2010-01-01

The Extent of Clientelism in Irish Politics: Evidence from Classifying Dáil Questions on a Local-National Dimension

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Richard Sinnott
University College Du

Niall O'Reilly
Technological University Dublin

The Extent of Clientelism in Irish Politics: Evidence from Classifying Dáil Questions on a Local-National Dimension

Sarah Jane Delany[1], Richard Sinnott[2], and Niall O'Reilly[3]

[1] Digital Media Centre, Dublin Institute of Technology,
sarahjane.delany@dit.ie
[2] Geary Institute, University College Dublin,
richard.sinnott@ucd.ie
[3] School of Computing, Dublin Institute of Technology

Abstract. The availability of the full text of Irish parliamentary questions offers opportunities for using machine learning techniques to examine the currently much discussed role of elected representatives (TDs) in the Irish parliamentary system. Bluntly, are TDs mainly national legislators or "constituency messenger boys"? This paper presents an initial investigation into the use of automated text classification techniques to categorise parliamentary questions from 1922 up to 2008 as national or local. The approach uses a bag of words representation, standard feature reduction methods and an SVM classifier. Initial results show there is very little evidence in the corpus of parliamentary questions in Dáil Éireann to support the view that the role of the TD is determined by mainly clientelist/parochial imperatives.

1 Introduction

There has been a long running debate about the role of the elected representative (Teachta Dála or TD) in the Irish parliamentary system. Given the economic and political crisis in which the country finds itself, it is not surprising that this debate is currently a matter of some intensity. The debate centres around the extent to which TDs are or should be predominantly national policy makers whose primary task is to legislate and to hold the government accountable, as opposed to being primarily local representatives looking after the needs of their constituencies. The issue has substantial implications for the related debate about reforming the electoral system of Proportional Representation by means of the Single Transferrable Vote (PR-STV), which is seen by many as the main cause of what is considered to be the excessively localist orientation of TDs.

The debate about what TDs actually do relies heavily on anecdotal evidence from TDs themselves and from journalists and other observers. The purpose of the present paper is to use supervised machine learning techniques to exploit the textual record of parliamentary questions as a more systematic source of data on the local versus national orientation of TDs. This will be done by applying automated text categorisation techniques to the text of questions raised in Dáil Éireann (the Irish Parliament) in order to classify them as either local or national questions. Previous research on parliamentary questions in Ireland [1, 2] has relied on manual content analysis of the questions. The disadvantage of this approach is that it limits the potential coverage of the inquiry, particularly over the all-important temporal dimension.

The remainder of this paper is organised as follows. Section 2 gives an overview of related work which uses automated techniques for text analysis and topic classification to address problems in the political science arena. Section 3 discusses the process of asking questions in Dáil Éireann and describes the information available on Dáil questions. Section 4 outlines the methodology used in the application of supervised learning techniques to the problem of classifying Dáil questions. The results and implications of this categorisation are discussed in Section 5 with conclusions and opportunities for future work outlined in Section 6.

2 Related Work

In the political science arena, the analysis and classification of parliamentary questions has been performed using manual content analysis [1–5] but there is little evidence of automated techniques being used. One example is the work done by Abdul-Nasir [6] who proposes extracting rules from historical parliamentary questions posed to the National Assembly of Pakistan to help with the task of identifying legal questions using admissibility rules.

Automated techniques for text analysis and topic classification have recently become popular in political science research [7–9]. Research in this area covers a variety of applications of statistical analysis and machine learning approaches including, for example, discovering patterns of lexical cohesion in political speech [10]; selecting words that capture partisan or other differences in political speech [11]; classifying party affiliation from political speech [12]; recognising citation sentences in electronically submitted public comments [13]. There has also been considerable interest in identifying sentiment and opinion from political speech [14–16].

The importance of supervised learning techniques to the analyst in political science is further evidenced in the Comparative Agendas Project (CAP)[4]. CAP brings together scholars across different countries who are collaborating on developing systematic indicators of issue attention within their countries' political systems. Comprehensive databases consisting of over 1.5 million records of government activities (e.g., laws, bills, parliamentary questions, prime ministerial speeches) over several decades have been collated. All are linked by a common policy topic classification system which allows new types of analyses of public policy dynamics over time. Supervised learning algorithms have been developed

[4] http://www.comparativeagendas.org/

and made available to the project to assist with the annotation, update and analysis of this data[5].

3 Irish Parliamentary Questions

As a prominent part of Dáil procedure, Question Time gives TDs the right to question ministers on any matter relating to public affairs or to administration that lies within their area of responsibility. According to the standing orders of the Dáil, the purpose of parliamentary questions is to elucidate matters of fact or of policy. Once the answer has been given by the minister, the deputy who submitted the question and other deputies have the right to ask supplementary questions for the further elucidation of the information requested.

Deputies must give a number of days notice of their intention to ask a question and must submit the text of the question to the Question Office where each question is checked for compatibility with standing orders. Questions can be put down for either oral or written answer. Currently questions are answered during a time slot that varies from one hour fifteen minutes to two hours depending on whether it is a Tuesday, Wednesday or Thursday. If a question is not reached within the time allocated for that day it can receive a written answer or can, at the request of the TD who put down the original question, be held over.

The foregoing brief account of the question procedure applies generally to Question Time as it has evolved since its inception in 1922. Minor and not so minor changes have been made from time to time, with a major set of reforms having been implemented in late 1986. The main 1986 changes were the introduction of two new categories of question (questions to the Taoiseach and "priority" questions asked by a member of a group of opposition deputies), the establishment of a daily rota to determine when individual ministers are to be available to answer questions and the use of a lottery to determine the sequence of questions (other than questions to the Taoiseach or priority questions) on the Dail Order Paper.

Overall, the parlimentary question process generates a large volume of text. The text of all Dáil proceedings, including parliamentary questions, is publicly available online[6]. This availability of the text of all questions and replies provides a substantial body of data on the parliamentary role and behaviour of Irish elected representatives that is highly relevant to contemporary debate about institutional reform.

4 Methodology

This section outlines the methodology used to gather the text of the Dáil questions and determine the most appropriate document representation. It also reports on the preliminary experimentation undertaken.

[5] http://www.comparativeagendas.org/text-tools

[6] http://www.oireachtas.ie/viewdoc.asp?fn=/documents/nav/debates.htm&CatID=50&m=d

4.1 Datasets

The text of the questions, responses and supplementary questions for all parliamentary questions from September 1922 to the end of December 2008 was retrieved from the Dáil Debates website[7] to form the corpus. Sometimes questions relate to a particular individual and are anonymised – the name is supplied to the Minister or the Department concerned but not given in the Dáil proceedings. These questions are categorised as 'personal' questions and were removed from the corpus. They can be considered to have local orientation but further analysis will be required to confirm this.

Some examples of questions are listed below:

> *Dr. O'Connell asked the Taoiseach the average wage of industrial workers.*
>
> *Mr. Spring asked the Minister for Finance the cause of the delay in commencing work on the boatslip at Fenit, Tralee, County Kerry.*
>
> *Deputy David Stanton asked the Tánaiste and Minister for Enterprise, Trade and Employment the funding that has been made available to the research frontiers programme in 2008; the amount that has been expended to date; and if she will make a statement on the matter.*
>
> *Deputy Michael Creed asked the Minister for Finance the reason for the delay in issuing a P21 to a person (details supplied) in County Cork.*

Table 1 gives the totals and the distribution of questions since the start of the third Dáil in September 1922 (usually considered to be the start of the period of normal parliamentary politics). The corpus was stored in XML format.

The initial training dataset was derived from existing political science research [1] which had analysed a random sample of parliamentary questions between the 16th and 19th Dáils, from 02/03/57 to 13/03/73 inclusive, categorising them as local or national questions. This set, which included 97 local and 167 national questions, formed the initial training set available for experimentation.

An additional dataset was generated from questions, randomly selected from the full corpus, which were categorised manually by political science experts via a website developed for the purpose. This dataset acted as a hold-out set and included 19 local and 49 national questions. A profile of this training and hold-out data, split across the different Dáils is also given in Table 1.

4.2 Document and Feature Representation

The questions were represented in the standard bag of words representation, with the textual content of the questions tokenised at non-letter characters. The feature values used reflected the frequency of occurrence of each word in the

[7] http://historical-debates.oireachtas.ie/en.toc.dail.html

Table 1: Number of questions associated with each Dáil across the datasets used.

Dáil	**Start Date**	**Corpus**	**Personal Questions**	**Training Set National**	**Training Set Local**	**Hold-Out Set National**	**Hold-Out Set Local**
3	09/09/22	1,172	-	-	-	-	-
4	19/09/23	5,076	-	-	-	-	1
5	23/06/27	521	-	-	-	-	1
6	11/10/27	4,482	-	-	-	-	-
7	09/03/32	642	-	-	-	-	-
8	08/02/33	4,532	-	-	-	1	-
9	21/07/37	1,080	-	-	-	-	-
10	30/06/38	5,065	-	-	-	-	1
11	01/07/43	1,246	-	-	-	-	-
12	09/06/44	5,634	-	-	-	1	1
13	18/02/48	9,768	1	-	-	-	2
14	13/06/51	9,835	-	-	-	1	2
15	02/06/54	5,608	6	-	-	-	-
16	02/03/57	11,228	32	26	16	1	-
17	11/10/61	14,576	127	38	30	-	-
18	21/04/65	18,507	522	49	29	-	-
19	20/07/69	19,575	939	54	22	4	1
20	14/03/73	17,937	1,931	-	-	3	-
21	05/07/77	19,507	3,781	-	-	1	1
22	30/06/81	2,209	1,779	-	-	-	-
23	09/06/82	4,207	4,638	-	-	-	-
24	04/12/82	24,123	23,800	-	-	2	1
25	10/03/87	18,082	6,842	-	-	1	2
26	29/06/89	32,290	7,437	-	-	1	1
27	17/12/92	55,076	10,288	-	-	7	-
28	26/06/97	84,347	23,148	-	-	10	3
29	06/06/02	110,898	34,331	-	-	14	2
30	14/06/07	36,751	14,192	-	-	2	-
Total		523,974	133,794	167	97	49	19

text of the question. In all experiments a Support Vector Machine classifier[8] with a linear kernel was used due to its appropriateness for text classification [18]. As the training dataset is quite imbalanced the performance was measured using average class accuracy, which is the average accuracy achieved across the national and local classes on a ten-fold cross validation of the training set.

Preliminary cross validation experiments showed that removing stopwords was beneficial to generalisation accuracy, whereas stemming was not. In addition including the response text in the tokenisation did not improve the gener-

[8] The implementation used was the SMO classifier in Weka [17].

alisation accuracy; in fact it had a significant detrimental effect. Experiments which included the response text in tokenisation but annotated similar words from the question text and response text as separate features performed better than using both indistinguishably but did not improve over the generalisation accuracy achieved on question text alone.

A series of experiments were run to investigate the stability of the classifier and get an estimate of overall generalisation accuracy. Table 2 reports on these experiments, which formed the basis for the selection of a classifier to use for the categorisation of all questions in the full corpus.

Table 2: Accuracy results for a SVM with different feature sets. Row (a) reports cross validation accuracy on the training set, (b) reports the accuracy on the hold-out set after training on the training set (c) reports cross validation accuracy across a combination of the training and hold-out examples.

		National Class Accuracy	**Local Class Accuracy**	**Average Class Accuracy**
(a) Cross validation on Training Set	Stopword Removal (SR)	89.8%	60.8%	75.3%
	SR + DFR < 50	89.8%	62.9%	76.3%
	SR + DFR < 500	86.8%	62.9%	74.8%
(b) Hold-out Set	SR	91.8%	47.4%	69.6%
	SR + DFR < 50	91.8%	47.4%	69.6%
	SR + DFR < 500	89.8%	52.6%	71.2%
(c) Cross validation on Training Set + Hold Out Set	SR	92.1%	66.4%	79.1%
	SR + DFR < 50	90.7%	67.2%	78.9%
	SR + DFR < 500	86.1%	69.8%	78.0%

The baseline experiment was to produce a bag of words from the question text and to remove stopwords. Although the training set is relatively small, the dimensionality of the feature space is large (2028 features) and has the potential to become significantly larger with the addition of more training examples. Document Frequency Reduction (DFR), a technique that removes words that occur in less than a specified number of documents in the training set, was used to reduce the dimensionality of the feature space. Given that the full corpus was available, rather than using the training set to determine the words to remove, the full corpus was used. Feature sets which included the removal of all words that occur in less than 50 questions (approx 0.0001% of the full corpus) and less than 500 questions (approx 0.001% of the full corpus) were also produced, which resulted in training sets of 1494 and 1046 features respectively.

The results of ten-fold cross validation accuracy on the baseline training set and reduced feature training sets is shown in Table 2 (denoted (a)). This table also includes the results of testing the hold-out set on each of the feature sets (denoted (b)) and finally it includes the cross validation accuracy of training on an extended set of questions derived from combining both the initial training set and the hold-out set (denoted (c)).

Looking at the results in Table 2, it is clear that it is much easier to identify national questions than local questions. This could be influenced by the fact that the training sets are imbalanced with a higher proportion of national examples, although it might be considered that the diversity in local questions make it a harder concept to learn. The class accuracies of local questions on the hold-out set trained on the different feature sets are poor, but the proportion of local questions in both the training and hold-out sets is low suggesting more labelled data is required. Training on the combined sets improves the local class accuracy but it is still lower than the accuracy achieved on the national questions.

5 Classifying Dáil Questions from 1922 - 2008

Based on the results achieved in the experimentation discussed above the classifier selected to classify the full corpus was an SVM classifier trained on both the initial training set and the hold-out set (a total of 332 questions) using a bag of words representation. Dimensionality reduction of the feature set was achieved by stopword removal and the removal of words that occur in less than 500 of the questions in the full corpus. Although this combination does not return the highest accuracy of results in Table 2, it achieves the most balanced results between national and local class accuracy figures. Personal questions were not included in the classification as our current assumption is that personal questions are intrinsically local in nature.

Figure 1 shows the results of classifying parliamentary questions over the entire period September 1922 - December 2008. The local category figures in both graphs in Figure 1 include those classified from the full corpus as local plus the personal questions which we assume to be of the local category.

The first graph in Figure 1 shows the volume of parliamentary questions over the period distinguishing between those with a national focus and those with a local focus. The contrast between the early and the contemporary periods is very striking. Prior to the middle of the 20th century, exploitation of the opportunity presented by parliamentary questions was extremely low tending to be in the hundreds rather than thousands per year. The rate of questioning picked up somewhat in the early 1960s but really took off only in the late 1980s, continuing to expand substantially thereafter. This may reflect the changes made to the question procedure brought about by the 1986 reforms. The second striking aspect is the indication that questions have never been predominantly local in focus and that the recent expansion in the use of the procedure is substantially more geared towards national rather than towards local issues.

Figure 1 elaborates on both of these points by displaying in the second graph the ratio of national to local questions over the entire period under consideration. It shows that there was a clear predominance of national questions from the early 1990s to 2008. Intriguingly, the results also show recurring periods marked by the predominance of national issues from as early as the 1930s with just two periods

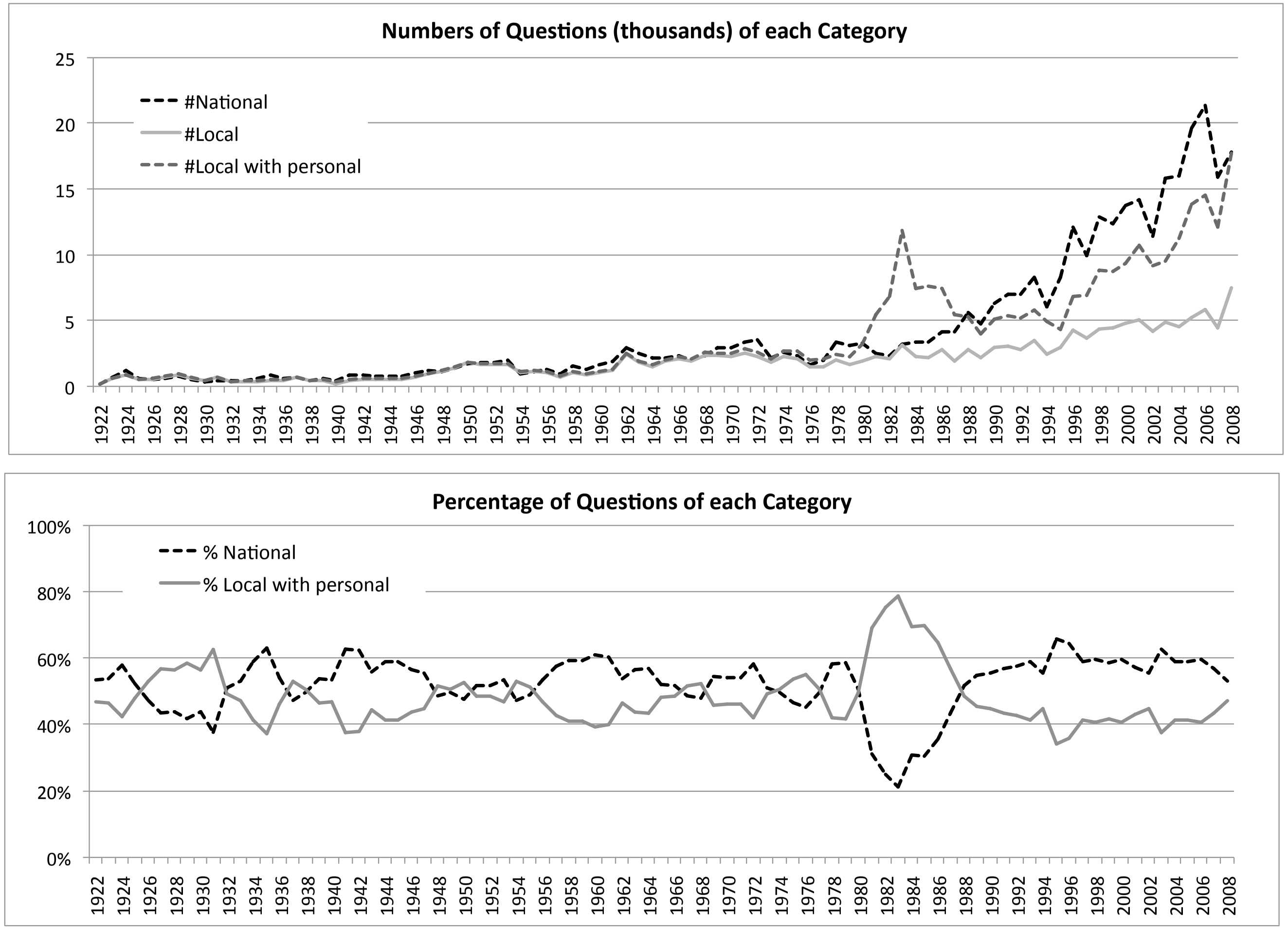

Fig. 1: Numbers and percentage of Dail questions each year by category

in which local topics clearly predominate (1926-32 and 1981-87). In addition to these two there are several other periods in which the proportion of local topics slightly exceeds the proportion categorised as national. This somewhat greater emphasis on local topics shows substantial correspondence with periods in which Fianna Fáil (traditionally the largest party in the Dáil) were in opposition.

Given the tentative nature of our findings, it would be unwise to draw definitive or far-reaching conclusions at this stage. One can argue, however, that there is little evidence in the corpus of parliamentary questions in Dáil Éireann to support the view that the role of the TD is determined by mainly clientelist/parochial imperatives and that there are some indications that this varies by party.

6 Conclusions and Future Work

This paper has presented an initial analysis of over 500,000 Dáil questions using automated text categorisation techniques. Using supervised machine learning techniques to assist in the analysis of parliamentary questions has much to offer, providing as it does evidence that is otherwise impossible to collect.

The generalisation accuracy achieved on this preliminary investigation could be considered lower than desirable. Future work will consider ways to improve the accuracy of the results. The length of the text in the questions tends to be quite short (the average length of question text in the corpus is approximately 50 words) so techniques that are appropriate for short text classification such as feature expansion techniques will be considered. Also n-gram tokenisation has proven successful in text classification [19] and may also be beneficial here.

The volume of training data used relative to the size of the corpus is very small and there is evidence that providing more training examples improved the results. This is a good example of a scenario that would benefit from using active learning to assist with labelling by targeting informative training examples. Using active learning will also assist in building an accurate classifier with the minimal amount of labelling effort and should help with improving the generalisation accuracy.

The focus of this investigation has been on classifying questions as local or national. There are other distinctions that are important to the experts in this area such as identifying whether questions are related to policy or administration. Future work in this area will consider this and other such distinctions. Finally, with improved accuracy, a potentially fruitful direction of further research would be to apply the method to the very similar system of parliamentary questions in the British House of Commons and to compare the evolution of the use of parliamentary questions in both systems over the same time period.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/RFP/CMSF718.

References

1. Sinnott, R.: A Study of Parliamentary Questions in Dáil Eireann. Master's thesis, Department of Ethics and Politics, University College Dublin (1972)
2. Martin, S.: Using Parliamentary Questions to Measure Constituency Orientation: An Application to the Irish Case. Working Papers in International Studies, Centre for International Studies, Dublin City University (2010)
3. Tapio, R.: Parliamentary Questions in the European Parliament: Representation, Information and Control. The Journal of Legislative Studies **2** (1996) 356–82
4. Padilla, A., Gibson, I.: Science moves to centre stage. Nature **403** (2000) 357–359
5. Brid, K.: Gendered Parliamentary Questions. British Journal of Politics and International Relations **7** (2005) 353–70
6. Abdul-Nasir, J.: Learning admissibility rules for questions posed to the National Assembly of Pakistan. Master's thesis, University of Oxford (1998)
7. Cardie, C., Wilkerson, J.: Guest Annotation for Political Science Research. Journal of Information Technology and Politics **5** (2008)
8. Munroe, B.L., Schrodt, P.: The Statistical Analysis of Political Text. Political Analysis **16** (2008) 351–355
9. Hillard, D., Purpura, S., Wilkerson, J.: Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. Journal of Information Technology and Politics **4** (2007) 31–46
10. Klebanov, B.B., Diermeier, D., Beigman, E.: Lexical Cohesion Analysis of Political Speech. Political Analysis **16** (2008) 447–463
11. Munroe, B.L., Colaresi, M.P., Quinn, K.M.: Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis **16** (2008) 372–403
12. Yu, B., Kaufmann, S., Diermeier, D.: Classifying Party Affiliation from Political Speech. Journal of Information Technology and Politics **5** (2008) 33–48
13. Arguello, J., Callan, J., Shulman, S.: Recognizing Citations in Public Comments. Journal of Information Technology and Politics **5** (2008) 49–71
14. Malouf, R., Mullen, T.: Taking sides: User Classification for Informal Online Political Discourse. Internet Research **18** (2008) 117–190
15. van Atteveldt, W., Kleinnijenhuis, J., Schlobach, S.: Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish between Positive and Negative Relations. Journal of Inf. Tech and Politics **5** (2008) 73–94
16. Yu, B., Kaufmann, S., Diermeier, D.: Exploring the Characteristics of Opinion Expressions for Political Opinion Classification. In: Procs of the 2008 Int. Conf. on Digital Government Research. (2008) 82–91
17. Witten, I.H., Frank, W.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann (2005)
18. Joachims, T.: Text Categorisation with Support Vector Machines: Learning with many relevant features. In Nedellec, C., Rouveirol, C., eds.: Procs of ECML-98. Number 1398 in LNCS (1998) 137–142
19. McNamee, P., Mayfield, J.: Character N-Gram Tokenization for European Language Text Retrieval. Inf. Retr. **7** (2004) 73–97