published by the Commission since 1975. While the database includes descriptors about the 'field of activity' to which a certain document belongs, several policy labels are often attached to a single document. Also, policy labels are completely missing for a considerable number of policy documents. In order to map the Commission's agenda-setting activity, we need a policy categorization scheme that is both comprehensive in scope and able to allocate documents into mutually exclusive categories. In this study, I investigated the usefulness of computer-assisted document classification methods for generating policy labels for documents in a way that is in line with these principles for policy categorization schemes.

The data for the classification analysis was extracted from the Commission's PreLex database. Each webpage in the database describes a policy document. I downloaded all webpages and extracted the titles of the documents as well as the 'field of activity' description. This procedure resulted in a dataset of 27,000 policy documents introduced by the Commission between 1975 and 2007. About 70 percent of the documents in this dataset included unambiguous policy descriptors. Thus, I used this part of the dataset to train and test the performance of the document classifier. The part of the dataset for which the policy category is known was randomly split into a test and a training set of equal size. The training set was used to train a support vector machine classifier, one of the most popular and best-performing methods currently used in automated document classification. I trained the support vector machine on several versions of the input data to identify the pre-processing strategy that yields the best results in terms of prediction accuracy. The prediction results were remarkably robust. Regardless of whether words entered the analysis in their raw form or were reduced to their word stems, the support vector machine always reached a prediction

accuracy of just below 75 percent. Similarly, how the individual terms were represented in the document-term input matrix for the analysis did hardly matter. Terms were represented in three different ways: First, by a binary value for their absence or presence in a document; second, by the unweighted frequency of their occurrence in a document; and third, by term frequencies weighted by their inverse document frequencies. In all instances, the prediction results varied only slightly.

Although the differences in the prediction success rate were very small, the results based on stemmed words and those based on binary word representations performed consistently better than the alternatives. Thus, I used the support vector machine trained on data based on this combination of pre-processing options to predict the policy category of documents for which policy labels were missing or ambiguous. Given that the policy category had to be estimated for only about 30 percent of the documents in the dataset, the overall misclassification rate is expected to be around 7.5 percent. A preliminary descriptive analysis of the Commission's agenda-setting activity showed several interesting developments that are generally in line with the received wisdom about the changes in the focus and the extent of the Commission's agenda-setting activity. The Commission's agenda-setting activity diversified considerably over time. While Agriculture dominated for a long time, the proportion of documents published in this area decreased steadily after the first reforms were introduced in the early 1990s. External Relations has seen the biggest growth in the proportion of documents introduced by the Commission, confirming the view that the EU is playing an increasingly active political role on the world stage. Also, the continuous extension of EU policy competences is reflected in the growth of documents issued in areas such as Environment, Justice and Home Affairs, and Transport and Telecommunications. This brief overview indicated an acceptable level of face validity of the policy category indicator generated with the help of the support vector machine classifier.

Still, further improvements are possible and should be investigated in future research. First, the current analysis treated documents to which more than one policy label was assigned the same way as documents for which labels were completely missing. For documents with more than one policy label, research should investigate the possibility of limiting the policy options in the classification analysis to one of the policy areas mentioned in the 'field of activity' descriptor. This restriction of the label options would rule out complete misclassifications of those documents and thus improve the prediction accuracy rate of the automated classification. Second, the incorporation of additional information, for example about the author of the document, could boost the proportion of documents correctly

classified. PreLex often includes a descriptor about the Commission Directorate General that was primarily responsible for drafting the document in question. This information could be highly indicative of the policy field. Finally, this study only examined the performance of support vector machines as classifiers. While existing experimental results show that support vector machines tend to perform better than other classifiers, this is not necessarily the case for each and every dataset. Also, the usefulness of combining the results of different classifiers for ensemble predictions (see e.g. Hillard *et al.* 2008) should be examined in future research.

References

Alesina, Alberto, Ignazio Angelino and Ludger Schuknecht (2005): What Does the European Union Do? *Public Choice* 123: 275-319.

Crombez, Christophe (1996): Legislative Procedures in the European Community. *British Journal of Political Science* 26(2): 199-228.

Daviter, Falk (2007): Policy Framing in the European Union. *Journal of European Public Policy* 14(4): 654-66.

Diermeier, Daniel , Jean-François Godbout, Bei Yu and Stefan Kaufmann (2006): Language and Ideology in Congress. *Unpublished working paper*.

Feinerer, Ingo (2009): Tm: Text Mining Package, R Package Version 0.4.

Gornitzka, Åse and Ulf Sverdrup (2008): Who Consults? The Configuration of Expert Groups in the European Union. *West European Politics* 31(4): 725 - 50.

Hillard, Dustin, Stephen Purpura and John Wilkerson (2008): Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics* 4(4): 31 - 46.

Høyland, Bjørn and Jean-François Godbout (2009): Debates and Votes in the European Parliament. *Unpublished working paper*.

Karatzoglou, Alexandros, David Meyer and Kurt Hornik (2006): Support Vector Machines in R. *Journal of Statistical Software* 15(9): 1-28.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik and Achim Zeileis (2004): Kernlab: An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9): 1-20.

Laver, Michael, Kenneth Benoit and John Garry (2003): Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review* 97(02): 311-31.

List, Christian (2004): A Model of Path-Dependence in Decisions over Multiple Propositions. *American Political Science Review* 98(3): 495-513.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008): *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Moravcsik, Andrew (1998): *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*. Ithaka: Cornell University Press.

Pennings, Paul and Hans Keman (2002): Towards a New Methodology of Estimating Party Policy Positions. *Quality and Quantity* 36(1): 55-79.

Porter, M. F. (1980): An Algorithm for Suffix Stripping. *Program* 14(3): 130-7.

Proksch, Sven-Oliver and Jonathan B. Slapin (2009): Position Taking in European Parliament Speeches. *Unpublished working paper*.

Purpura, Stephen and Dustin Hillard (2006): Automated Classification of Congressional Legislation. *Proceedings of the 2006 International Conference on Digital Government Research*. San Diego: ACM.

Quinny, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radevk (2008): How to Analyze Political Attention with Minimal Assumptions and Costs. *Unpublished working paper*.

R Development Core Team (2009): R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Shepsle, Kenneth A. and Mark S. Bonchek (1997): *Analyzing Politics: Rationality, Behavior, and Institutions*. New York: W. W. Norton.

Slapin, Jonathan B. and Sven-Oliver Proksch (2008): A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science* 52(3): 705-22.

Steunenberg, Bernard (1994): Decision Making under Different Institutional Arrangements: Legislation by the European Community. *Journal of Institutional and Theoretical Economics* 150(4): 642-69.

Stone Sweet, Alec and Thomas L. Brunell (1998): Constructing a Supranational Constitution: Dispute Resolution and Governance in the European Community. *American Political Science Review* 92(1): 63-81.

Tsebelis, George and Geoffrey Garrett (2001): The Institutional Foundations of Intergovernmentalism and Supranationalism in the European Union. *International Organization* 55(2): 357-90.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier (2008): Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics* 5(1): 33 - 48.

Appendix

Table A: Recoding of PreLex policy field descriptions

New label	Original label in PreLex
Agriculture	Agriculture
Budget	Budget
External Relations	Bilateral agreements Common, foreign and security policy External relations Multilateral relations
Development Policy	Development policy
Environment	Environment
Internal Market	Competition policy Industrial policy Internal market Intellectual property law Law relating to undertakings Public contracts Right of establishment
Energy	Energy
Justice and Home Affairs	Justice and home affairs Justice, freedom and security
Economic and Monetary Policy	Economic and monetary policy Economic policy Monetary policy
Fisheries	Fisheries
Transport and Telecommunications	Trans-European networks Transport policy Dissemination of information
General Affairs	Administration European citizenship General, financial and institutional matters Provisions governing the institutions
Research	Science and research
Regional Policy	Regional policy
Financial Affairs	Company law Free movement of capital Taxation
Education and Culture	Culture Education and training Science, information, education and culture Sport
Health	Health protection
Consumer Policy	Consumers
Customs Union	Customs union
Commercial Policy	Commercial policy
Social Policy	Freedom of movement for workers

Table B: Agreement between PreLex policy classifications and support vector machine predictions

											Predicted												
True	**Agri**	**Bud**	**Ext**	**Dev**	**Envi**	**Int**	**Ener**	**JHA**	**EMP**	**Fish**	**Tel**	**Gen**	**Res**	**Reg**	**Fin**	**Edu**	**Heal**	**Cons**	**Cust**	**Com**	**Soc**	**Total**	**R**
Agri	1900	11	29	7	13	19	1	0	2	4	7	4	4	0	1	0	0	3	20	5	8	2038	0.93
Bud	29	1042	3	4	2	4	0	0	4	0	4	7	1	1	0	2	0	0	0	0	3	1106	0.94
Ext	67	2	505	21	5	19	0	0	8	1	11	12	3	0	0	0	0	0	23	39	9	725	0.70
Dev	39	3	58	251	1	12	0	0	2	0	6	3	2	0	0	0	0	0	6	5	1	389	0.65
Envi	34	3	10	7	193	27	1	0	1	0	9	2	3	0	0	0	2	3	1	1	4	301	0.64
Int	98	5	22	9	10	358	2	0	2	0	30	3	8	0	0	1	0	1	5	2	22	578	0.62
Ener	30	3	10	1	5	14	163	0	1	0	4	3	8	0	0	0	0	1	2	1	2	248	0.66
JHA	30	1	13	2	2	18	0	17	1	0	10	2	0	0	0	0	0	0	0	1	8	105	0.16
EMP	33	11	28	2	0	25	1	0	146	1	10	2	0	0	0	0	0	1	5	0	9	274	0.53
Fish	28	0	7	0	2	7	0	0	0	480	1	0	2	0	0	0	0	0	5	2	2	536	0.90
Tel	48	7	19	9	11	51	0	1	6	1	242	2	9	0	0	0	0	0	5	1	26	438	0.55
Gen	56	6	5	1	2	36	3	0	10	0	5	149	0	0	0	1	0	0	1	0	14	289	0.52
Res	12	3	11	2	8	11	1	0	3	0	8	1	144	0	0	0	0	0	0	0	7	211	0.68
Reg	13	7	4	5	0	12	1	0	4	0	1	2	0	30	0	0	0	0	0	0	2	81	0.37
Fin	37	2	11	1	1	5	1	0	1	0	1	1	0	0	126	1	0	0	12	1	9	210	0.60
Edu	17	1	3	0	0	8	0	0	6	0	4	5	2	0	0	46	0	0	0	0	22	114	0.40
Heal	11	0	1	1	0	5	0	0	0	1	3	0	0	0	0	1	5	0	0	0	19	47	0.11
Cons	4	0	0	0	4	13	0	0	0	0	4	0	0	0	0	0	0	38	0	0	2	65	0.58
Cust	57	0	34	7	4	14	1	0	2	0	6	0	0	0	4	0	0	0	702	19	2	852	0.82
Com	66	0	112	0	1	11	0	0	0	3	4	0	0	0	0	0	0	0	22	488	1	708	0.69
Soc	41	5	7	1	3	24	2	0	4	1	10	6	3	0	0	10	7	0	0	0	221	345	0.64
Total	2650	1112	892	331	267	693	177	18	203	492	380	204	189	31	131	62	14	47	809	565	393	9660	
P	0.72	0.94	0.57	0.76	0.72	0.52	0.92	0.94	0.72	0.98	0.64	0.73	0.76	0.97	0.96	0.74	0.36	0.81	0.87	0.86	0.56		0.75

Notes: R = Recall = Prediction accuracy: Percentage of proposals of category X predicted to be members of X (column percentage of diagonal cell entry), P = Precision: Percentage of proposals predicted to be members of X that are actually members of X (row percentage of diagonal cell entry). Agri = Agriculture, Bud = Budget, Ext = External Relations, Dev = Development Policy, Envi = Environment, Int = Internal Market, Ener = Energy, JHA = Justice and Home Affairs, EMP = Economic and Monetary Policy, Fish = Fisheries, Tel = Transport and Telecommunications, Gen = General Affairs, Res = Research, Reg = Regional Policy, Fin = Financial Affairs, Edu = Education and Culture, Heal = Health, Cons = Consumer Policy, Cust = Customs Union, Com = Commercial Policy, Soc = Social Policy.

Table C: Relative frequency of Commission documents in different policy fields, 1975-2007

Year	Agri	Bud	Ext	Dev	Envi	Int	Ener	JHA	EMP	Fish	Tel	Gen	Res	Reg	Fin	Edu	Heal	Cons	Cust	Com	Soc
1975	32.2	5.9	7.0	5.1	2.9	11.0	5.5	0.0	0.7	0.4	7.0	0.7	3.3	0.4	1.8	0.4	0.0	0.0	11.0	0.4	4.4
1976	24.7	9.6	8.1	11.8	3.3	8.7	5.1	0.0	2.1	1.8	3.6	2.1	3.0	0.3	2.1	0.0	0.3	0.3	10.2	0.0	2.7
1977	33.1	□.0	4.6	10.7	1.0	5.0	4.2	0.0	2.1	5.2	4.2	1.7	4.2	1.2	0.6	0.4	0.2	0.6	9.8	0.8	5.6
1978	28.3	8.1	5.5	9.3	2.9	5.8	5.2	0.0	2.3	4.1	5.5	1.2	5.8	0.9	0.9	1.5	0.3	0.6	9.5	0.0	2.6
1979	33.4	10.7	4.6	6.7	2.9	6.5	3.1	0.0	2.1	8.1	2.3	1.3	1.7	2.1	0.8	0.0	0.4	1.5	7.1	0.6	4.2
1980	27.9	7.1	7.4	7.1	2.8	6.5	3.2	0.0	2.4	6.5	3.4	1.6	2.2	1.9	1.3	0.0	0.3	0.3	13.7	1.2	3.4
1981	33.5	12.6	3.6	8.6	2.6	7.4	3.6	0.0	3.3	5.7	3.8	1.7	2.4	1.2	1.4	0.0	0.0	1.0	3.6	0.0	4.3
1982	33.1	16.2	3.9	7.3	1.5	3.7	5.6	0.0	2.1	4.4	3.1	1.2	3.3	2.3	1.2	0.8	0.6	0.2	4.1	1.9	3.5
1983	24.9	15.5	2.5	5.4	4.4	7.4	5.9	0.0	2.2	2.0	6.2	1.5	9.1	1.0	2.7	0.3	0.3	1.0	2.5	1.7	3.7
1984	27.8	8.0	6.0	5.5	1.6	7.7	3.6	0.0	2.1	5.9	4.6	2.1	2.2	2.5	1.4	0.0	0.0	0.4	12.2	2.8	3.8
1985	29.7	6.1	7.7	4.5	4.0	5.7	3.0	0.0	1.0	5.3	3.2	1.0	3.1	0.6	1.3	0.7	0.0	0.2	16.6	3.4	3.1
1986	29.4	6.8	9.5	3.9	2.5	8.6	2.2	0.0	1.3	5.8	3.7	0.7	2.4	0.9	0.6	0.1	0.3	0.4	14.1	3.7	3.4
1987	28.1	10.7	8.2	2.2	2.5	4.6	1.7	0.0	1.6	6.2	3.0	0.8	3.1	0.4	1.1	0.3	0.1	0.9	18.2	3.0	3.3
1988	30.5	8.2	4.9	2.0	3.7	6.2	2.4	0.0	1.1	4.0	6.5	1.7	3.9	0.3	0.4	1.4	0.4	0.8	13.3	3.5	4.8
1989	30.8	7.8	7.7	2.3	3.5	5.7	1.8	0.0	1.7	7.2	6.2	1.3	2.4	0.1	1.5	1.2	0.3	0.4	12.7	2.7	2.8
1990	25.7	7.9	12.0	2.9	2.7	8.7	1.4	0.0	1.3	5.8	4.7	1.7	2.8	0.3	1.6	0.7	0.4	0.6	11.1	4.0	3.6
1991	26.5	8.7	12.3	4.0	3.2	6.4	1.4	0.1	1.8	5.4	4.0	2.1	2.5	0.3	0.7	1.3	0.1	0.3	8.5	7.3	3.2
1992	24.1	8.5	8.6	2.5	3.4	8.9	1.6	0.0	2.2	7.8	4.8	1.8	2.2	0.3	2.9	1.1	0.1	0.5	8.9	6.8	3.0
1993	21.8	7.5	6.3	1.8	2.1	5.7	1.3	0.2	4.0	8.4	5.0	3.2	1.8	0.6	2.1	1.8	0.4	0.7	10.8	11.8	2.7
1994	20.4	9.4	11.9	3.1	3.5	6.1	1.6	0.0	3.1	7.8	3.5	1.9	3.8	0.2	2.7	2.0	0.5	0.7	6.5	7.8	3.6
1995	18.9	9.8	11.4	3.5	3.2	6.6	2.9	0.0	2.0	8.2	4.1	2.5	1.9	0.6	2.1	1.4	0.5	0.4	7.5	8.5	4.1
1996	18.6	10.6	8.3	3.2	3.7	9.8	2.0	0.2	2.8	8.6	5.3	3.2	1.6	0.3	1.6	1.5	0.3	0.4	3.7	9.1	5.3
1997	14.3	11.4	8.1	2.9	3.7	8.5	1.9	0.6	3.7	6.6	6.3	3.9	1.4	0.5	4.1	2.6	1.2	0.3	3.7	10.8	3.9
1998	12.6	9.8	14.6	2.4	5.0	7.3	1.7	0.8	4.1	5.8	4.4	4.0	1.8	0.6	2.1	1.8	0.8	1.0	4.2	9.5	5.9
1999	12.4	12.5	15.0	4.0	4.9	7.1	1.5	1.7	2.6	3.7	5.5	3.3	2.1	0.7	2.0	1.2	1.7	0.3	3.5	10.9	3.5
2000	9.6	11.9	10.2	4.2	6.2	7.3	1.0	2.2	3.5	4.2	5.5	4.7	1.1	0.6	2.4	2.6	1.4	0.5	4.0	12.4	4.6
2001	8.9	11.6	13.5	3.7	5.3	7.9	2.2	4.1	2.4	4.9	5.6	4.4	3.3	0.7	2.8	1.4	1.3	0.8	1.6	9.9	3.8
2002	9.7	11.6	9.3	3.9	3.7	7.7	1.9	2.5	2.7	5.8	6.4	4.1	1.2	1.1	2.4	1.6	0.4	0.3	0.9	19.0	4.0
2003	11.0	10.0	11.7	3.3	5.4	9.0	1.3	3.9	2.7	5.7	6.9	2.9	2.3	0.6	1.7	2.7	0.8	0.5	2.2	9.9	5.6
2004	7.0	10.2	14.6	3.1	4.4	6.6	0.7	4.1	8.0	4.4	5.3	4.6	1.7	1.2	3.5	1.5	0.5	0.2	2.6	12.2	3.7
2005	14.7	9.5	14.7	3.3	4.5	8.8	0.3	1.6	4.8	3.6	9.0	2.4	1.8	0.1	2.0	1.5	0.4	0.3	2.1	7.9	6.5
2006	14.7	8.6	17.8	3.5	5.4	9.8	1.3	0.4	5.9	3.4	7.1	1.3	0.9	0.3	1.8	1.0	0.1	0.5	1.8	7.3	7.0
2007	14.6	8.1	20.2	2.3	4.9	9.9	1.2	0.3	4.9	4.8	9.1	1.3	1.1	0.1	1.1	1.0	0.1	0.3	1.7	6.0	7.1
Mean	20.5	9.5	10.3	3.9	3.7	7.4	2.1	0.9	2.9	5.6	5.2	2.4	2.4	0.7	1.8	1.2	0.5	0.5	7.1	7.0	4.3

Table D: Absolute frequency of Commission documents in different policy fields, 1975-2007

Year	Agri	Bud	Ext	Dev	Envi	Int	Ener	JHA	EMP	Fish	Tel	Gen	Res	Reg	Fin	Edu	Heal	Cons	Cust	Com	Soc	Total
1975	88	16	19	14	8	30	15	0	2	1	19	2	9	1	5	1	0	0	30	1	12	273
1976	82	32	27	39	11	29	17	0	7	6	12	7	10	1	7	0	1	1	34	0	9	332
1977	173	26	24	56	5	26	22	0	11	27	22	9	22	6	3	2	1	3	51	4	29	522
1978	98	28	19	32	10	20	18	0	8	14	19	4	20	3	3	5	1	2	33	0	9	346
1979	160	51	22	32	14	31	15	0	10	39	11	6	8	10	4	0	2	7	34	3	20	479
1980	190	48	50	48	19	44	22	0	16	44	23	11	15	13	9	0	2	2	93	8	23	680
1981	141	53	15	36	11	31	15	0	14	24	16	7	10	5	6	0	0	4	15	0	18	421
1982	160	78	19	35	7	18	27	0	10	21	15	6	16	11	6	4	3	1	20	9	17	483
1983	101	63	10	22	18	30	24	0	9	8	25	6	37	4	11	1	1	4	10	7	15	406
1984	256	74	55	51	15	71	33	0	19	54	42	19	20	23	13	0	0	4	112	26	35	922
1985	270	55	70	41	36	52	27	0	9	48	29	9	28	5	12	6	0	2	151	31	28	909
1986	296	68	96	39	25	87	22	0	13	58	37	7	24	9	6	1	3	4	142	37	34	1,008
1987	254	97	74	20	23	42	15	0	14	56	27	7	28	4	10	3	1	8	165	27	30	905
1988	290	78	47	19	35	59	23	0	10	38	62	16	37	3	4	13	4	8	127	33	46	952
1989	240	61	60	18	27	44	14	0	13	56	48	10	19	1	12	9	2	3	99	21	22	779
1990	255	78	119	29	27	86	14	0	13	57	47	17	28	3	16	7	4	6	110	40	36	992
1991	233	76	108	35	28	56	12	1	16	47	35	18	22	3	6	11	1	3	75	64	28	878
1992	232	82	83	24	33	86	15	0	21	75	46	17	21	3	28	11	1	5	86	65	29	963
1993	186	64	54	15	18	49	11	2	34	72	43	27	15	5	18	15	3	6	92	101	23	853
1994	183	84	107	28	31	55	14	0	28	70	31	17	34	2	24	18	4	6	58	70	32	896
1995	177	92	107	33	30	62	27	0	19	77	38	23	18	6	20	13	5	4	70	80	38	939
1996	178	102	80	31	35	94	19	2	27	82	51	31	15	3	15	14	3	4	35	87	51	959
1997	137	109	77	28	35	81	18	6	35	63	60	37	13	5	39	25	11	3	35	103	37	957
1998	134	104	156	25	53	78	18	8	44	62	47	43	19	6	22	19	8	11	45	101	63	1,066
1999	114	115	138	37	45	65	14	16	24	34	51	30	19	6	18	11	16	3	32	100	32	920
2000	105	130	111	46	67	79	11	24	38	46	60	51	12	7	26	28	15	5	44	135	50	1,090
2001	91	118	138	38	54	81	22	42	24	50	57	45	34	7	28	14	13	8	16	101	39	1,020
2002	91	108	87	36	35	72	18	23	25	54	60	38	11	10	22	15	4	3	8	178	37	935
2003	106	96	112	32	52	86	12	37	26	55	66	28	22	6	16	26	8	5	21	95	54	961
2004	81	117	168	36	50	76	8	47	92	51	61	53	19	14	40	17	6	2	30	140	42	1,150
2005	145	94	145	32	44	87	3	16	47	35	89	24	18	1	20	15	4	3	21	78	64	985
2006	155	91	188	37	57	103	14	4	62	36	75	14	9	3	19	11	1	5	19	77	74	1,054
2007	156	87	216	24	52	106	13	3	52	51	97	14	12	1	12	11	1	3	18	64	76	1,069
Total	5,558	2,575	2,801	1,068	1,010	2,016	572	231	792	1,511	1,421	653	644	190	500	326	129	138	1,931	1,886	1,152	27,104