Cross-Domain Topic Classification for Political Texts

Moritz Osnabrügge[1], Elliott Ash[2], and Massimo Morelli[3]

[1]*Durham University, School of Government and International Affairs, Durham, United Kingdom. Email: moritz.osnabruegge@durham.ac.uk*
[2]*ETH Zurich, Center for Law & Economics, Zurich, Switzerland. Email: ashe@ethz.ch*
[3]*Bocconi University, Department of Social and Political Sciences, Milan, Italy. Email: massimo.morelli@unibocconi.it*

Abstract

We introduce and assess the use of supervised learning in cross-domain topic classification. In this approach, an algorithm learns to classify topics in a labeled source corpus and then extrapolates topics in an unlabeled target corpus from another domain. The ability to use existing training data makes this method significantly more efficient than within-domain supervised learning. It also has three advantages over unsupervised topic models: the method can be more specifically targeted to a research question and the resulting topics are easier to validate and interpret. We demonstrate the method using the case of labeled party platforms (source corpus) and unlabeled parliamentary speeches (target corpus). In addition to the standard within-domain error metrics, we further validate the cross-domain performance by labeling a subset of target-corpus documents. We find that the classifier accurately assigns topics in the parliamentary speeches, although accuracy varies substantially by topic. We also propose tools diagnosing cross-domain classification. To illustrate the usefulness of the method, we present two case studies on how electoral rules and parliamentarian gender influence the choice of speech topics.

Keywords: cross-domain classification, supervised learning, text analysis, manifesto corpus, parliamentary speeches, electoral reform, debate participation

1 Introduction

Social scientists have expended significant resources to hand-code political text data. For example, the Comparative Agendas Project and the Manifesto Project have coded many documents across a variety of politically relevant categories (Budge *et al.* 2001; Jones and Baumgartner 2005). Scholars have used this hand-coded data to measure quantities of interest in studies on party competition, legislative politics, and political stability, amongst others (e.g., Tsebelis 1999; Tavits and Letki 2009; Böhmelt *et al.* 2016). Meanwhile, an increasing number of studies are using a hand-coded subsample of text data to automatically code unlabeled documents using supervised-learning techniques (Grimmer and Stewart 2013). These studies train and test supervised-learning models on certain types of text documents, such as tweets or legislative texts, to classify unlabeled text documents of the same type to the same categories (e.g., Hopkins and King 2010; Workman 2015; Anastasopoulos and Bertelli 2020; Barberá *et al.* 2021). We call this approach *within-domain* supervised learning.

In this paper, we study and assess the use of supervised-learning approaches to *cross-domain* topic classification of political texts. In this approach, the supervised-learning models are trained and tested based on a certain type of text document to classify texts belonging to a different type of text document. Compared to within-domain classification, cross-domain classification significantly reduces the data collection costs because researchers can use existing training data. In contrast to unsupervised learning approaches such as latent dirichlet allocation (LDA), we can validate our model using well-established metrics (Hastie, Tibshirani, and Friedman 2009; Wilkerson and Casas 2017).

Political Analysis (2021)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
Moritz Osnabrügge

Edited by
John Doe

We use existing hand-annotated party platforms from the manifesto corpus as a source corpus to categorize the topics of parliamentary speeches. Manifestos and parliamentary speeches are related to each other because both types of documents focus on the policy priorities of parties or party members. However, they also have multiple differences. While manifestos detail parties' electoral programs, speeches are statements of individual parliamentarians held after parliamentary elections (e.g., Laver, Benoit, and Garry 2003; König, Marbach, and Osnabrügge 2013; Proksch and Slapin 2014). Cross-domain analysis could be used in future research to compare priorities expressed in party manifestos, speeches, coalition agreements, legislative texts and social media data.

Methodologically, we start by training a machine classifier to learn topics from the annotated party platforms in the manifesto corpus. The manifesto corpus is one of the largest and most widely used hand-annotated corpora of political documents (e.g., Slapin and Proksch 2008; Mikhaylov, Laver, and Benoit 2012; Zirn *et al.* 2016). It includes over 115,000 English-language statements labeled according to 44 narrow topics and 8 broad topics.

After training the topic predictor using machine learning, we use it to classify topics in a corpus of parliamentary speech transcripts from the New Zealand Parliament. This target corpus encompasses the universe of parliamentary speeches for the period from 1987 to 2002. We calculate both the most likely topic and the entire distribution of topic probabilities for each speech.

To validate that the topic prediction works in the new domain, we compare predictions to those made by an expert coder for 4,165 parliamentary speeches. The coder received training from the Manifesto Project and coded speeches based on manifesto categories following the project's guidelines. We find that the accuracy is similar to the expected accuracy inherent in human coder misclassification. We assess the replicability of our findings by asking three additional coders to code a subset of the speeches. For additional robustness, we show that the topic predictions have similar accuracy in speeches by U.S. Congressmen.

We propose two tools for diagnosing cross-domain classification without annotating the target corpus. First, we establish that the within-domain accuracy by topic is a strong predictor of cross-domain accuracy. Second, we introduce a diagnostic metric, *feature congruence*, which captures the semantic equivalence of a classified topic in the source corpus and the target corpus. Higher feature congruence predicts greater cross-domain accuracy for a given topic. Together, these diagnostic steps identify which topics can or cannot be reliably classified in a target domain, before expensive additional hand annotation is undertaken.

Finally, we illustrate two applications of cross-domain supervised learning. First, we study the consequences of New Zealand's 1993 electoral reform, which changed the system from first-past-the-post to mixed-member proportional representation (Vowles *et al.* 2002). We find that the reform significantly increased attention toward the issue of political authority, which includes discussions about political stability and party competence. In the second application, we study whether the gender of New Zealand parliamentarians is related to their debate participation on certain topics. We find that women speak more about welfare, while men talk more about external relations (Bäck and Debus 2019).

Our paper adds to the literature by demonstrating and assessing the opportunities to use supervised learning for cross-domain text classification in a setting that is relevant for applied political scientists. Recent work uses machine-learning models to analyze text across domains, but focuses on settings in which the source and target corpora differ more substantially. Burscher, Vliegenthart, and De Vreese (2015) assess the cross-domain classification of Dutch-language news articles and parliamentary questions using issue categories from the Comparative Agendas Project. Yan *et al.* (2019) use data on U.S. Congress speeches and media platforms to predict partisanship across domains. These studies find that cross-domain classification does not work well in their respective contexts.

Table 1. Summary of design factors for topic classification methods

	Dictionaries: Customs	Dictionaries: Generic	Topic Modeling	Supervised Learning: Within-Domain	Supervised Learning: Cross-Domain
Design Efficiency	Low	High	High	Low	High
Annotation Efficiency	High	High	High	Low	Moderate
Specificity	High	Moderate	Low	High	Moderate
Interpretability	High	High	Moderate	High	High
Validatability	Low	Low	Low	High	High

Notes. Overview of the pros and cons of the main methods for topic classification in political science. The columns list the methods, while the rows list the design factors underlying the choice of method.

Next, we present a toolkit that can be used to assess and validate the cross-domain classification of topics. In addition to inspecting within- and cross-domain classification accuracy metrics, we review their performance by topic and propose two tools for diagnosing cross-domain classification. Our tools can help conserve scarce empirical resources by helping researchers focus on analyzing topics that can be reliably measured in a target domain. Finally, we share our code, trained models, and hand-annotated data, which researchers can use to assess alternative methods for cross-domain classification.

2 Background: Topic Classification in Political Science

In political science there are three main approaches to categorizing individual documents (Grimmer and Stewart 2013) – lexicon-based pattern matching, unsupervised topic models, and supervised-learning classifiers. This section discusses the pros and cons of these methods, as well as those of supervised learning for cross-domain classification.

Building upon Quinn *et al.* (2010), Table 1 summarizes the main approaches along five design factors. First, design efficiency assesses the amount of time an expert researcher would spend designing a classification system. Second, annotation efficiency denotes the time needed to annotate documents. Third, specificity refers to how much the system can be targeted toward answering specific questions or exploring particular features in the data. Interpretability summarizes how straightforward it is to interpret the resulting topic classifications. Finally, validatability refers to the feasibility of validating topics – that is, checking whether the classifier is correctly grouping topics.

The dictionary- or lexicon-based approach works by searching for particular textual patterns in the text to assign topics. Researchers can create their own dictionaries by identifying words related to the topic of interest. For example, previous work identifies a lists of words related to "women" to detect documents on issues or topics that are important for women (e.g., Pearson and Dancey 2011). Alternatively, researchers can use existing generic dictionaries, such as the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker *et al.* 2015).

Custom dictionaries entail significant up-front costs for the researcher to build the tags dictionary, but after that the annotation costs are zero. They have high specificity in the sense that they give the researcher full control over the dimensions of text they would like to target. For instance, if one is interested in women's issues, one can search for that topic. The method is also highly interpretable because the tags already contain expert knowledge and can be easily inspected.

Generic dictionaries (like custom dictionaries) have the advantages of negligible annotation costs and high interpretability. One can easily read the full list of terms to see what is going on. The advantage of generic dictionaries is the much lower up-front design time, as previous researchers have already produced and validated them. The tradeoff is a significant loss in specificity, as one

can only measure the dimensions of text that are available in the dictionaries.

The major downsides of both custom and generic dictionaries involve their highly constrained representation of language and limited validatability. The lexicon tags are unavoidably subjective, over-inclusive, and under-inclusive. For example, politicians use the word "women" in many contexts that are unrelated to the substantive representation of women's issues. Some documents will have tags from multiple categories, and many documents will have no tags. There is no easy way to deal with these cases. One cannot tell how well the labels work without significant investing significantly in labeling the documents, which defeats the purpose of using dictionaries (the low annotation costs) (see also Barberá *et al.* 2021).

The next major approach to text classification is topic modeling, such as latent dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), the expressed agenda model (Grimmer 2010), the dynamic topic model (Quinn *et al.* 2010), or the structural topic model (Roberts *et al.* 2013). These algorithms provide a form of interpretable dimension reduction in which documents are transformed from high-dimension counts over words to low-dimensional shares over topics. Topic models are a powerful tool because they often produce intuitive, interpretable topics without any labeled training data (Catalinac 2016; Greene and Cross 2017).

The major advantage of topic models, as mentioned above, is that they do not require any labeled training data to classify documents into categories. Moreover, the design costs are very low, as for example in LDA the only major design choice is how many topics, and most other steps are automated. In turn, there are zero annotation costs, as a trained topic model can instantly produce a set of topic probabilities for any given document. It allows documents to have multiple topics, and properly deals with all documents.

However, unsupervised topic models have limitations (Wilkerson and Casas 2017; Denny and Spirling 2018). The topics are learned directly from the data, so specificity is low and the topics may or may not represent the language dimensions in which the researcher is interested. In particular, topic models do not work on multilingual corpora because the learned topics will not match up across languages. Interpreting topics requires additional work after estimating the model, and the produced topics can be sensitive to perturbations in the data, such as the steps taken in text pre-processing, featurization, and the number of topics chosen (Denny and Spirling 2018). This sensitivity is problematic because there are no unified diagnostics for validating topic models (e.g., Grimmer and Stewart 2013; Roberts, Stewart, and Tingley 2016).

A third approach to classifying text is supervised learning, where researchers randomly sample some of the documents and hand-annotate the topics to create a labeled training dataset (e.g., Drutman and Hopkins 2013; Workman 2015; Barberá *et al.* 2021). With a set of labeled documents in hand, one can use machine learning to encode the relationships between text features and topics. The trained machine-learning model can then automatically classify the topics in unlabeled data. Most political science studies use supervised learning for within-domain classification.

Supervised learning has several major advantages. First, like custom dictionaries the system can be highly targeted toward classifying any dimension of the text that the researcher is interested in (high specificity). This semantic targeting includes the capacity to produce labeled corpora in multiple languages. Second, the topics are highly interpretable because one can read the codebook provided to annotators. One can also look at example documents for each category. Third, and perhaps most importantly, supervised-learning classifiers can be rigorously validated (Denny and Spirling 2018). The standard machine-learning approach involves dividing the annotated data into a training set and a test set and assessing how well the classifier works in held-out test data (e.g., Hastie, Tibshirani, and Friedman 2009; Peterson and Spirling 2018). The classification accuracy metrics in the test set provide a good estimate of how well the classifier will work in the unlabeled documents.

The supervised-learning approach has two general shortcomings related to costs. First, the

researcher must design a set of topics and build detailed documentation and codebooks for annotators. Second, the annotators must spend a significant amount time being trained, performing the annotations, and comparing their results with those of other annotators. Recent innovations have emerged to reduce these costs, such as active learning and the use of crowdsourcing for hand-annotation (Benoit *et al.* 2016; Miller, Linder, and Mebane 2020). Still, for most applications, within-domain supervised learning requires a large investment of time and money to hand code enough documents to make a classifier useful.

We focus on supervised learning for cross-domain classification, in which a text classifier is built in one domain (the source corpus) and applied in another domain (the target corpus). While recent papers have explored the usefulness of cross-domain supervised learning (Burscher, Vliegenthart, and De Vreese 2015; Yan *et al.* 2019), there is very little evidence on its relative performance.

Analogous to the move from custom dictionaries to generic dictionaries, the cross-domain approach inherits some of the benefits of supervised learning for within-domain classification. For instance, cross-domain supervised learning exhibits higher levels of interpretability than unsupervised topic modeling. The classifier is interpretable in the same way as in within-domain supervised learning, as one can read the annotation codebooks and examine sample documents (but this time in the target corpus). Unlike dictionary methods and unsupervised topic modeling, cross-domain supervised learning can be validated using well-established metrics, such as classification accuracy (Hastie, Tibshirani, and Friedman 2009; Géron 2017). More specifically, the classifier can be validated by annotating a sample of documents in the target corpus to compute cross-domain test-set accuracy metrics.

The main advantage of cross-domain classification over within-domain classification is that researchers can draw on existing labeled corpora as training data. This reduces the design costs to zero, since the researcher borrows the complete schema and codebooks of the original system. The annotation costs are largely eliminated, as the previous annotations are used to train the classifier. That said, some annotations are needed in supervised learning for cross-domain classification (as reflected in Table 1) to validate, rather than build, the classifier.

A disadvantage of cross-domain supervised learning, which is again analogous to the shift from custom to generic dictionaries, is a loss of specificity relative to within-domain classification. The set of questions and policies that one can analyze with supervised learning for cross-domain classification depends on the availability of existing labeled datasets. As the availability of training data expands, including consistent multilingual corpora, this disadvantage should become less relevant.

3 Cross-Domain Classification

This section outlines the implementation of a supervised learning approach for cross-domain classification of political texts. We trained a classifier using an annotated source corpus and applied it to an unlabeled target corpus. First, we prepared the source corpus and target corpus for machine learning. Second, we trained a machine-learning model based on the source corpus. Third, we predicted the topics of the texts of the target corpus. Fourth, we annotated the target corpus and evaluated the model's performance.

3.1 Source Corpus: Manifesto Project Party Platforms

Our source corpus consists of party platforms annotated by the Manifesto Project. We accessed the English-language manifesto statements from the following countries: Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States (Krause *et al.* 2018). Using manifesto statements from all of these countries increases our training sample size, which is important for machine classification tasks. The data set has $N_S = 115,410$ rows of annotated policy statements,

where S indicates the "source" corpus.

Each statement includes a hand-annotated topic code. The statement "and reduce global warming emissions" refers, for example, to the environment (category 501), while the statement "We can't afford another dose of Labour" relates to political authority (category 305). The Manifesto Project usually has one trained coder for each country. Appendix Section A.1 shows an example statement for each topic.

We pre-processed the topic codes k following two specifications. The 44-topic specification ($K = 44$) takes into account all topics and merges categories that focus on the same topic but in a different direction (positive/negative) because we are interested in the topic rather than then sentiment of the text. For example, we combined the categories "per607 Multiculturalism: Positive" and "per608 Multiculturalism: Negative" to create one "Multiculturalism" topic. This procedure generates a sample of 44 categories.

For the 8-topic specification ($K = 8$), we merged all categories into eight major topics following the Manifesto codebook (Budge *et al.* 2001) – external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society, social groups, and no topic. This specification merged multiple similar categories, which facilitates the interpretation of the results. Appendix Section A.1 contains additional information about the corpus and this process, as well as example snippets for each topic.

3.2 Corpus Pre-processing

Before training the machine-learning model, we took standard pre-processing steps to transform the text data into a document-term matrix. First, we removed uninformative features – stopwords, punctuation, and capitalization. We then took the left-over words and constructed N-grams (phrases) up to length three – words, bigrams, and trigrams. We dropped N-grams that appeared in fewer than 10 documents, as they contain little predictive information. We also dropped those that appeared in more than 40 percent of the documents, as these are likely specific to manifestos and not distinctive to specific topics. Finally, we computed term-frequency/inverse-document-frequency (TF-IDF) weights for each N-gram, treating each manifesto statement as a document. The resulting document-term matrix has $M = 19,734$ columns, with each column is indexed by j.[1]

3.3 Machine Classifier Training

The next step was to train a machine-learning model based on the document-term matrix of the source corpus. A range of machine-learning classifiers could be used for cross-domain classification. We employed the regularized multinomial logistic regression model, a widely used multiclass prediction model (Hastie, Tibshirani, and Friedman 2009; Géron 2017). We trained separate models for the 8-topic and 44-topic labels.

We used the standard approach to machine classification. First, we split the data into a 75 percent training and 25 percent held-out test set. Second, we learned hyperparameters (the regularization penalty and class weighting) by conducting a three-fold cross-validation grid search in the training set. This procedure determined that the best parameters were an inverse of the regularization strength equal to two and no weighting of the categories (for both 44-topic and 8-topic models).[2]

We experimented with other machine-learning algorithms, including a random forest, gradient boosting, and a neural net. These models performed similarly in classifying manifesto statements to topics. For example, Appendix Section B.5 reports similar results for gradient boosting. However,

1. We experimented with other text-preprocessing steps and found that the performance is similar. See Appendix B.4.
2. We implemented the logistic regression model using Python's Scikit-Learn package (Pedregosa *et al.* 2011). To solve the optimization problems, we used the Newton-conjugate gradient solver.

these models have more hyperparameters to tune and take much longer to train. Thus, we decided to use the regularized multinomial logistic regression model for this analysis.

3.4 Target Corpus: New Zealand Parliamentary Speeches

Our target corpus consists of speeches delivered by members of the New Zealand Parliament from 1987 to 2002. We chose this target corpus for three reasons. First, New Zealand's Parliament uses the English language, which facilitates analysis because the manifesto corpus includes a large number of hand-annotated English statements. Second, the trained classifier is applicable to text data from other English-language parliaments, such as the U.S. Congress and U.K. Parliament. Third, this period in New Zealand is empirically interesting because of an electoral reform in 1993, which we examine to assess the usefulness of cross-domain supervised learning.

We extracted the speech data from the *Hansard*, which is the official record of the New Zealand Parliament. We removed speeches given by the speaker of parliament and his/her deputy. Next, we removed short oral contributions from 'government member(s)' and 'opposition member(s)' without further information on the name of the speakers. We also dropped speeches with fewer than 40 characters excluding numbers (Peterson and Spirling 2018) and Maori-language speeches for which an official translation was not provided. The final dataset contains $N_T = 290,456$ documents, where T indicates the target corpus. Appendix Section A.2 provides additional information on the data.

3.5 Predicting Topics in Target Corpus

The text classifier described in Subsection 3.3 can be applied to any snippet of text. The prediction pipeline takes a string of text and outputs a probability distribution over topic categories. In this application, we first used the vocabulary and the document frequencies from the source corpus in the pre-processing step. These were applied to vectorize each parliamentary speech, producing N-gram frequencies for speech i. These frequencies are the features used to compute the model predictions.

Second, we took the logistic regression classifiers for the 44-topics and 8-topic specifications, trained them on the full 100 percent sample of manifesto statements, and applied them to the feature vectors for each parliamentary speech. The model outputs a set of predicted probabilities across topic classes based on the logistic regression coefficients learned in the source corpus. Summary tabulations of the speech topics by year and in total are reported in Appendix Section B.3.

The predicted probabilities sum to one and can be interpreted as the shares of a document that is allocated to each topic. Alternatively, one can take the highest-probability topic and mark a speech as having only that topic. The choice between a single-topic or multiple-topic representation will depend on the downstream empirical task.

3.6 Annotation of Target Corpus Validation Set

To validate the cross-domain predictions, we arranged for the hand coding of a sample of documents in the target corpus. We follow previous work that has used human judgment to validate the statistical analysis of political documents (Lowe and Benoit 2013). We hired the manifesto coder for New Zealand, who was trained by the Manifesto Project and has coded the manifestos from New Zealand for the manifesto project.

The coder annotated a random sample of 4,165 parliamentary speeches in one manifesto category. We only gave the coder the text of the speech and no meta-data such as the date or speaker. We asked the coder to code based on the Manifesto Project. We annotated topics by speech as our downstream empirical analysis is at the speech level, which allowed us to obtain

much more data than sentence-level annotations. This approach is in line with Barberá *et al.* (2021, 28-29), who find that breaking larger text units into sentences does not improve classification performance. The annotations took a total of 52.5 hours.

We hired three additional coders to assess inter-coder reliability within the New Zealand target corpus (Mikhaylov, Laver, and Benoit 2012). Like the main coder, these coders also received training from the Manifesto Project in English-language platforms; we asked them to code according to the Manifesto Project methodology. The coders were not experts on New Zealand politics, however. We drew a random sample of 250 speeches from the 4,165 speeches annotated by the first coder. Each of the three secondary coders annotated the same subsample of 250 speeches, which gave us four annotations in total.

Finally, to assess the method's potential broader generalization, we also hand annotated a corpus of congressional speeches from the United States. We hired the Manifesto Project coder for the United States and asked him to code a random sample of 150 speeches from the House of Representatives. The sample was drawn from all speeches contained in the *Congressional Record* for the period August 1987 through July 2002. All five coders assigned each speech to one manifesto category.

3.7 Model Performance Evaluation

We followed standard machine-learning approaches to evaluate the performance of our machine classifier. For the within-domain performance, we assessed the predictive performance in a 25-percent held-out test sample. For the cross-domain performance, we compared the machine predictions to the new annotations provided by the human coder. We also examined performance by class, because researchers are likely to be interested in particular topics for any given empirical application. In the appendix, we use bootstrapping to assess the robustness of the metrics to sampling variation.

We report a variety of metrics to evaluate and understand model performance. First, we report the simple (top-1) accuracy. This is the proportion of predicted topics (that is, the topic with the highest predicted probability) in the test set that are also the "true" topic as selected by human annotation. Note that simple accuracy is equal to the model's micro-weighted aggregate precision, recall, and F1 score. As the manifesto corpus includes multiple similar categories (e.g., economic goals and economic growth), we do not only report how often the true topic is correctly ranked first, as well as more broadly how often it is highly ranked and within the top few topics by predicted probability. We therefore calculated the top-3 (and top-5) accuracy – the proportion of observations for which the true class (from the hand annotations) is within the top three (or five) categories as ranked by their predicted probability defined by the machine classifier.

Simple accuracy summed across test samples, such that categories with more documents in the test sample are weighted higher in the metrics. Therefore, inaccurate predictions in the less frequent categories could be missed. To provide a more rounded aggregate report, we computed the balanced accuracy, which is the (un-weighted) average recall (fraction of true-class documents correctly identified) across output categories. Finally, we report the macro-weighted F1 score, which is the (un-weighted) average of the F1 scores (harmonic mean of precision and recall) across all categories.

4 Classification Results

This section reports the results on the performance of our classifier. We show that it works in-domain, in that it can reliably reproduce the hand-coded topic labels in the manifesto corpus. We also assess its performance in the domain of New Zealand parliamentary speeches.

Table 2. Overview of classifier performance in test set

	Within-Domain		Cross-Domain	
	44 topics	8 topics	44 topics	8 topics
Top-1 accuracy / F1 micro	0.538	0.641	0.410	0.507
Top-3 accuracy	0.766	0.908	0.650	0.816
Top-5 accuracy	0.841	0.975	0.747	0.916
Balanced accuracy	0.388	0.504	0.265	0.451
F1 macro	0.417	0.523	0.261	0.450

Notes. Within-Domain denotes the performance of a classifier trained on manifesto statements and evaluated on (held-out) manifesto statements. *Cross-Domain* describes the performance of a classifier trained on manifesto statements and evaluated on parliamentary speeches. 44 topics and 8 topics refer to the models with the narrow and broad output classes. Top-1, top-3 and top-5 accuracies, balanced accuracy and F1 macro are the performance metrics.

4.1 Aggregate Performance

The classification results are summarized in Table 2. Columns 1 and 2 report the results for the within-domain (source-source) predictions (manifesto-trained model applied to manifesto test corpus), while columns 3 and 4 report the results for the cross-domain (source-target) predictions (manifesto-trained model applied to the corpus of newly annotated New Zealand parliamentary speeches). Within each test corpus, we report metrics for 44 narrow topics (columns 1 and 3) and eight broad topics (columns 2 and 4).

Column 1 reports the 44-topic within-domain specification. It shows that the trained model predicts the correct category label 53.8 percent of the time. As expected, this is worse than the training-sample prediction (71 percent accurate), as the model somewhat over-fits the training data. As there are 44 topic labels to be assigned, choosing randomly would be correct about 2 percent of the time. Choosing the top category would be correct 13 percent of the time.

The within-domain efficacy of the classifier is further demonstrated in the top-3 accuracy (76.6 percent) and the top-5 accuracy (84.1 percent). These metrics show that even when the true class is not picked as having the highest probability, it is usually highly ranked. So if one is using predicted probabilities in an empirical analysis, one can have some confidence that they contain information about textual variation in policy dimensions.

To qualify these statements, we also report balanced accuracy (0.388) and macro-weighted F1 (0.417). These worse numbers reflect that, perhaps unsurprisingly, the less frequent categories are more likely to be mis-classified. Hence, empirical analyses of less frequent topics should be undertaken with caution.

We obtained better performance in the eight-topic within-domain prediction (column 2) due to the smaller number of classes that the machine needs to assign. The test-sample accuracy is 64.1 percent, closer to the in-sample accuracy of 76 percent and significantly better than guessing randomly (12.5 percent accuracy) or guessing the top category (31 percent accuracy). The top-3 accuracy (91 percent) is similarly encouraging. As before, the balanced accuracy (0.504) and F1 score (0.523) indicate lower performance for the less frequent classes.

Next we consider the cross-domain results (columns 3 and 4). Examining the results on the 44 topics, we find an overall top-1 accuracy of 0.410. This is significantly better than guessing at random (an accuracy of 2 percent) or guessing the most common hand-annotated class (accuracy of 19 percent). Perhaps unsurprisingly, the accuracy is lower than the in-domain accuracy (54 percent). The top-3 accuracy (0.650) and top-5 accuracy (0.747) are even more encouraging.

At the bottom of the table, the metrics for the eight-topic specification are also relatively encouraging. The overall top-1 performance is 51 percent, which is not that much less than the

within-domain accuracy of 0.641. It is much better than guessing randomly (0.125) or the most common class (0.256). The top-3 and top-5 accuracies are 0.816 and 0.916. This increase in accuracy compared to the 44-topic specification is similar to the increase we achieved using within-domain supervised learning.

Appendix Section B.11 uses a bootstrapping procedure to assess the sampling variation in the within-domain metrics. We produce a distribution of the metrics from models trained on resampled subsets of the manifesto corpus. We find that the mean and median accuracy across the samples are identical to the baseline accuracy (to two decimal places), with a standard deviation 0.003 (relative to a mean accuracy 0.539 for 44 topics and 0.645 for eight topics). Thus the metrics are not sensitive to training-set sampling variation.

To contextualize these results, we compare them to the performance of human manifesto coders reported by Mikhaylov, Laver, and Benoit (2012). The authors compare the Manifesto Project's master coding of 179 quasi-sentences to human coding. The quasi-sentences come from a manifesto from the United Kingdom and New Zealand. They find significant human coder error in the manifesto data: they aggregate all manifesto statements into three categories and report accuracies between 0.59 and 0.70. Unsurprisingly, coders' accuracy without aggregating into three broad categories is lower. Our within-domain and cross-domain classification accuracies are quite good in comparison.

4.2 Performance by Topic

For most empirical applications, one would be interested in analyzing variation in particular topics. Therefore, it is important to assess the variation in predictive performance across topics. To illustrate this type of evaluation, we report topic-level metrics. Further details on this issue can be found in Sections B.1 and B.2 in the Appendix.

First, we build confusion matrices using the 8-topic specification for the within-domain (Table 3a) and cross-domain (Table 3b) predictions. In Table 3, rows index true categories while columns index predicted categories. In Table 3a, a document is a test-set manifesto statement; in Table 3b a document is a hand-annotated New Zealand parliamentary speech. The numbers in the cells capture how often the model classified a document from the row class to the column class. The color (ranging from white to yellow to green) reflects the relative within-row frequency; darker colors indicate that the cell has more weight than other cells in the row. A good classifier will result in a confusion matrix with the highest counts on the diagonal.

For example, the first row of Table 3a shows that for the topic economy, the within-domain model correctly classified 5,382 of 7,306 manifesto statements; 1,091 economy statements were incorrectly classified as welfare and quality of life, while 833 were incorrectly assigned to one of the other five categories (besides no topic). These numbers correspond to a topic-specific recall (top-1 accuracy) of 0.737, reported in the right-most column. The economy column reports the counts for each true topic that was (mis-)classified as economy. For example, 1,015 welfare documents and 601 political system documents were mis-classified as economy. This is perhaps not too surprising given the potential semantic overlaps in discussions of these topics.

At the bottom of the column (and that of each topic), we report the ratio of total predicted count to total true count, which tells us how well the model replicates the distribution of topics. A value of 1 would mean that the distribution is the same; less than 1 indicates that the predictions are under-representative for this topic; greater than 1 means the predictions are over-representative for this topic. For the topic of economy, a predicted-to-true ratio of 1.106 means the predicted frequency of this topic is reasonably similar to the true frequency in the held-out test set.

As a whole, Table 3a shows that the within-domain model effectively replicates the annotated classes (besides the infrequent "other topic" class). The true category is selected most often across all topics. The minimum recall is a decent 0.422 (social groups), going all the way up to 0.776

Table 3. Classifier performance with eight topics: confusion matrices

(a) Within-domain predictions for party platforms

	Economy	External relations	Fabric of society	Freedom & democracy	Political system	Social groups	Welfare & quality of life	No topic / other	Total true	Recall
Economy	5,382	101	138	32	292	270	1,091	0	**7,306**	**0.737**
External relations	175	1,226	139	78	73	42	169	2	**1,904**	**0.644**
Fabric of society	254	103	1,816	84	181	121	609	1	**3,169**	**0.573**
Freedom & democracy	120	74	152	604	219	52	163	1	**1,385**	**0.436**
Political system	601	63	239	140	1,271	70	628	1	**3,013**	**0.422**
Social groups	458	36	200	44	113	1,185	775	0	**2,811**	**0.422**
Welfare & quality of life	1,015	65	319	59	238	331	7,006	1	**9,034**	**0.776**
No topic / other	73	9	24	10	36	12	61	6	**231**	**0.026**
Total predicted	**8,078**	**1,677**	**3,027**	**1,051**	**2,423**	**2,083**	**10,502**	**12**		
Total predicted / total true	**1.106**	**0.881**	**0.955**	**0.759**	**0.804**	**0.741**	**1.162**	**0.052**		

(b) Cross-domain predictions for parliamentary speeches

	Economy	External relations	Fabric of society	Freedom & democracy	Political system	Social groups	Welfare & quality of life	No topic / other	Total true	Recall
Economy	389	8	22	24	180	33	63	1	**720**	**0.540**
External relations	7	53	9	7	9	1	8	0	**94**	**0.564**
Fabric of society	18	4	239	44	76	23	28	0	**432**	**0.553**
Freedom & democracy	26	4	48	202	201	13	51	0	**545**	**0.371**
Political system	86	14	92	136	612	15	113	0	**1,068**	**0.573**
Social groups	41	3	23	19	62	123	54	0	**325**	**0.378**
Welfare & quality of life	56	0	28	26	153	34	492	0	**789**	**0.624**
No topic / other	33	5	20	16	58	10	49	1	**192**	**0.005**
Total predicted	**656**	**91**	**481**	**474**	**1,351**	**252**	**858**	**2**		
Total predicted / total true	**0.911**	**0.968**	**1.113**	**0.870**	**1.265**	**0.775**	**1.087**	**0.010**		

Notes. Table 3a refers to a model trained on manifesto statements; the predictions are from held-out manifesto statements. Table 3b refers to the same model, but the predictions are from newly annotated parliamentary speeches. The numbers in the cells represent the count of the number of row-class instances classified to the column class. The colors reflect the relative within-row frequency; darker green indicates higher counts. The rightmost column reports recall for the class. The bottom row gives the ratio of the number of predictions to the number of true instances in the class.

(welfare and quality of life). The most common misclassifications are somewhat intuitive. For example, many statements are classified into the welfare category, which could reflect that it is the most numerous category and is somewhat broad in its definition. Looking at the bottom row, meanwhile, we can see that overall, the distribution of topics is replicated quite well. Economy and welfare are slightly over-represented, while the other categories (especially social groups) are somewhat under-represented.

Next we consider the cross-domain eight-topic confusion matrix in Table 3b. The format of this matrix is the same as in Table 3a, except that the predictions are made in the target corpus (New Zealand Parliament speeches) and comparisons are made to our new human annotations. Overall, the results are quite encouraging about how well the model generalizes to the new corpus. Within each category, the correct class has by far the highest number of retrieved documents. In the economy topic, for example, the model correctly identifies 389 out of 720 speeches, corresponding to a recall of 0.540. The errors are somewhat evenly distributed, with the second-most-selected topic, political system, has 180 documents. In the economy column, the most frequent topic that is mis-construed as economy is also political system. This is likely because these topics are often discussed in the same speech.

The lowest recall of 0.371 (freedom and democracy) is very similar to the lowest recall of a topic in the within-domain model (0.422). The recall does not rise above 0.624, however. Because this data is at the speech level (rather than at the statement level, as in the source corpus), and speeches can touch on multiple topics, this relative decrease in performance is perhaps not too surprising. The tendency of misclassifications also looks different: while the within-domain model tends to categorize them as related to welfare, the cross-domain model tends to put documents into the political system category.

The relative distribution of predictions (Table 3b, bottom row) is also quite encouraging. The distribution is comparable to the distribution predicted by the within-domain model. But there are some interesting differences. For example, while economy is over-represented in the party platforms, the topic is under-represented in the parliamentary speeches.

For the 44-topic specification, confusion matrices are somewhat unwieldy, so we report the metrics in a table. In Table 4, each row is a topic, as indicated in the first column. Then there are two sets of columns corresponding to the within-domain and cross-domain classifiers. Within these column groups, the first column (N) denotes the number of documents (statements or speeches, respectively) in the annotated test set. The remaining columns indicate topic-specific accuracy – top-1, top-3, and top-5, respectively. As mentioned above, top-1 accuracy is equivalent to class-level recall. Finally, the right-most column (Ratio) reports the ratio of cross-domain top-1 accuracy to within-domain top-1 accuracy. The table is sorted by cross-domain top-1 accuracy, from highest to lowest.[3]

Overall, the 44-topic metrics produce a more mixed picture of our classifier's performance. Some topics have quite poor within-domain performance. For example, economic goals has 0.026 accuracy, while anti-imperialism has 0.048 accuracy. Yet some topics are highly distinctive and easy to classify: welfare state expansion has 0.772 top-1 and 0.940 top-3 accuracy, for example, while the corresponding figures for education are 0.769 and 0.905. The other topics are somewhere in between; the overall within-domain average accuracy is 0.538 (as indicated in Table 2).

Some of these poor-performing categories can be explained by the Manifesto Project's codebook choices. Topics like anti-imperialism (along with corporatism / mixed economy, keynesian demand

3. Appendix Section B.11 assesses the robustness of the topic-level within-domain metrics to sampling variation using a bootstrapping procedure. We do not observe notable differences in the mean/median recall or precision relative to the baseline reported values. The standard deviation in the metrics is sensitive to the frequency of the topics, however: infrequent topics have quite high standard deviations, especially in topic precision. Again, this finding cautions against using infrequent source-corpus topics for empirical analysis using cross-domain learning.

Table 4. Classifier performance with 44 topics

	Within-Domain				Cross-Domain				Ratio
	N	Top-1	Top-3	Top-5	N	Top-1	Top-3	Top-5	
Education	1,817	0.769	0.905	0.937	177	0.746	0.910	0.955	0.970
Law and order	1,310	0.698	0.879	0.931	158	0.715	0.892	0.943	1.024
Welfare state expansion	3,800	0.772	0.940	0.972	368	0.685	0.897	0.948	0.887
Political authority	1,016	0.460	0.720	0.829	775	0.570	0.831	0.895	1.239
Military	624	0.625	0.804	0.869	47	0.553	0.809	0.915	0.885
Environmental protection	1,504	0.696	0.880	0.924	90	0.522	0.756	0.867	0.793
Underprivileged minority groups	392	0.224	0.533	0.691	10	0.500	0.800	0.900	2.232
Agriculture and farmers	747	0.584	0.791	0.863	87	0.494	0.713	0.816	0.846
Internationalism	659	0.537	0.795	0.873	37	0.486	0.676	0.784	0.905
Culture	498	0.570	0.785	0.845	43	0.465	0.698	0.791	0.816
Democracy	684	0.449	0.725	0.808	305	0.449	0.748	0.856	1.000
Economic growth	823	0.473	0.693	0.790	104	0.404	0.673	0.837	0.854
Multiculturalism	433	0.446	0.704	0.790	103	0.398	0.631	0.709	0.892
Technology and infrastructure	2,152	0.704	0.901	0.941	113	0.398	0.628	0.788	0.565
Labour groups	904	0.596	0.823	0.888	188	0.383	0.681	0.766	0.643
Non-economic demographic groups	686	0.252	0.631	0.786	37	0.378	0.676	0.838	1.500
Nationalisation	184	0.435	0.614	0.668	32	0.344	0.531	0.656	0.791
Economic orthodoxy	475	0.516	0.716	0.781	136	0.331	0.566	0.728	0.641
Market regulation	852	0.421	0.678	0.799	114	0.298	0.553	0.719	0.708
Government and admin efficiency	1,037	0.453	0.754	0.856	191	0.267	0.681	0.796	0.589
National way of life	649	0.362	0.622	0.741	61	0.262	0.689	0.787	0.724
Equality	1,415	0.462	0.789	0.886	111	0.261	0.712	0.883	0.565
Protectionism	310	0.419	0.629	0.723	59	0.254	0.441	0.559	0.606
Centralization	773	0.405	0.682	0.788	52	0.250	0.654	0.712	0.617
Incentives	869	0.513	0.753	0.846	47	0.234	0.447	0.638	0.456
Traditional morality	442	0.391	0.701	0.805	67	0.194	0.388	0.522	0.496
Free market economy	427	0.237	0.487	0.623	73	0.096	0.233	0.397	0.405
Freedom and human rights	546	0.385	0.643	0.742	78	0.064	0.372	0.564	0.166
Political corruption	187	0.273	0.508	0.610	50	0.060	0.180	0.340	0.220
Civic mindedness	335	0.263	0.454	0.579	43	0.047	0.233	0.349	0.179
Constitutionalism	155	0.252	0.587	0.665	162	0.012	0.099	0.173	0.048
No topic	231	0.043	0.199	0.242	192	0.010	0.047	0.120	0.233
Anti-growth economy	581	0.234	0.719	0.811	13	0.000	0.308	0.462	0.000
Anti-imperialism	21	0.048	0.048	0.048	3	0.000	0.000	0.000	0.000
Controlled economy	124	0.306	0.500	0.589	9	0.000	0.444	0.556	0.000
Corporatism/mixed economy	45	0.044	0.111	0.200	11	0.000	0.000	0.000	0.000
Economic goals	234	0.026	0.192	0.295	7	0.000	0.000	0.000	0.000
Economic planning	147	0.116	0.272	0.422	1	0.000	0.000	0.000	0.000
Foreign special relationships	161	0.224	0.547	0.671	5	0.000	0.000	0.000	0.000
Keynesian demand management	39	0.077	0.154	0.205	1	0.000	0.000	0.000	0.000
Middle class and professional groups	82	0.293	0.463	0.500	3	0.000	0.000	0.000	0.000
Peace	115	0.409	0.539	0.600	2	0.000	0.000	0.500	0.000
European Union	324	0.586	0.775	0.830					
Marxist analysis	44	0.045	0.205	0.250					
Total	28,853	0.538	0.766	0.841	4,165	0.410	0.650	0.747	0.762

Notes. Rows are topics, and columns are metrics. *Within-domain* means a model trained on manifesto statements and evaluated on held-out manifesto statements. *Cross-domain* indicates the manifesto-trained model evaluated on newly annotated New Zealand parliament speeches. N denotes the number of documents that are hand-labeled as that category. Top-1, top-3, and top-5 are the accuracy metrics. Ratio is the cross-domain top-1 accuracy, divided by the within-domain top-1 accuracy.

management, Marxist analysis, and middle class and professional groups) are rare (at least in English-language party platforms) and could probably be folded into other, broader topics. Some topic pairs are difficult to distinguish semantically, such as economic goals vs. economic growth, and the machine classifier tends to fold the smaller category into the larger one. This type of subtle distinction is unlikely to play an important role in downstream empirical applications.

The cross-domain performance is slightly worse overall. The distribution of the gap between within-domain and cross-domain is shown in Appendix Figure A11. Unsurprisingly, any topic that the within-domain classifier could not categorize also performs poorly in the cross-domain classifier. We revisit this issue below as a potential diagnostic tool. Some topics (European Union and Marxist analysis) do not feature in the parliamentary speeches, so we cannot compute metrics for them. Ten topics, although infrequent, have zero accuracy. A number of other topics have quite poor performance, with the classifier even failing to rank the correct topic within the top-5 most of the time. These metrics demonstrate the importance of some target-corpus validation, as machine-coded data on these poor-performing topics should not be used for any empirical analysis.

If we limit to the top rows of the table (based on the sort), the cross-domain classification is quite good. The performance for education (top-1 accuracy = 0.746, top-3 = 0.910) is about the same as for the within-domain classifier. A handful of topics perform even better cross-domain than within-domain: law and order (top-1 = 0.715), political authority (top-1 = 0.570), underprivileged minority groups (top-1 = 0.500), and non-economic demographic groups (top-1 = 0.378). Of the 44 topics, seven are ranked first correctly at least half of the time, and all have good top-3 / top-5 accuracy. For 23 topics, the correct topic is ranked within the top-3 at least half of the time.

The relevance of the variation will depend on the downstream empirical task. Most of the categories with bad accuracy are quite rare. In our first application below, we focus on political authority – one of the better-performing topics for cross-domain learning.

4.3 Interpreting the Model Predictions

To increase our confidence that the model is properly identifying topics in the target corpus, we undertook further analysis to interpret the model predictions. First, we read the ten parliamentary speeches with the highest probability of belonging to each topic, using both the 44- and 8-topic specifications. In general, the speeches corresponded very well to the specified topics, and we saw no evidence that they were driven by correlated features. Appendix Section B.6 includes text snippets for each of these topics.

To more systematically analyze the connection between text features and predicted topics, we created a feature importance measure to identify which phrases are significantly correlated with topics in the source and target corpora. We use a simple metric computed from ordinary least squares (OLS) regressions. Formally, for each output topic class k and each N-gram feature j, we run a bivariate OLS regression model

$$\hat{p}_i^k = \alpha + \beta_{jk} x_i^j + \epsilon_i, \ \forall j, k \tag{1}$$

where $\hat{p}_i^k$ is the predicted probability that document i is about topic k and x_i^j is the relative frequency of N-gram j in document i. These regressions generate a dataset of coefficients $\hat{\beta}_{jk}$ and associated standard errors $\widehat{\text{s.e.}}_{jk}$, separately for the manifesto statements and the parliamentary speeches.

To identify statistically significant features for each topic, we compute the t-statistic $\hat{\tau}_{jk} = \hat{\beta}_{jk} / \widehat{\text{s.e.}}_{jk}$ in both the source and target corpora. Important features have a high (absolute value) t-statistic. To help further ensure that the features are interpretable, we constrain the vocabulary to a set of idiomatic noun phrases. This procedure is described in detail in Appendix B.7, which also includes word clouds depicting the top-ranked phrases and t-statistics.

Figure 1 illustrates the results of this analysis using scatter plots. Each plot focuses on one of the eight topics k. In each plot, the vertical axis indexes the t-statistic $\hat{\tau}_{jk}^{S}$ for k in the manifesto platform statements, while the horizontal axis indexes $\hat{\tau}_{jk}^{T}$ for k in the New Zealand parliamentary speeches. Each dot on the plot corresponds to an N-gram j, printed as a marker label. The vocabulary is filtered to the intersection of N-grams that are predictive for at least one topic in either corpus. Looking at the plots by topic, we find that there is a strong relationship in general between the t-stats in the source and target corpus. This is reassuring evidence that the content of the topics is similar in the manifesto and the parliamentary speech data. Hence, the topics in both data sources can be interpreted in a similar manner. The exception is no topic (Panel h), which intuitively would be less well-defined in terms of political language.

4.4 Diagnostic Tools for Cross-Domain Classification

Human annotation in the target corpus is costly, so it is useful to have preliminary diagnostic tools available using only within-domain metrics. We recommend two approaches to producing diagnostics without annotating any documents in the target corpus. While these diagnostics do not provide a clear rule regarding which analyses will work, they offer useful inputs that can be considered along with other relevant factors.

First, one can use the within-domain performance to predict cross-domain performance. As already mentioned, topics with poor within-domain accuracy also tend to have poor cross-domain accuracy (see also Yan *et al.* 2019). Appendix Figure A12 clearly shows that within-domain and cross-domain accuracy are highly correlated (correlation coefficient = 0.79). Therefore, within-domain metrics can help researchers assess which cross-domain topic measurements are likely to work well for empirical analysis.

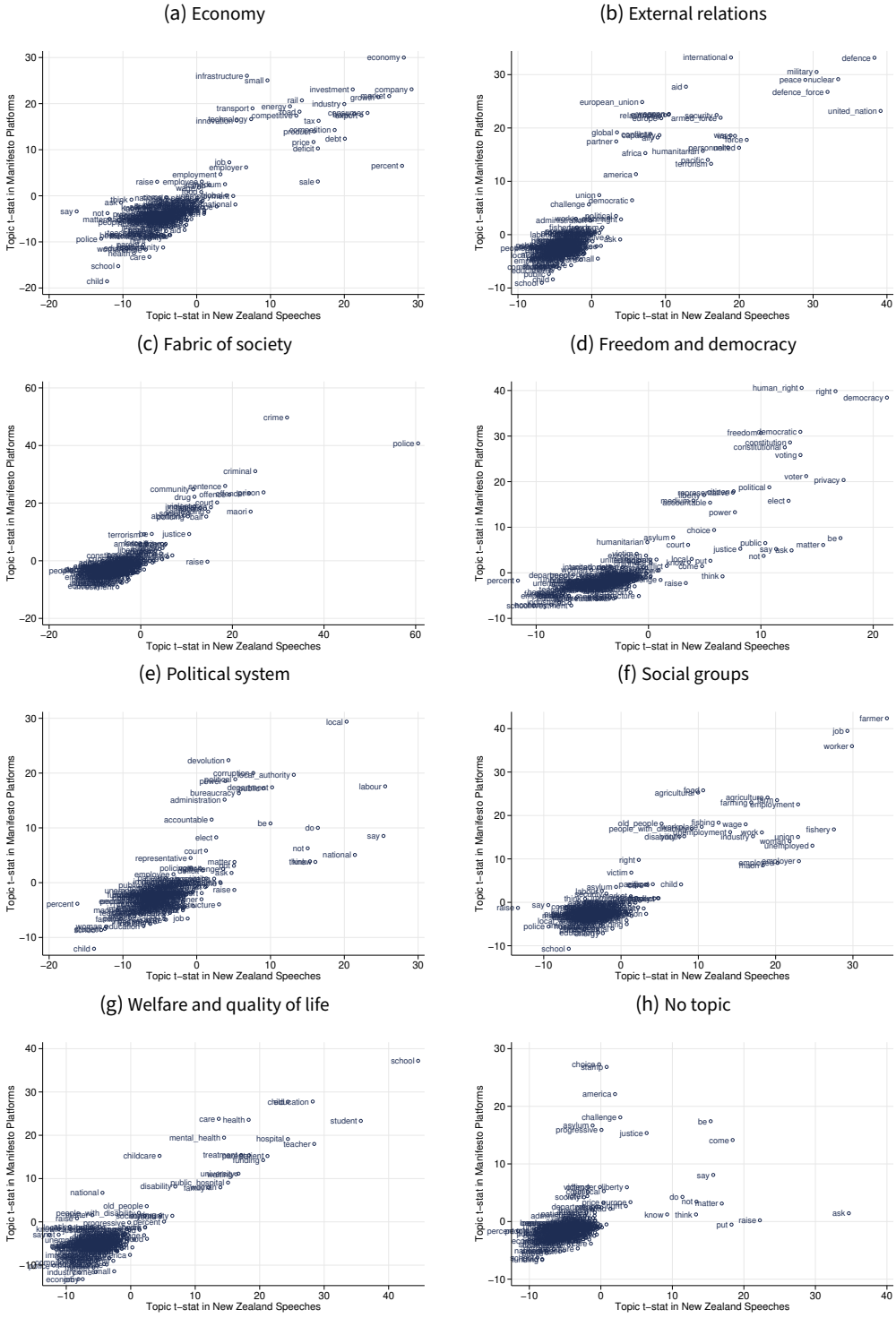
Second, we use the feature importance metrics from Section 4.3 to compute a measure of feature congruence by topic. Formally, we define the feature congruence of topic k as $F_k = \text{corr}(\hat{\tau}_{jk}^{S}, \hat{\tau}_{jk}^{T})$, where corr() is the Pearson's correlation and $\hat{\tau}_{jk}$ gives the estimated OLS t-statistic for N-gram j's frequency regressed on topic k's predicted probability (from the model trained on manifesto statements). As above, S and T indicate manifesto statements (source corpus) and parliamentary speeches (target corpus), respectively. This congruence measure is the correlation of the topic's feature importance weights from the cross-domain model with those from the within-domain model (at the N-gram level). A strong correlation indicates that the model predictions are associated with similar language in both domains.

Figure 2 illustrates the extent to which F_k can predict cross-domain performance. The vertical axis denotes the top-3 cross-domain classification accuracy, and the horizontal axis is the feature congruence measure. A dot corresponds to one of the topics with the associated label attached. We find that feature congruence is highly correlated with cross-domain classification accuracy (correlation coefficient = 0.73). For example, the education, welfare, and law/order topics have both high accuracy and high across-domain feature-importance congruence. Meanwhile, a number of topics (e.g. no topic, free market economy) have both low accuracy and low congruence.[4]

These diagnostic tools can be used to filter out topics to produce more precise overall measurements. For example, if we only keep topics that are above the median in top-3 within-domain accuracy and feature congruence, the top-3 cross-domain balanced accuracy increases from 0.44 to 0.68.

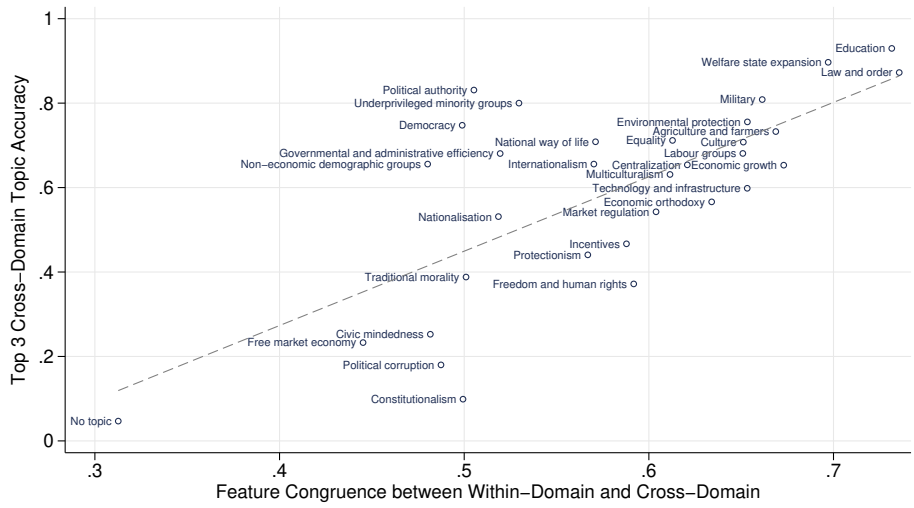
4. A qualitatively similar plot using the top-1 accuracy is in Appendix B.8. We also show a top-1 accuracy plot for the eight-class topic model.

Figure 1. N-gram correlations with topics for source and target corpus



Notes. Scatter plot for the eight topics, showing the t-stats of N-grams in the manifesto corpus (vertical axis) against the t-stat in the speech data.

Figure 2. Feature congruence and cross-domain classification accuracy



Notes. Scatter plot for the 44 topics, showing each topic's top-3 cross-domain classification accuracy (vertical axis) against the feature congruence, as defined in the text.

4.5 Inter-Coder Reliability and Application to other Countries

Appendix Section B.9 reports the results of our inter-coder reliability analysis. We find that the human coders agree with each other at about the same rate that our machine classifier agrees with the manifesto annotations. The cross-domain classifier accuracy does not vary substantially across human coders.

The classifier also works quite well in a (smaller) corpus of U.S. congressional speeches (see Appendix Section B.10). Using the classifier trained on 44 topics, we find that the top-1 accuracy is 0.440. When we predict eight-topics, the top-1 accuracy increases to 0.520. These numbers are comparable to those computed using the new hand annotations for New Zealand, and they suggest that the cross-domain classifier could work in other contexts besides the main application implemented here.

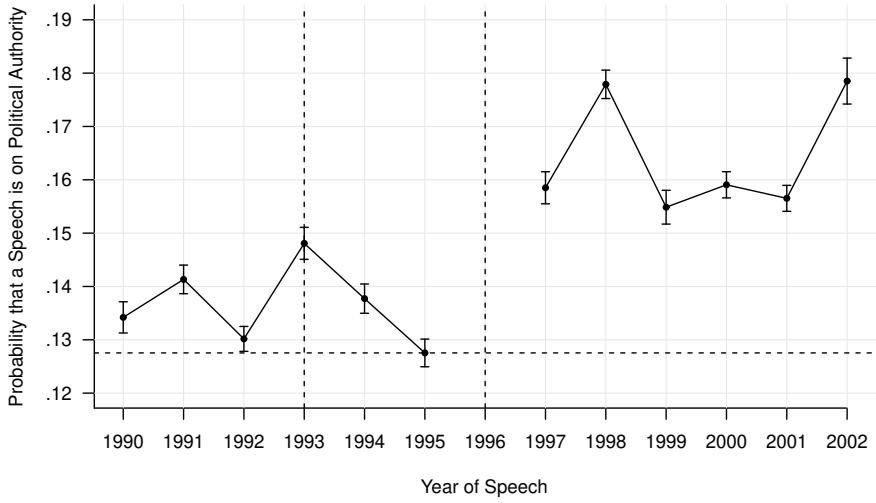
5 Empirical Applications

Cross-domain topic classification has broad potential scope for interesting empirical applications. In what follows, we illustrate two case studies. The first examines the effect of New Zealand's electoral reform on topics discussed in parliament. The second investigates the topics of parliamentary speeches broken down by speaker's gender.

5.1 Effect of Electoral Reform on Speech Topics

Previous research suggests that the 1993 electoral reform fundamentally changed the country's parliamentary practices. Unlike in the pre-reform period, afterwards parties had to form coalition and minority governments, which are generally associated with lower stability (Powell 2000; Vowles *et al.* 2002). Furthermore, parliamentary Standing Orders of the parliament were revised to explicitly reference parties and give them an important role in allocating speaking time (Proksch and Slapin 2014). At the same time, parliamentarians had little experience with proportional representation and had to adapt to the new system (Taagepera and Shugart 1989). Multiple parliamentarians also split from their party to form new parties. In light of the political developments described above, our main theoretical expectation is that the reform increased discussions about political authority,

Figure 3. Effect of electoral reform on political authority



Notes. Vertical dashed lines indicate the year that the reform passed (1993) and went into effect (1996). The horizontal dashed line indicates the outcome mean in 1995. The bars illustrate 95% confidence intervals.

which covers issues related to stability and party competence. Some example speeches on this topic are shown in Appendix Section B.6.

This application illustrates the usefulness of cross-domain learning in three ways. First, this analysis would be difficult to conduct using a lexicon approach because there are no established lexicons for this topic: issues of stability and competence are somewhat abstract and context dependent. The associated words in our target domain (Appendix Figure A9 Panel (f)) are specific to New Zealand and would produce many false positives (e.g. the party names, and verbs like "say" and "promise"). Second, this analysis cannot be conducted with an unsupervised topic model because of the specificity issue: topic models cannot target a particular dimension of speech, such as political authority. This topic would only appear (and therefore be measurable).

It would only be through good luck, and potentially significant manipulation of the features and model hyperparameters, that this topic would show up and therefore be measurable. Third and finally, the advantage over supervised learning is that one does not have to undertake expensive handcoding to label enough speeches on this topic.

Figure 3 plots the average probability that a speech focuses on political authority for the years 1990–2002. Note that the reform was passed in 1993 and the new rules were used for the first time at the 1996 elections. The figure illustrates a clear discrete increase after the reform, relative to beforehand. The average probability that a speech focuses on political authority increases by about 0.03 from a pre-reform baseline probability of 0.13. In Appendix Section C.2 we show that the result is statistically significant in a fixed-effects OLS regression framework. The results are also robust in a bootstrap analysis using multiple models trained on different subsets of the training data.

5.2 Effect of Gender on Speech Topics

A vibrant literature in comparative politics examines the policy content of parliamentary debates (Proksch and Slapin 2014). For example, Bäck and Debus (2019) study how speech topics vary according to the gender of the speaking parliamentarians. Using data from seven countries (Czech Republic, Estonia, Finland, Germany, Ireland, Norway, and Sweden), the authors assigned topics via manual coding of the debate segment title. They find that women parliamentarians talk less often about topics that are stereotypically associated with men.

The approach in Bäck and Debus (2019) is reasonable in their context, yet exemplifies the challenges of topic classification in parliamentary speeches. Manual coding is feasible in these countries. However, the manual coding is costly and debate segment information that allows the topics of all speeches to be inferred is not available in many databases and corpora. Coding topics at the speech level is an even more intensive effort. Lexicon-based approaches would not work because there is not an established/validated dictionary for measuring topics across languages. Similarly, unsupervised topic models would not work in this case because they cannot estimate consistent topics across multiple languages. Supervised learning is expensive because one would need to create a training dataset for each parliament.

The cross-domain learning approach is well suited to this setting, as manual annotation of documents is not necessary. The manifesto corpus is available in multiple languages and would allow the assignment of a consistent set of topics across parliamentary speeches in these languages. While multilingual cross-domain learning is beyond the scope of this paper, we illustrate its feasibility by extending the analysis from Bäck and Debus (2019) to New Zealand's parliament.

Specifically, we examine differences in the content of speeches by gender. Our results suggest that men tend to devote a greater share to topics such as external relations, which is in line with Bäck and Debus (2019). We also find that women devote a larger share of their speeches to talk about welfare (see Tables A17 and A18 in the 7).

6 Concluding Remarks

This paper has studied supervised learning for the cross-domain classification of political texts. This method combines the low cost of unsupervised topic models with the high interpretability and validatability of within-domain supervised classifiers. In an era of large and growing public annotated datasets, we expect that the applicability of this method will continue to expand. We have demonstrated how to use this method in the context of the manifesto corpus and parliamentary speeches from New Zealand. We used a multinomial logistic classifier to learn topics in the source corpus and predict topics in the target corpus. We showed how to validate the method using explanation methods in the target corpus, and more importantly, using human annotations of a subset of target corpus documents. To illustrate the empirical relevance of the method, we used our predicted topics to analyze the effects of New Zealand's 1993 electoral reform and debate participation.

We make three recommendations for scholars interested in applying supervised learning for cross-domain classification tasks. First, applied researchers should carefully assess whether the categories of the source corpus capture the concepts of interest. In other words, researchers need to consider whether they can test their theoretical expectations using the categories of the source corpus. Second, researchers need to determine whether the source and target corpora are similar enough to ensure good performance. They can inspect existing evidence on cross-domain classification to assess their application, and compare the predictive features, as done in Figure 1 above. Third, we recommend hand-coding a subset of the target corpus to assess model performance. This can be done using crowd-sourcing or expert coders. We suggest using cross-domain classification when the cross-domain performance is similar to within-domain performance.

Supervised learning for cross-domain classification has the potential to increase our understanding of political phenomena. An important advantage of this method is that the ability to estimate the same topics across documents and countries. Cross-domain supervised learning allows us to study how closely a manifesto's priorities match those in other documents such as speeches, party press releases, coalition agreements, legislative texts, and social media data (e.g., Thomson *et al.* 2017). For example, researchers may want to investigate whether populist parties act more or less in line with their manifesto priorities than non-populist parties. Moreover, cross-domain classification can be used to improve the performance of existing measures of policy

positions. For example, the tool could be used to distinguish ideological from non-ideological topics, which might improve the performance of existing methods (e.g., Slapin and Proksch 2008).

Future research may further improve the performance of cross-domain classification by using alternative models or coding schemes and by providing additional training to coders. The supervised learning algorithms used here do not directly take into account the different data distributions in the source and target corpora. Nor did we provide any special training to coders; we used the existing manifesto coding scheme, which was not developed for cross-domain text classification. As political scientists have invested significant resources in hand coding data, we hope that our work encourages further research on supervised learning and transfer learning in this field.

Funding

This work was supported by the European Research Council (advanced grant 694583).

Acknowledgements

For helpful comments and suggestions, we thank Kenneth Benoit, Amy Catalinac, Daniele Durante, Sara B. Hobolt, Michael Laver, Stefan Müller, Andrew Peterson, Sebastian Thieme, Matia Vannoni, Jack Vowles and our audiences at the ASQPS conference, the MPSA conference, the Berlin Social Science Center, Bocconi University, ETH Zurich, the London School of Economics and Political Science, New York University, the University of Essex and the New Zealand Parliament. We thank Matthew Gibbons, David Bracken, Pandanus Petter, Yael Reiss and Tove Wikelhut for annotating speeches. Samriddhi Jain, Linda Samsinger and Meet Vora provided excellent research assistance.

Data Availability Statement

The replication materials for this paper can be found at Osnabrügge, Ash, and Morelli (2021).

Supplementary Material

(This is dummy text) For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.xxxx.xx.

References

Anastasopoulos, J. L., and A. Bertelli. 2020. "Understanding Delegation Through Machine Learning: A Method and Application to the European Union." *American Political Science Review* 114 (1): 291–301.

Bäck, H., and M. Debus. 2019. "When Do Women Speak? A Comparative Analysis of the Role of Gender of Legislative Debates." *Political Studies* 67 (3): 576–596.

Barberá, P., A. E. Boydstun, S. Linn, R. McMahon, and J. Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–295.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Böhmelt, T., L. Ezrow, R. Lehrer, and H. Ward. 2016. "Party Policy Diffusion." *American Political Science Review* 110 (2): 397–410.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, and E. Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998.* Oxford: Oxford University Press.

Burscher, B., R. Vliegenthart, and C. H. De Vreese. 2015. "Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts?" *The ANNALS of the American Academy of Political and Social Science* 659 (1): 122–131.

Catalinac, A. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *Journal of Politics* 78 (1): 1–18.

Denny, M. J., and A. Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why it Matters, when it Misleads, and what to Do about it." *Political Analysis* 26 (2): 168–189.

Drutman, L., and D. J. Hopkins. 2013. "The Inside View: Using the Enron E-Mail Archive to Understand Corporate Political Attention." *Legislative Studies Quarterly* 38 (1): 5–30.

Géron, A. 2017. *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* Sebastopol: O'Reilly.

Greene, D., and J. P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25 (1): 77–94.

Grimmer, J. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18 (1): 1–35.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Hopkins, D. J., and G. King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Jones, B. S., and F. R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems.* Chicago: University Chicago Press.

König, T., M. Marbach, and M. Osnabrügge. 2013. "Estimating Party Positions across Countries and Time - A Dynamic Latent Variable Model for Manifesto Data." *Political Analysis* 21 (4): 468–491.

Krause, W., P. Lehmann, J. Lewandowski, T. Matthieß, N. Merz, and S. Regel. 2018. *Manifesto Corpus. Version: 2018-2. Berlin: WZB Berlin Social Science Center.*

Laver, M., K. Benoit, and J. Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311–331.

Lowe, W., and K. Benoit. 2013. "Validating Estimates of Latent Traits from Textual Data Using Human Judgement as a Benchmark." *Political Analysis* 21 (3): 298–313.

Mikhaylov, S., M. Laver, and K. Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20 (1): 78–91.

Miller, B., F. Linder, and W. R. Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." *Political Analysis* 28 (4): 532–551.

Osnabrügge, M., E. Ash, and M. Morelli. 2021. *Replication Data for: Cross-Domain Topic Classification for Political Texts.* https://doi.org/10.7910/DVN/Z0X7SU, Harvard Dataverse, V1.

Pearson, K., and L. Dancey. 2011. "Speaking for the Underrepresented in the House of Representatives. Voicing Women's Interests in a Partisan Era." *Politics & Gender* 7 (4): 493–519.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, *et al.* 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12:2825–2830.

Pennebaker, J. W., R. J. Booth, R. L. Boyd, and M. E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015.* Austin, TX: Pennebaker Conglomerates (www.LIWC.net).

Peterson, A., and A. Spirling. 2018. “Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems.” *Political Analysis* 26 (1): 120–128.

Powell, G. B. 2000. *Elections as Instruments of Democracy.* New Haven: Yale University Press.

Proksch, S.-O., and J. Slapin. 2014. *The Politics of Parliamentary Debate. Parties, Rebels, and Representation.* Cambridge: Cambridge University Press.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54 (1): 209–228.

Roberts, M. E., B. M. Stewart, and D. Tingley. 2016. “Navigating the Local Models of Big Data: The Case of Topic Models,” edited by R. M. Alvarez, vol. Computational Social Science, 51–97. Cambridge: Cambridge University Press.

Roberts, M. E., B. M. Stewart, D. Tingley, and E. M. Airoldi. 2013. “The Structural Topic Model and Applied Social Science.” *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.*

Slapin, J. B., and S.-O. Proksch. 2008. “A Scaling Model for Estimating Time-series Party Positions from Texts.” *American Journal of Political Science* 52 (3): 705–722.

Taagepera, R., and M. S. Shugart. 1989. *Seats and Votes.* New Haven: Yale University Press.

Tavits, M., and N. Letki. 2009. “When Left Is Right: Party Ideology and Policy in Post-communist Europe.” *American Political Science Review* 103 (4): 555–569.

Thomson, R., T. Royed, E. Naurin, J. Artés, R. Costello, L. Ennser-Jedenastik, M. Ferguson, *et al.* 2017. “The Fulfillment of Parties’ Election Pledges: A Comparative Study on the Impact of Power Sharing.” *American Journal of Political Science* 61 (3): 527–542.

Tsebelis, G. 1999. “Veto Players and Law Production in Parliamentary Democracies: An Empirical Analysis.” *American Political Science Review* 93 (3): 591–608.

Vowles, J., P. Aimer, J. Karp, S. Banducci, R. Miller, and A. Sullivan. 2002. *Proportional Representation on Trial.* Auckland: Auckland University Press.

Wilkerson, J., and A. Casas. 2017. “Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges.” *Annual Review of Political Science* 20 (1): 529–544.

Workman, S. 2015. *The Dynamics of Bureaucracy in the U.S. Government.* Cambridge: Cambridge University Press.

Yan, H., D. Sanmay, A. Lavoie, S. Li, and B. Sinclair. 2019. “The Congressional Classification Challenge: Domain Specificity and Partisan Identity.” *Proceedings of the 2019 ACM Conference on Economics and Computation,* 71–89.

Zirn, C., G. Glavaš, F. Nanni, J. Eichorst, and H. Stuckenschmidt. 2016. “Classifying Topics and Detecting Topic Shifts in Political Manifestos.” In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text,* 88–93. Dubrovnik, Croatia, July.

Appendix:
Cross-Domain Topic Classification for Political Texts

Moritz Osnabrügge*
Elliott Ash†
Massimo Morelli‡

*Durham University, School of Government and International Affairs, Durham, United Kingdom.
†ETH Zurich, Center for Law & Economics, Zurich, Switzerland.
‡Bocconi University, Department of Social and Political Sciences, Milan, Italy.

Contents

A Data

A.1 Manifesto Corpus Categories and Topics

We classify speeches to topics that are based on the Manifesto Project categories (Budge et al., 2001; Klingemann et al., 2006; Krause et al., 2018). In the following, we provide a list of manifesto categories. The detailed description of the manifesto categories can be found in the manifesto codebook.[1] The manifesto codebook categorizes the categories to seven main topics: external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society, social groups, and a no topic category. We also illustrate one example of a manifesto statement for each category.

External Relations

- per101 Foreign Special Relationships: Positive

  example: The United States government's policy is that there is one China, as reflected in the three communiqués and the Taiwan Relations Act.

- per102 Foreign Special Relationships: Negative

  example: Every responsible nation recognized this threat, and knew it could not go on forever.

- per103 Anti-Imperialism

  example: and Iraqis allowed their self-determination.

- per104 Military: Positive

[1] Manifesto Project Dataset. 2015. Codebook. https://manifestoproject.wzb.eu/down/documentation (accessed on November 10, 2017).

example: To guard against foreign involvement in our elections, we call for vigilance regarding online credit card contributions to candidates and campaigns.

- per105 Military: Negative

  example: and being nuclear free.

- per106 Peace

  example: We would work to support a negotiated settlement between Russia and Ukraine,

- per107 Internationalism: Positive

  example: support, through the United Nations framework, socially just reform, including democratic and economic reforms, in countries where governments are engaged in human rights abuses.

- per108 European Community/Union: Positive

  example: Roaming Charges within the EU: Fine Gael strongly supports the principle that roaming charges should be harmonised across the EU and we will work in Europe to speed up progress in this regard.

- per109 Internationalism: Negative

  example: The current Administration's way of dealing with these violations of world trade standards has been a virtual surrender.

- per110 European Community/Union: Negative

  example: They discriminate in favour of EU citizens and against the rest of the world.

Freedom and democracy

- per201 Freedom and Human Rights

  example: The rights of citizenship do not stop at the ballot box.

- per202 Democracy

  example: These goals require that we build a democratic developmental state capable of mobilising all sectors and boldly intervening in the economy in favour of workers and the poor.

- per203 Constitutionalism: Positive

  example: COPE will work to ensure full respect for our Constitution and the rule of law.

- per204 Constitutionalism: Negative

  example: You can vote to retain the old First Past the Post system, or you can vote for the Mixed Member Proportional system - MMP.

- per301 Decentralization

  example: The restructured Regional Development Australia network will navigate federal government funding programs, according to our fair share commitment, including a restructured Better Regions Program.

- per302 Centralisation

  example: A new relationship between the regional administrations and the National Government;

- per303 Government and Admin Efficiency

  example: allow no major reorganisation of local government in our first term in office.

- per304 Political Corruption

  example: These are the people who should be appointed to boards not political cronies.

- per305 Political Authority

  example: Unlike Labour, I'm not prepared to mislead the New Zealand public about the situation this country finds itself in.

Economy

- per401 Free Market Economy

  example: Moreover, the inflation tax is regressive

- per402 Incentives: Positive

  example: who maintain their homes well and have lower maintenance costs as a result.

- per403 Market Regulation

  example: Increasing costs of production affect the competitiveness of companies in domestic and international markets, while consumers have to pay higher prices.

- per404 Economic Planning

  example: Develop a national plan to address the impacts of climate change on rural communities and regional industries

- per405 Corporatism/Mixed Economy

  example: encouraging workplace reform through promoting the shared interests of employers and employees in building efficient, cohesive, profitable and competitive enterprises

- per406 Protectionism: Positive

  example: and the prohibition on overseas investment in housing will help stabilise house prices.

- per407 Protectionism: Negative

  example: We will introduce a Welsh Development Agency (WDA) for the 21st century, tasked with boosting Welsh trade.

- per408 Economic Goals

  example: Labour will improve the quality of information available on the Irish housing market by requiring that the selling price of all dwellings is recorded in a publicly available, national housing price database.

- per409 Keynesian Demand Management

  example: The measures we propose as alternatives to austerity will, by halting and reversing the cuts to public services, restore lost jobs and create new ones.

- per410 Economic Growth: Positive

  example: Only Federal Labor has a plan to ensure a prosperous future.

- per411 Technology and Infrastructure: Positive

  example: Roads to Recovery funding for local roads has been increased by 83 per cent, to $3.2 billion over the next four years

- per412 Controlled Economy

  example: It will be managed to protect NZ jobs, give certainty to farmers and manufacturers, and cut out ups and downs in the NZ economy.

- per413 Nationalisation

  example: Create a fully licenced state enterprise bank

- per414 Economic Orthodoxy

  example: Our plan can be entirely implemented within the ten-year transport budget set out in the draft 2015-2025 Government Policy Statement on Transport Funding.

- per415 Marxist Analysis

  example: Mana rejects so called market solutions to climate change and environmental degradation.

- per416 Anti-Growth Economy: Positive

  example: Until now, economic forces have seen it cheaper to throw away a cheap item rather than fix it.

Welfare and quality of life

- per501 Environmental Protection

  example: At present New Zealanders are amongst the highest per capita emitters in the world

- per502 Culture: Positive

  example: Over the next five years, we will continue to promote the development of a wide range of local sports facilities around the country, to improve participation and provide healthy social outlets for people of all ages.

- per503 Equality: Positive

  example: Meanwhile, to help the poorest students now, we will immediately restore maintenance grants.

- per504 Welfare State Expansion

  example: Ensure that staff ratios in aged residential care facilities are set at

appropriate levels for safety and care,

- per505 Welfare State Limitation

  example: It was not intended to make people dependent on the state, reward irresponsible or dysfunctional behaviour, or contribute to the breakdown of the two-parent family.

- per506 Education Expansion

  example: This new funding will not impact those students hoping to study in the humanities.

- per507 Education Limitation

  example: That represents more than 4 percent of GDP devoted to K-12 education in 2011-2012.

Fabric of society

- per601 National Way of Life: Positive

  example: This country is growing up, and I want it to see it owned and controlled by New Zealanders in every possible sphere.

- per602 National Way of Life: Negative

  example: and give Welsh speakers the right to use their language at all levels and for its status to be internationally recognised.

- per603 Traditional Morality: Positive

  example: The integrity of our country's foreign assistance program has been compromised by the current Administration's attempt to impose on foreign recipients, especially the peoples of Africa, its own radical social agenda

- per604 Traditional Morality: Negative

  example: The referendum on marriage equality was a historic victory for the rights of gay and lesbian people

- per605 Law and Order: Positive

  example: We want an acknowledgement that government is prepared to do everything necessary to protect the innocent and vulnerable from the reign of terror imposed on them by legions of brigands and thugs.

- per606 Civic Mindedness: Positive

  example: Every percentage drop represents an assault on the ties that bind us together,

- per607 Multiculturalism: Positive

  example: Since then we have seen our nation build on a diverse heritage of cultures that has contributed to our proud nation.

- per608 Multiculturalism: Negative

  example: Restructuring of the Maori Affairs portfolio and establishment of Te Puni Kokiri the Ministry of Maori Development: this has enabled the successful return of social assistance and training programmes to mainstream Government departments.

Social groups

- per701 Labour Groups: Positive

  example: It also means providing support that has been proven to work, like work experience placements that help them get a first foot on the career ladder.

- per702 Labour Groups: Negative

  example: We will tackle intimidation of non-striking workers;

- per703 Agriculture and Farmers: Positive

  example: We have huge ambitions for our farming industry:

- per704 Middle Class and Professional Groups

  example: We will reduce the tax burden for those on low and middle incomes Income tax and USC

- per705 Underprivileged Minority Groups

  example: It is unacceptable that a significant gender pay gap still exists in our economy.

- per706 Non-economic Demographic Groups

  example: Women in Politics: Fine Gael recognises that there needs to be a substantial increase in the number of women in politics.

For the 44-topic specification, we merged manifesto categories on the same topic, but different direction (positive/negative). For example, we combined the categories "per607 Multiculturalism: Positive" and "per608 Multiculturalism: Negative" to create one "Multiculturalism" topic.

For the 8-topic specification, we merged the categories to 8 larger topics following the manifesto codebook. These topics, denoted as domains in the codebook, are: external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society, and social groups. We also add the no topic category, which the manifesto project codes as "0". These are manifesto statements that could not be coded to the substantive categories.

A.2 New Zealand Parliamentary Data

We use the Hansard as a source to identify the parliamentary speeches. This document offers a verbatim record of parliamentary speeches.[2] The New Zealand in-house service to report on debates was established in 1867 and in 1899 the reports became "substantially verbatim" (Edwards, 2015, 8). The goal of Hansard is to give the public un-biased information on parliamentary speeches. The name Hansard has its origins in England. Thomas Curzon Hansard compiled the debates of the House of Commons. In England, the Parliament took control of the reporting in 1909 (Ralphs, 2009). The Hansard is an established source in political science and has mainly been used to study parliamentary speeches held in the House of Commons (Peterson and Spirling, 2018; Spirling, 2016). With few exceptions, scholars have not yet studied parliamentary debates in New Zealand using quantitative text analysis (e.g., Curran et al., 2018).

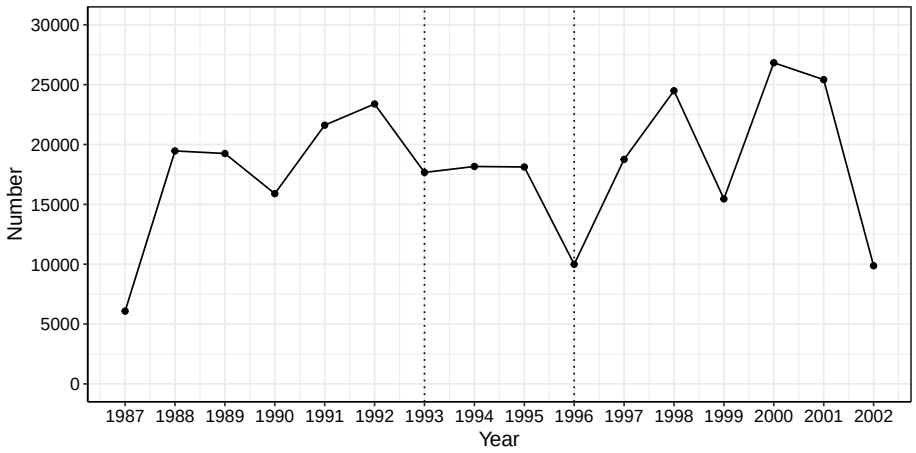
We access the data via the database provider *The Knowledge Basket*[3]. The data came in a series of zip archives, which included sets of files in HTML format. We built the data as follows. First, a set of Python scripts parsed the HTML and extracted the speeches along with corresponding meta-data, most importantly the speaker. We identified 437,865 speeches in total in the period from 1987 until 2002 and applied a set of filters as follows. We remove speeches from the Speaker and Deputy Speaker and study speeches that include at least 40 characters. We also removed short oral contributions by 'government member(s)' and 'opposition member(s)'. We identified and removed 31 speeches held in Maori language.

Our final corpus takes into account 290,456 oral contributions. Figure A1 illustrates the number of speeches by year. The number of speeches appears to be relatively

[2]Additional information can be found on the webpage of the New Zealand Parliament: https://www.parliament.nz/en/pb/hansard-debates/what-is-hansard/ (accessed on July 5, 2017).

[3]http://www.knowledge-basket.co.nz (accessed on July 7, 2017).

Figure A1: Number of speeches by year



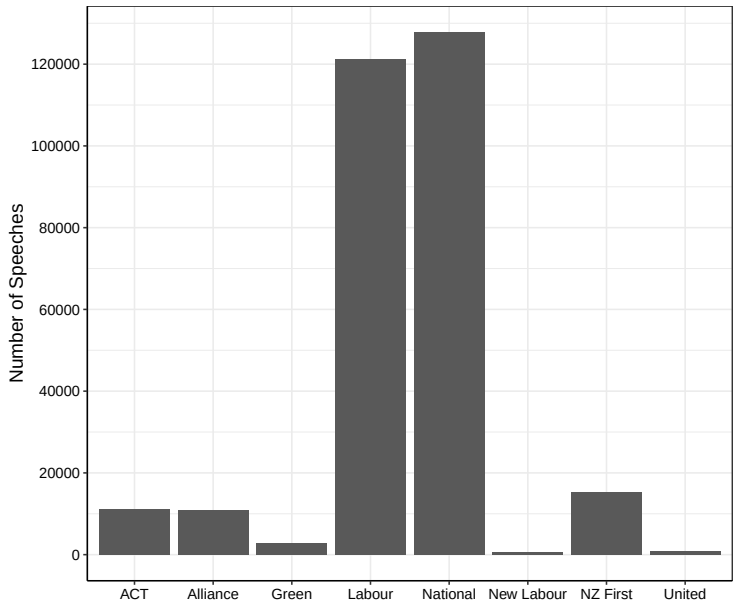
Notes. The dashed lines refers to the decision to introduce the reform (1993) and the first elections held under the new electoral system (1996).

constant over time with a slight increase after the 1996 electoral reform. We observe that in elections years the number of speeches is lower than in non-election years because the parliament has fewer sessions.

Figure A2 shows the number of speeches by party. We observe that the number of speeches correlates with the number of seats held in the New Zealand Parliament. The two parties with the largest number of speeches are the National and Labour parties followed by New Zealand First and ACT. Our data involves speeches from eight main parties: ACT, Alliance, Green, Labour, National, New Labour, NZ First, United New Zealand. The National party is a conservative and the ACT a liberal party. On the left side of the political spectrum, the New Zealand party system exhibits the Labour party as well as New Labour and the Alliance party. New Zealand First is a right-wing populist party and United New Zealand is located in the political center (Miller, 2005; Hayward, 2015, chapter 8). Note that we measure the party of parliamentary at the beginning of a legislative period.

We also collect data on individual-level characteristics of parliamentarians. We

Figure A2: Number of speeches by party



collect data from the Information Service of the New Zealand Parliament and the New Zealand Electoral Commission. Our meta-data includes information on the gender, election mechanism, party membership, prime minister position, committee chairmanship, amongst others. Table A1 provides descriptive statistics on the variables used in the regression analysis.

Table A1: Summary statistics of variables

Variable	Mean	Std. Dev.	Min.	Max.
Logarithm of political authority	-2.924	1.704	-18.654	-0.001
Political authority (indicator variable)	0.254	0.435	0	1
Reform	0.417	-	0	1
List parliamentarian	0.181	-	0	1
Female	0.173	-	0	1
Ethnicity minority	0.082	-	0	1
Prime Minister	0.026	-	0	1
Committee chair	0.123	-	0	1
Question	0.468	-	0	1
General debate	0.044	-	0	1
Administrative speech	0.013	-	0	1
Committee of the whole House stage	0.104	-	0	1
Government party	0.537	-	0	1
New parliamentarian	0.132	-	0	1

B Additional Material on Classification

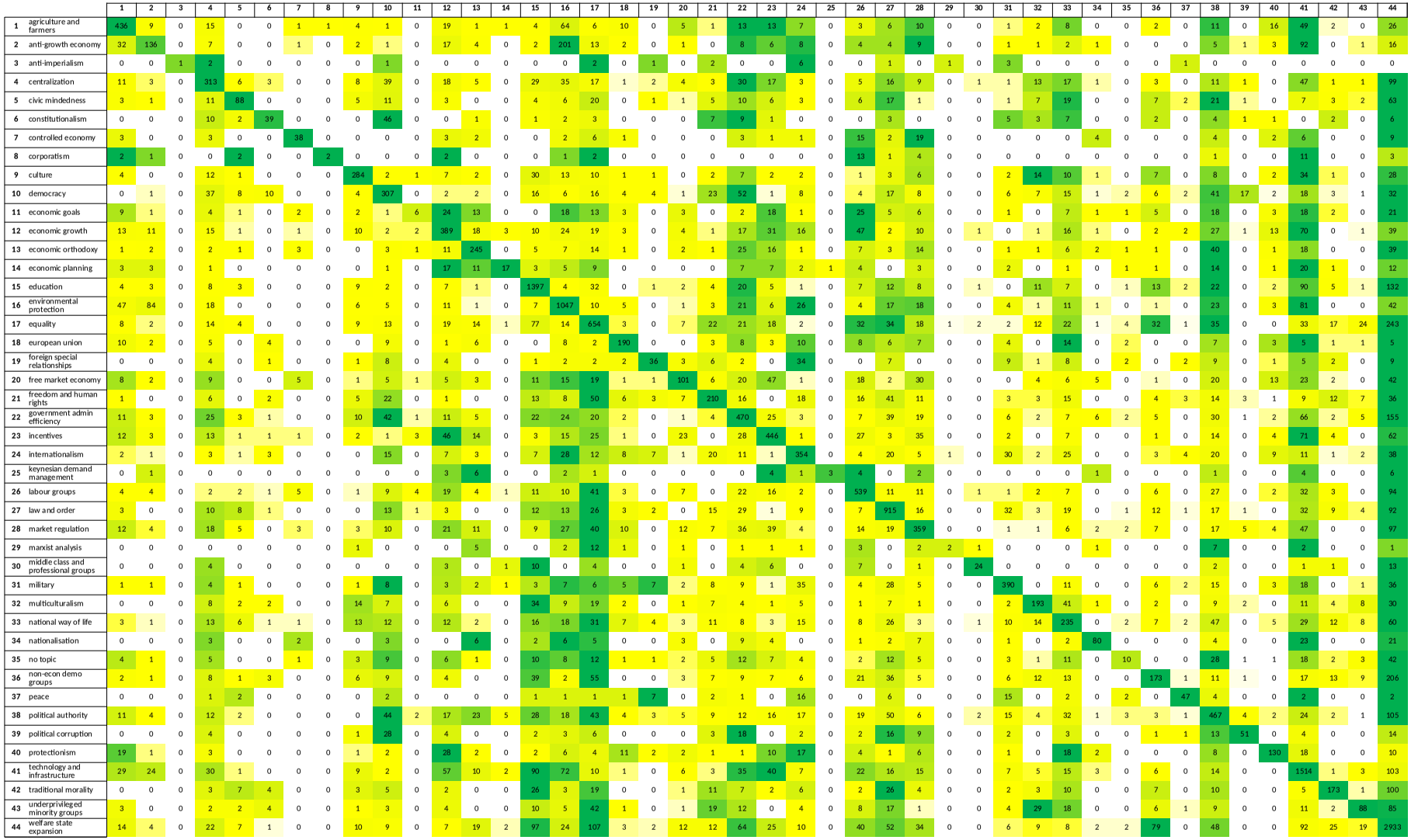
B.1 Within-Domain Classification: Confusion Matrix

Table A3 presents the confusion matrix of the 44 topics. The training and test set come from the manifesto corpus. The training set included 75% of the English quasi-sentences of the manifesto corpus and the training set encompassed 25% of the corpus.

B.2 Within-Domain Classification: Precision and Recall

As described in the manuscript, we train our classifier based on the manifesto corpus. Our training data includes 75% of the English statements of manifesto corpus and the held-out test sample 25%. Hence, the training and test data are based on the manifesto corpus. In the following tables we present the precision and recall by topic. Table A2 focuses on the 44-topic specification and Table A3 summarizes the metrics from 8-topic specification. Note that recall is the same as top-1 accuracy indicated in the confusion

Figure A3: Confusion matrix (44 topics)



matrices in the main text.

Table A2: Precision and recall (44 topics)

Topic	Precision	Recall
Agriculture and farmers	0.60	0.58
Anti-growth economy	0.43	0.23
Anti-imperialism	1.00	0.05
Centralization	0.45	0.40
Civic mindedness	0.52	0.26
Constitutionalism	0.48	0.25
Controlled economy	0.59	0.31
Corporatism	0.67	0.04
Culture	0.66	0.57
Democracy	0.43	0.45
Economic goals	0.27	0.03
Economic growth	0.47	0.47
Economic orthodoxy	0.55	0.52
Economic planning	0.50	0.12
Education	0.68	0.77
Environmental protection	0.59	0.70
Equality	0.45	0.46
European Union	0.64	0.59
Foreign special relationships	0.42	0.22
Free market economy	0.44	0.24
Freedom and human rights	0.47	0.38
Governmental and administrative efficiency	0.43	0.45
Incentives	0.52	0.51
Internationalism	0.53	0.54
Keynesian demand management	0.75	0.08
Labour groups	0.56	0.60
Law and order	0.61	0.70
Market regulation	0.48	0.42
Marxist analysis	0.40	0.05
Middle class and professional groups	0.71	0.29
Military	0.67	0.62
Multiculturalism	0.53	0.45
National way of life	0.35	0.36
Nationalisation	0.68	0.43
No topic	0.26	0.04
Non-economic demographic groups	0.42	0.25
Peace	0.63	0.41
Political authority	0.40	0.46
Political corruption	0.56	0.27
Protectionism	0.57	0.42
Technology and infrastructure	0.56	0.70
Traditional morality	0.56	0.39
Underprivileged minority groups	0.46	0.22
Welfare state expansion	0.56	0.77

Notes. The rows are topics and the columns refer to the precision and recall metrics.

Table A3: Precision and recall (8 topics)

Topic	Precision	Recall
Economy	0.67	0.74
External relations	0.73	0.64
Fabric of society	0.60	0.57
Freedom and democracy	0.57	0.44
No topic	0.50	0.03
Political system	0.52	0.42
Social groups	0.57	0.42
Welfare and quality of life	0.67	0.78

Notes. The rows are topics and the columns refer to the precision and recall metrics.

B.3 Number of Speeches by Policy Areas and Legislative Period

Table A4 summarizes the number of speeches by topic and legislative period. More specifically, the table captures the most likely topic using the results from the 44-topic specification. Table A5 summarizes the absolute frequencies for the 8-topic specification.

B.4 Alternative Text-Preprocessing

We also implemented the classifier using alternative text-preprocessing steps. In the following, we report one specific alternative text featurization. We use the Snowball stemmer to stem the manifesto corpus and focus on uni- and bigrams. We apply the tf-idf transformation and focus on n-grams that occur at least 10 times and not more than in 40% of the annotated manifesto data. We remove stopwords. Taking into account the 44-topic specification, the top-1 accuracy is 0.530, the F1-score (macro) is 0.407 and the balanced accuracy is 0.380. In case we use the 8-topic specification, the top-1 accuracy is 0.638, the F1-score macro corresponds to 0.515, the balanced accuracy is equal to 0.498.

Table A4: Number of speeches by topic and legislative period (44-topic specification)

topic	1987-1990	1990-1993	1993-1996	1996-1999	1999-2002	N
Agriculture and farmers	1787	1229	776	1199	914	5905
Anti-growth economy	77	178	67	57	154	533
Centralization	1132	987	720	640	882	4361
Civic mindedness	105	129	104	179	216	733
Constitutionalism	72	83	80	100	130	465
Controlled economy	44	19	12	20	22	117
Corporatism	2	0	0	0	2	4
Culture	425	588	428	318	388	2147
Democracy	7792	7760	5436	7619	7716	36323
Economic goals	27	22	16	17	13	95
Economic growth	1246	1531	839	697	1058	5371
Economic orthodoxy	1893	1241	826	1012	616	5588
Economic planning	13	21	13	8	26	81
Education	2436	3474	2738	2332	2892	13872
Environmental protection	1559	1789	1097	1247	1777	7469
Equality	1633	1900	1363	1682	1772	8350
European Union	104	72	82	54	58	370
Foreign special relationships	19	27	16	28	37	127
Free market economy	298	391	231	367	376	1663
Freedom and human rights	459	507	350	541	624	2481
Governmental and administrative efficiency	4548	4474	3033	3308	3402	18765
Incentives	975	640	251	553	563	2982
Internationalism	413	420	412	435	633	2313
Labour groups	1647	1813	984	1154	1798	7396
Law and order	2958	3158	2244	3305	3208	14873
Market regulation	1620	1430	1003	1683	1369	7105
Middle class and professional groups	0	6	2	5	3	16
Military	671	466	434	475	1106	3152
Multiculturalism	726	784	668	1093	469	3740
National way of life	798	807	803	681	938	4027
Nationalisation	645	251	188	381	296	1761
No topic	436	400	271	350	377	1834
Non-economic demographic groups	574	788	671	648	337	3018
Peace	11	12	15	28	27	93
Political authority	12439	15118	10735	17379	18012	73683
Political corruption	101	55	42	54	46	298
Protectionism	630	363	308	398	508	2207
Technology and infrastructure	1799	1896	1239	1041	1888	7863
Traditional morality	420	418	349	653	706	2546
Underprivileged minority groups	256	267	224	263	266	1276
Welfare state expansion	6037	8855	7142	6864	6555	35453

Notes. The rows are topics and the columns are legislative periods. The cells refer to the absolute number of speeches.

Table A5: Number of speeches by topic and legislative period (8-topic specification)

topic	1987-1990	1990-1993	1993-1996	1996-1999	1999-2002	N
Economy	12377	10796	6749	8439	9601	47962
External relations	1360	1104	1084	1128	1948	6624
Fabric of society	5674	6474	5245	7600	7014	32007
Freedom and democracy	6473	6576	4688	6888	7049	31674
No topic	83	70	53	83	107	396
Political system	18104	20290	14380	20606	21386	94766
Social groups	4261	4433	2862	3317	3389	18262
Welfare and quality of life	10495	14626	11151	10807	11686	58765

Notes. The rows are topics and the columns are legislative periods. The cells refer to the absolute number of speeches.

B.5 Alternative Classifier: Gradient Boosting

We implemented extreme gradient boosting using the xgboost and scikit-learn python modules (Chen and Guestrin, 2016; Pedregosa et al., 2011). We identified the hyperparameters using three-fold cross-validated randomized search. After training the classifier on the training data with 44 topics, we achieve a classification accuracy of 0.512, F1-score macro equals 0.402 and the balanced accuracy is 0.381. We proceeded in the same manner in the analysis of 8 topics. The classification accuracy increases to 0.634, the balanced accuracy is 0.496 and the F1 macro score is 0.516. The classification performance using three-fold randomized search is similar to the multinomial logistic regression model with L2 penalty.

B.6 Text Snippets of Speeches by Topic

In the following, we present text snippets of speeches by topics. These snippets come from speeches that are highly related to the topics. Table A6 focuses on the 44-topic specification and Table A7 on the 8-topic specification. In Table A6 we present multiple examples of speeches that are highly related to the political authority topic.

Table A6: Examples: 44 topics

Topic	Text
Agriculture and farmers	Some years ago there was a massive outbreak of eczema that depleted flocks and reduced farmers' incomes. That was followed by a drought of serious proportions that affected our flocks and herds of cattle. Then there was cyclone Bola—a disaster of unbelievable proportions. Even today, years after that event, farmers are still clearing up the mess, and, above all, are trying to pay the bills.
Anti-growth economy	Yes, it suggests two ways—progressive pricing and reduced fixed-line charges for the electricity supplied—but it basically does not rationally address the issue of getting greater levels of sustainable energy across the energy sector, nor does it make any suggestion to increase energy sustainability in the wider energy sense, such as relative to the percent of New Zealand's totally unsustainable energy requirement that is used by the transport sector.
Centralization	The Government insists that in local authorities with more than people ward systems will ensure that people in each local authority area will be represented. The Government will not have local authorities elected by just the more prosperous part of the community. It will have a system that will enable people to choose. Those people will not have to live in the area, but will need to be responsive to all of the communities in their area in order to function. Ward committees will be responsible for local interests in those particular areas.
Civic mindedness	The role of Business in the Community is to get the business sector involved in community projects. The project will not operate within the confines of a given community but will be a nationwide organisation. The organisation will be complementary to the efforts of community-based agencies such as the Hamilton enterprise agency. The organisation is supported by the community employment development unit. The person promoting the Business in the Community concept in New Zealand is Dr Graeme Craig of Hamilton, one of the founders of Woolrest International Ltd.
Constitutionalism	I rise to support the Bill. The issue is the shape and content of the constitution of this country, and I believe that nothing has changed since I studied constitutional law. As a young law student I was taught that the constitution resides with the people.
Controlled economy	What has the Government already done about the minimum wage?
Corporatism	The Government has received a welcome approach from Business New Zealand and the Council of Trade Unions to forge a tripartite social partnership to improve the quality and relevance of, and accessibility to, the workplace learning that is on offer. The Government respects the independence of both of our partners. We acknowledge that on some issues there are policy differences, but on the broad policy of workplace learning there is common ground. I welcome the approach by Business New Zealand and the Council of Trade Unions to work with the Government on those issues.
Culture	I move, That the Arts Council of New Zealand Bill be introduced. New Zealand's well-being is closely linked to the vitality of its cultural life. That is the reason that this Government supports the arts, as do all Governments in developed countries. It is now years since the Queen Elizabeth the Second Arts Council of New Zealand Act was passed.
Democracy	In little more than hours the people of the greatest democracy in the Americas, the United States, will vote for their new President. The man who will be responsible for leading the free world in the fight against socialism, and who will be responsible for ensuring the security of this planet, will be elected by, probably, one-eighth of the American people. In the United States—as in some other countries but certainly not in New Zealand or Australia—registration is not compulsory.
Economic goals	They show that unemployment rates in those regions have actually declined faster than in regions where unemployment rates were already rather lower. This Government has done exceptionally well in terms of lowering unemployment rates, both in deprived regions and amongst target groups in the community.
Economic growth	The OECD is optimistic about New Zealand's growth prospects. Its assessment is: "Over the coming years there will be a gradual but sustained growth in economic activity driven by exports, buoyant terms of trade, strong productivity gains, and a recovery in profits and business confidence."
Economic orthodoxy	What is the Reserve Bank's latest forecast for the Government's financial balance out-turn for the - fiscal year, and has he been advised whether it is consistent with his target of a Budget surplus?

Economic planning	The Alliance policy would be, first, to have an economic development strategy for New Zealand, so that we would be both beefing-up our social and economic infrastructures in the short term and developing a long-term plan. We would be looking at new technology-based industries.
Education	I recently visited a school in the West Coast area. I read the Education Review Office report prior to that visit, and I was taken aback by the comments made by the evaluator of the school about a particular teacher who, the report stated, was putting the education of the children at risk. She had an infant class, and the teacher who had those children after her had worked very hard every year for years to get those children up to scratch, up to the average level. She was going to resign because she was sick and tired of having to pick up after another teacher.
Environmental protection	I move, That the Wildlife (Penalties) Bill be now read a second time. I am pleased to be able to rise this evening to speak on my member's Bill, the Wildlife (Penalties) Bill, because its purpose is to increase the penalties for various offences under the Wildlife Act. But, more important, this Bill is about protecting the biodiversity of New Zealand. The Bill seeks to update the penalties for offences under the original Act, because the penalties have not been reviewed since . The Wildlife Act provides a protection system for wildlife, with some species absolutely protected and some partially protected.
Equality	How will those women receive pay justice after the abolition of the Employment Equity Act?
European Union	Is the Minister aware of the increasing centralisation of European trade issues in European Community bodies, such as the commission that the council administers and the European Parliament, and is it not Government policy to give priority to that area of our trade interests?
Foreign special relationships	I am delighted to bring the member up to speed, because the Government has been doing a lot of work in North Asia and on its relationship with North America. Sir Frank Holmes has been involved in that work.
Free market economy	It means that, if one holds on to a property and the property has a high capital cost— and one has allowed the person who is buying the property either to buy it or he or she is given that property free for the next years— there is a bonus in discounting in order to get the capital money in.
Freedom and human rights	I move, That the Human Rights Amendment Bill be now read a first time. The bill provides for a greater public sector accountability, in compliance with human rights obligations, and strengthens our human rights institutional framework and dispute-resolution procedures. This legislation is the long-awaited outcome of the Consistency audit project that was cancelled under the previous Government but revived under this Government.
Governmental and administrative efficiency	I move, That the Department of Justice (Restructuring) Bill be introduced. The Bill continues the ongoing process of restructuring the Department of Justice. That restructuring has six main elements. A new Office of Treaty Settlements will replace the old Treaty of Waitangi policy unit, which was set up within the department on 1 January 1995. The office will be responsible directly to the Minister in charge of Treaty of Waitangi Negotiations, but will organisationally be part of the new Ministry of Justice.
Incentives	Mineral mining companies receive taxation advantages by way of concessions and incentives. Current incentives provide for a deduction of exploration and developmental expenditure. A company may also qualify for a deduction in advance for the amount of exploration and development expenditure it expects to incur in the next two income years. The tax payable by mining companies in respect of their income is two-thirds of the amount payable by other companies.
Internationalism	Kia ora, talofa lava and warm Pacific Island greetings. The thoughts and prayers of all New Zealanders will be with the people of the United States as they start to come to terms with the full extent of the tragedy that has unfolded today. This is a time for all of us to reflect on the vulnerability of nations, security, and life. My prayers go out to all who are suffering at this time. Speakers before me have touched on the many issues that this act of terrorism has brought forward.
Keynesenian demand management	Will the Minister acknowledge that the increase in the goods and services tax was brought about by his inability to control spending, and that he cannot therefore visit his spending demands on others?

Labour groups	The real purpose of the Bill is to attack trade unions, and, specifically, to attack the New Zealand workers' union over the conduct of the union membership ballot. That is the excuse for the introduction of the Bill. The ballot was conducted under very difficult conditions, and I shall comment about that matter soon.
Law and order	I welcome this Bill and I am pleased that it has reached its second reading. It is appropriate to congratulate the Minister of Justice on the work that he has put into the Bill, with the support of the Minister of Police and members of the Government caucus, who have been right behind the Bill the whole way. The Bill tilts the balance of the justice system against the serious or persistent offender, and the rights of the public will now prevail over the rights of such offenders.
Market regulation	As I said, the Green Party remains very sceptical about industry self-regulation, given that no other country relies on it for the strategically important electricity industry. But we are supporting this bill, because we believe that it represents an improvement on the mess that the previous Government—Max Bradford, in particular—left us with. Self-regulation will lead to the major industry players wheeling and dealing around market rules to maximise their corporate benefit, and there are checks and balances to make sure that generators, retailers, and network companies hold each other to account in that process.
Middle class and professional groups	What other factors are impacting on low and middle income householders?
Military	I move, That this House take note of the Government's defence announcement. The Government today is announcing its defence plan to ensure that New Zealand has a modern, sustainable defence force matched to New Zealand's needs. Providing appropriately for the security and defence of the nation is a core responsibility of the Government.
Multiculturalism	The Government's current policy on providing Maori language and culture tuition in schools is clearly stated in Te Urupare Rangapu: "To provide for the Maori Language and culture to receive an equitable allocation of resources and a fair opportunity to develop having due regard to the contribution being made by Maori Language and culture toward the development of a unique New Zealand Identity.".
National way of life	I am one of the old-fashioned Tory members of Parliament in this House who believes in God. I support the constitutional monarchy. In my lifetime, I do not want to see the flag changed, or the national anthem altered or shortened. I do not want to see the name of the country changed. I think that this country has seen so much change in recent times. We have changed the electoral system, for the worse. We are abandoning the Privy Council, for the worse. The royal honours system is under siege, for the worse. The republican debate in this country is for the worse.
Nationalisation	Will the sale of Trans Power assets "amount to a privatisation because many of the supply companies were now privatised" as stated by the Alliance energy spokesperson; if not, why not?
No topic	The Leader of the House has not discussed the details of this matter with me.
Non-economic demographic groups	What specific measures does the Minister intend to take to assist Maori women, given that the rate of unemployment for pakeha women is about one-quarter of the rate for Maori women?
Peace	There has been mixed progress in the peace process since the Townsville peace agreement last year. The international peace monitoring team, to which New Zealand contributes personnel—and Mr Sowry will remember that that issue came before the House to be discussed—has worked effectively. It has helped to sustain confidence in the peace process.
Political authority	It did not simply want a change in the person to whom the salary of the Prime Minister was paid; it wanted a real change away from sleaze, away from incompetence, and away from the appalling track record of this coalition Government. The country has got none of those things. The first thing that Mrs Shipley did was to wait a month before she assumed the prime ministership; a month of further inaction, a month of a lack of direction. Then, instead of doing what the country wanted her to do, which was to say to her junior coalition partner that it was the junior partner and that its antics would not be tolerated any longer in this country, Jenny Shipley continued to do what she had accused her predecessor of doing—swallowing dead rats. The only difference was that she did not wash them down with whisky.

Political authority	I spent a lot of time doing that, like people over here, who now scoff when they say to themselves: "Hey, hold on. We argued about improving politics. We argued about honesty, and accountability, and trust." Think about it: weeks ago New Zealand First was saying that National was out; they were off; they were bad. Tau would not work with Shipley; would not be anywhere near a Government that had any of these National people in it. Now he says it is a little slip of the tongue and youthful exaggeration, and he is back in here. I think he has given politicians a bad name. I think he will have to work hard over the next little while to try to improve that situation after all the promises we all made about what was happening with MMP. I want to say something about the Alliance. The Alliance is not to blame for this deal not going through. The Alliance did exactly what it said it would do prior to the election. It stuck to its word.
Political authority	What an extraordinary speech! I wonder whether anyone out there listening to that said: "My word! That's the kind of man I want in Cabinet. We'll have to change the Government again and get him back.". I rather think that there were not any. Tonight we are discussing the last Budget of a dead Government and of a dead party—and I underline the words "dead party". When one looks across at the other side of the House, my goodness, there is not a wriggle left in them.
Political authority	I apologise. An objective observer has to say that conditions are fundamentally different from how they were when this minority National Government won its first confidence vote. At that time it looked very unstable, and that is the difference. The difference is stability. Ever since National made the disastrous mistake of doing a coalition deal with Mr Peters and New Zealand First, the Government looked unstable. The ACT party, for years, had been saying to National that there would be no stability while Mr Peters had his hands on the Treasury. ACT said to National: "Sack Mr Peters. Govern as a minority party." We said that if National returned to its own policies, we would support a minority Government from the cross benches. ACT predicted that a minority National Government would be more stable, and that is exactly what has happened. It is stability that is the difference. Investors hate instability. People who owe money on mortgages hate instability. Instability is bad for business—and I say to Mr Peters that people hate members interjecting while they are walking around the House. Instability is also bad for jobs. Stability is delivering New Zealand lower interest rates, just as ACT predicted. We predicted that if National sacked Mr Peters, then interest rates would fall, and that is exactly what happened. This summer has demonstrated that a minority National Government, supported by ACT holding the balance of power, can deliver stability. Some things, of course, have not changed. The Labour Party remains negative. I have seen seven Opposition leaders in my time but I have never seen a leader as relentlessly negative as Helen Clark. She must take lessons when she is in Africa. How could anybody be so negative, day in, day out? It could get into the Guinness Book of Records. She does not have a positive word to say about anything. It is all negative, negative, negative. The only plan Labour had was to hold a snap election so that it could get elected without any policy. The ACT party, by saying that we are opposed to instability and opposed to the "Italianisation" of New Zealand politics, has thwarted her plans.
Political authority	Helen Clark's political adviser Judy Keall has just confirmed to this House that Jim Anderton will never be the Treasurer, but I do not think that she knows. She did not know before the o'clock news that Helen Clark would spring this on her. This Alliance is a party of convenience that started with members of Parliament and is now down to members. They will take their places at the Cabinet table under a Labour-Alliance Government.
Political corruption	No, they are not. The bribery and corruption referred to in the Crimes Act is the bribery and corruption as done by State servants. State servants are barred by this Act from accepting bribes in relation to their work or from corruption in relation to the Official Information Act.
Protectionism	They say that they are in favour of balance in their protection policy, but not yet. The Opposition wants to go to heaven, but is not prepared to die. This agreement is far from dying. The agreement has exceeded the expectations the Government had when it went into the negotiations. By July we will have established the world's most comprehensive free-trade area for goods, and the agreement that the Government reached on services goes beyond any other bilateral agreement in the world.

Technology and infrastructure	The first stages of a number of key projects are progressing well. Developments such as the Grafton Gully connection to the port, the North Shore busway, and the "Spaghetti Junction" improvements should be ready for construction to start over the next months.
Traditional morality	They can't even agree on moral issues: homosexuality and abortion.
Underprivileged minority groups	What has he done to ensure that Pacific people's immigration concerns are being addressed?
Welfare state expansion	Today, I want to talk about the key commitments that we have made in health. We came to office with seven key pledges that we made to the people of New Zealand, and one of those pledges was to focus on patients, not profits, and to cut waiting time for surgery.

Notes. The cell entries include selected text snippets that are highly related to the 44 topics.

Table A7: Examples: 8 topics

Topic	**Text**
Economy	In her statement on Tuesday the Prime Minister talked of the need for economic transformation. In discussing that, she talked further of the need for exports, modernisation, and expanding tourism. They are all very worthy ideals that we would support. There was not one word in that statement—and I have not heard one word from the Government since—about the transport systems that will be required to deliver those goals. Will these exports that will help to transform our economy all be moved around by e-commerce?
External relations	The Minister is very fiery, and he knows that morale among the troops is not high. I accept that the present Minister of Defence—perhaps unlike the former Minister of Defence—is interested in the troops.
Fabric of society	I move, That the Criminal Investigations (Blood Samples - Burglary Suspects) Amendment Bill be now read a first time. This bill will add a new weapon to the arsenal of the New Zealand Police to gather evidence against burglars, secure convictions, and expand DNA testing to burglary suspects. This bill will change the law to give the police the power to take compulsory DNA samples from burglary suspects in order to secure convictions.
Freedom and democracy	I welcome the debate on electoral reform that has been going on throughout New Zealand for the past year or so. I think that it is very good for the democratic process that people learn a lot more about their Parliament and how its members are elected. Indeed, the debate has enabled them to do that, although until now most of it has related to support for a change in the way in which members are elected, rather than an explanation of the present first-past-the-post system.
No topic	The Leader of the House has not discussed the details of this matter with me.
Political system	I shall just refer to the article to refresh my memory. That is permitted under the Standing Orders. The Contractors Federation (Inc.) said that it complained in a letter to the honourable member for Kaimai, who is the chairman of the select committee that the federation appeared before. That letter stated that the federation knew that its briefing to the select committee would be a circus and a farce and would have no effect on the decisions the Government intended to make.
Social groups	Yes, of all the farm workers throughout the country participated in the ballot. That is democracy Labour Government style. Even the Minister of Agriculture would acknowledge that if only workers out of participate in a ballot it is logical to assume that there is obviously no dissatisfaction on the part of workers with their conditions of employment.
Welfare and quality of life	Yes, it has and I shall talk a little about it. Teacher salaries make up about percent of total expenditure on education, so there is little to save unless one starts to hit teacher salaries. The Minister of Labour, the Minister of Education, and the Associate Minister of Education have all said that they want to use the Employment Contracts Act to slash teachers' wages and to reduce spending on education.

Notes. The cell entries include selected text snippets that are highly related to the 8 topics.

B.7 Interpreting Topics

We examine the phrases that are positively correlated to the topics in the data of manifestos and speeches. Note that we focus on the test set of the manifesto data. We then apply the following methodological approach to both corpora.

First, we extract informative phrases from the text (see e.g., Handler et al., 2016; Denny and Spirling, 2018). This approach allows us to recognize key phrases such as "new zealand", and treat them as single tokens. In addition, we implemented the following text-preprocessing techniques to reduce the computational resources needed to calculate the word clouds. We tag parts of the text and identify phrases with up to four words using tag patterns, which results in a collection of noun phrases (and a few verb phrases).[4] We remove upper-case and punctuation and then lemmatize the tokens to remove uninformative word endings. We filter out rare sequences that appear in fewer than 20 speeches fewer than 30 times in total. We rank the phrases by their relative collocation (point-wise mutual information) to get key phrases.[5] We filter out a set of policy-irrelevant words (e.g., names).

Then we calculate t-statistics on the correlation between phrase frequency and topic probabilities in both corpora. We produce scatterplots comparing the t-stats of the n-grams appearing in both corpora. The scatterplots focus on n-grams that have

[4]We use the following tag patterns: 'A', 'N', 'V', 'P', 'C', 'D', 'AN', 'NN', 'VN', 'VV', 'NV', 'VP', 'NNN', 'AAN', 'ANN', 'NAN', 'NPN', 'VAN', 'VNN', 'AVN', 'VVN', 'VPN', 'ANV', 'NVV', 'VDN', 'VVV', 'NNV', 'VVP', 'VAV', 'VVN', 'NCN', 'VCV', 'ACA', 'PAN', 'NCVN', 'ANNN', 'NNNN', 'NPNN', 'AANN', 'ANNN', 'ANPN', 'NNPN', 'NPAN', 'ACAN', 'NCNN', 'NNCN', 'ANCN', 'NCAN', 'PDAN', 'PNPN', 'VDNN', 'VDAN', 'VVDN'. A: adverbs and adjectives, C: conjunctions, D: pronouns, N: nouns, P: prepositions, V: verbs.

[5]This is done by calculating the geometric mean of the pointwise mutual information criterion, since this metric can be calculated using the absolute rather than the relative frequencies. We set the following minimum levels: bigrams: 0.004, trigrams: 0.003, quadgrams: 0.002.

a high t-statistics in either the manifesto or parliamentary speech data: we created a new standardized variable that captures the largest t-statistics for each n-gram and focus on all n-grams above the 90th percentile.

To provide further evidence, we provide in the following word clouds on the phrases with the highest t-statistic in the parliamentary speech data. Figure A4 refers to the phrases that are positively related to the eight topics. The first topic is economy and is associated to words such as company, industry, market, bank and economy. The second topic is external relations and is positively related to words such as military, defence, nuclear and peacekeeping. The topic fabric of society is related to words like offence, policy, prison, ciminal, crime and prison. The next topic is freedom and democracy and includes words such as democracy, right, voting system, voting, clause and right. The topic political system is highly related to words such as national, labour, local, authority, leadership. Furthermore, multiple verbs on political action such as do or want are related positively to this topic. The topic is also associated with the words leadership, vision and politics. The topic social groups is associated with employment, famer, fishery, worker, employee and fishery union. The words patient, student, school, education, parent, hospital are highly related to the topic welfare and quality of life. Finally, the no topic category relates to words that are not related to the previous topics, such as prepared, raise, ask, agree, matter and be.

Figures A5, A6, A7, A8, A9, and A10 illustrate word clouds with phrases that are positively related to the probability that a speech belongs to one of the 44 topics. The word clouds summarize the phrases with the largest t-statistic.

Figure A4: Wordclouds of manifesto topics (8 topics)

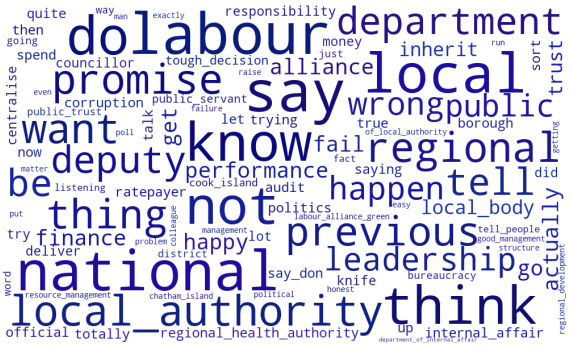
(a) economy

(b) external relations

(c) fabric of society

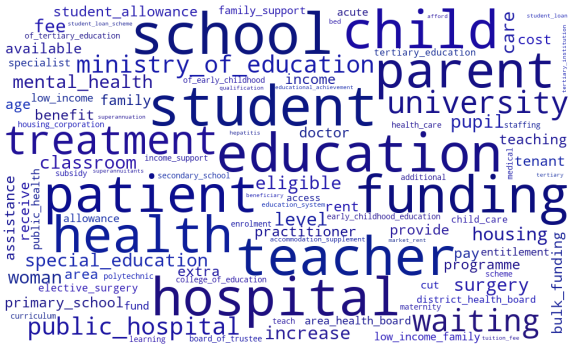
(d) freedom and democracy

(e) political system



(f) social groups

(g) welfare and quality of life

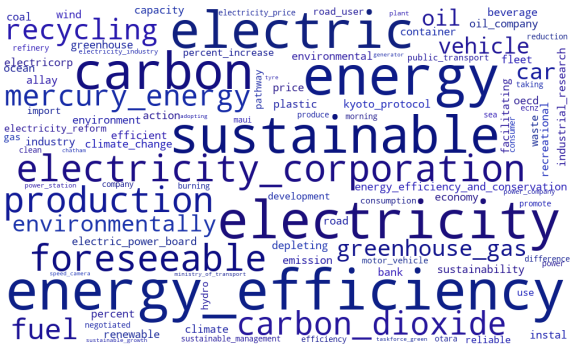


(h) no topic

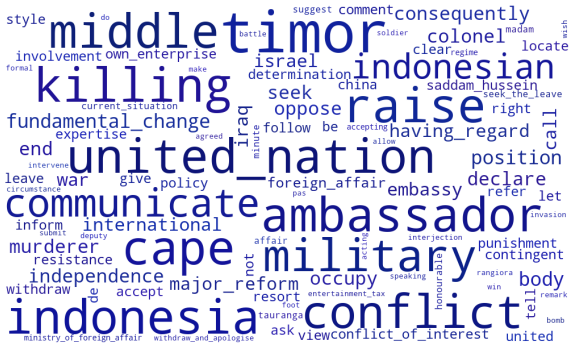
Figure A5: Wordclouds for topics I (44 topics)

(a) agriculture

(b) anti-growth economy



(c) anti-imperialism

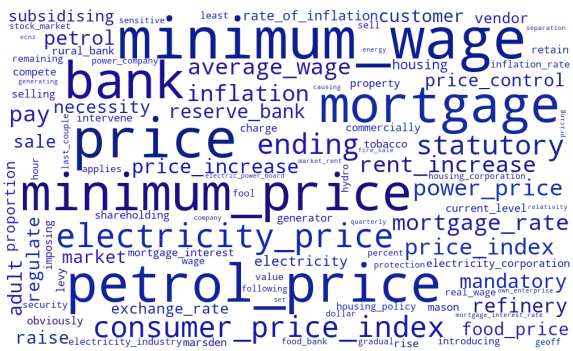


(d) centralization

(e) civic mindedness

(f) constitutionalism

(g) controlled economy



(h) corporatism

Figure A6: Wordclouds for topics II (44 topics)

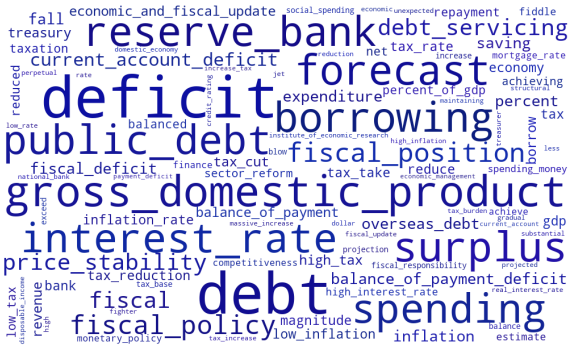
(a) culture

(b) democracy

(c) economic goals

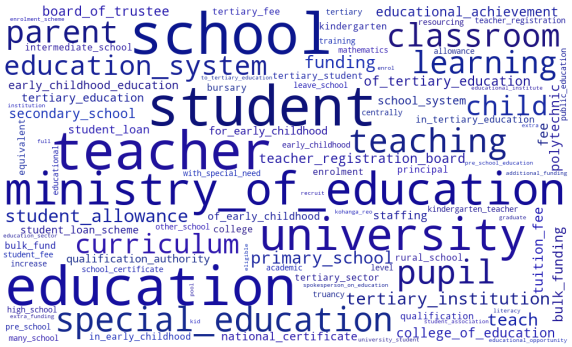
(d) economic growth

(e) economic orthodoxy



(f) economic planning

(g) education



(h) environmental protection

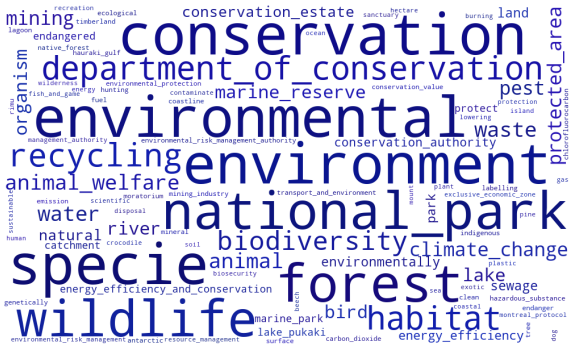


Figure A7: Wordclouds for topics III (44 topics)

(a) equality

(b) european union

(c) foreign special relationships

(d) free market economy

(e) freedom and human rights

(f) governmental and administrative efficiency

(g) incentives

(h) internationalism

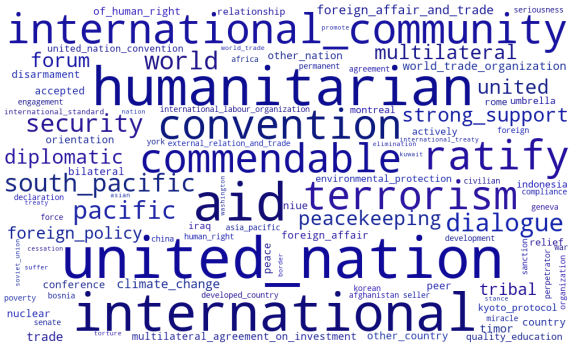


Figure A8: Wordclouds for topics IV (44 topics)

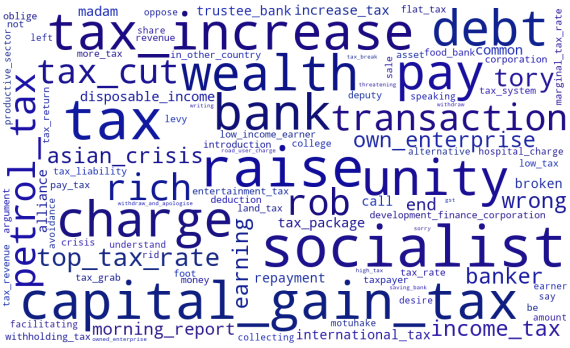
(a) keynesian demand management

(b) labour groups

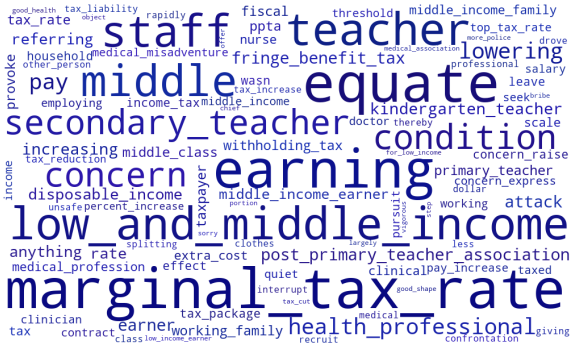
(c) law and order

(d) market regulation

(e) marxist analysis



(f) middle class and professional groups



(g) military

(h) multiculturalism

Figure A9: Wordclouds for topics V (44 topics)

(a) national way of life

(b) nationalisation

(c) no topic

(d) non-economic demographic groups

(e) peace

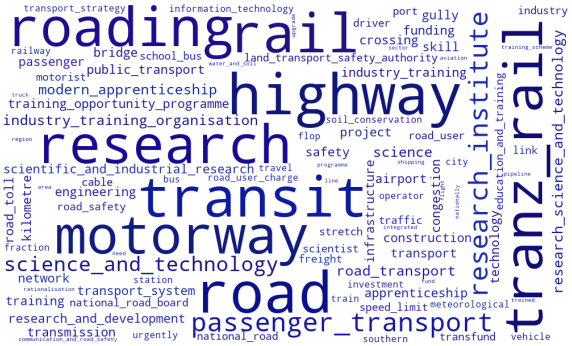
(f) political authority

(g) political corruption

(h) protectionism

Figure A10: Wordclouds for topics VI (44 topics)

(a) technology and infrastructure

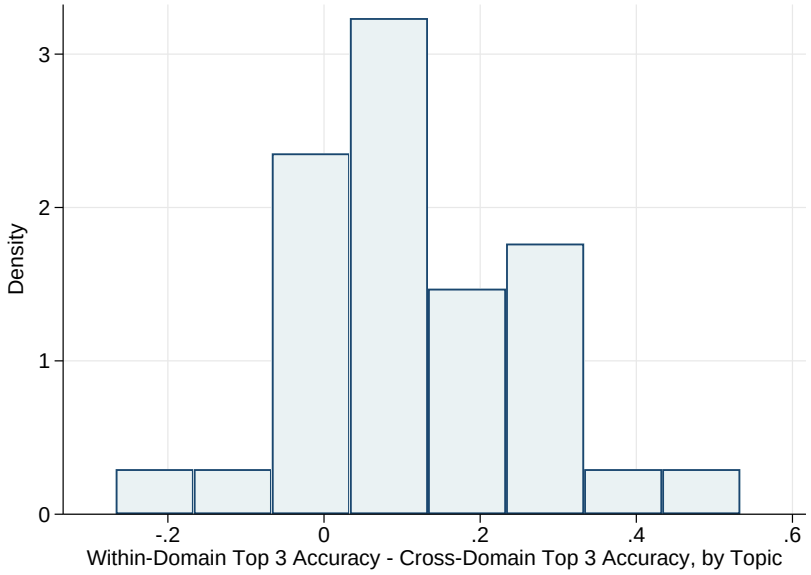


(b) traditional morality

(c) underpriviliged minority groups

(d) welfare state expansion

Figure A11: Distribution of difference between within-domain accuracy and cross-domain accuracy

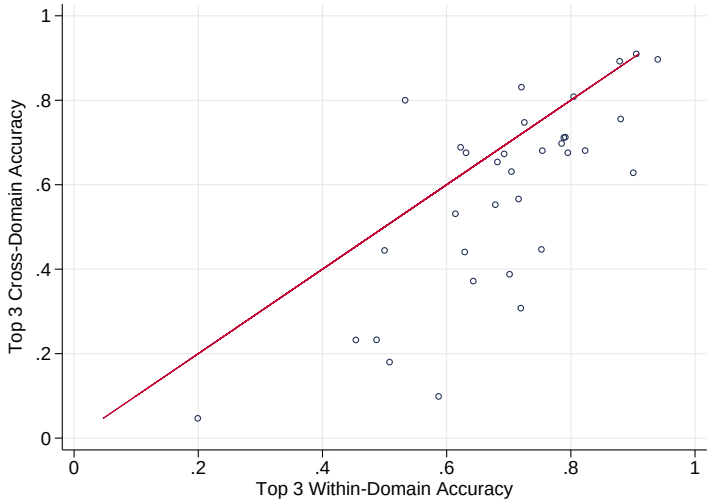


Notes. Histogram by topic of within-domain accuracy minus cross-domain accuracy. Positive values mean that cross-domain accuracy tends to be lower.

B.8 Diagnosing Cross-Domain Classification

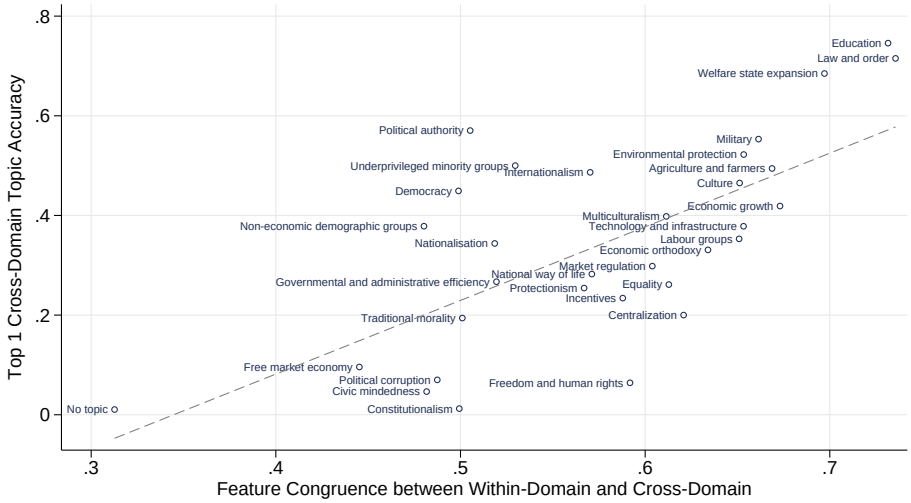
This section reports figures on diagnosing cross-domain accuracy using within-domain accuracy and feature congruence. Figure A11 shows the distribution of differences between the within-domain and cross-domain accuracy. This is mostly positive, reflecting that accuracy decreases when going cross-domain. Figure A12 shows that within-domain and cross-domain accuracy are highly correlated (correlation coefficient = 0.77). Figure A13 plots feature congruence against top-1 accuracy in the 44-topic specification. Figure A14 shows the analogous figure for top-5 accuracy. Finally, Figure A15 plots feature congruence against top-1 accuracy in the 8-topic specification.

Figure A12: Within-domain accuracy predicts cross-domain accuracy by topic



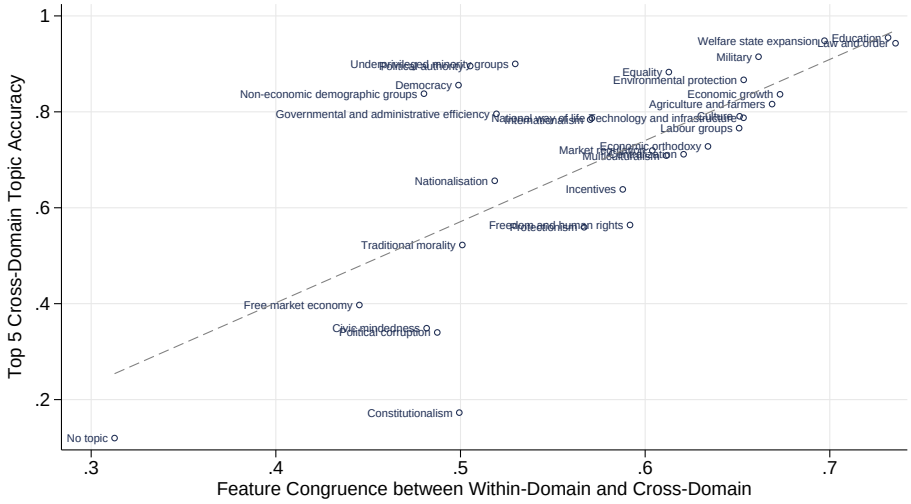
Notes. Scatter plot for the 44 topics, showing the topic's top-3 cross-domain classification accuracy (vertical axis) against the top-3 within-domain accuracy (horizontal axis). The red 45-degree line shows the relationship if within-domain and cross-domain accuracy were equal.

Figure A13: Feature congruence and cross-domain accuracy, 44-topics, top-1 accuracy



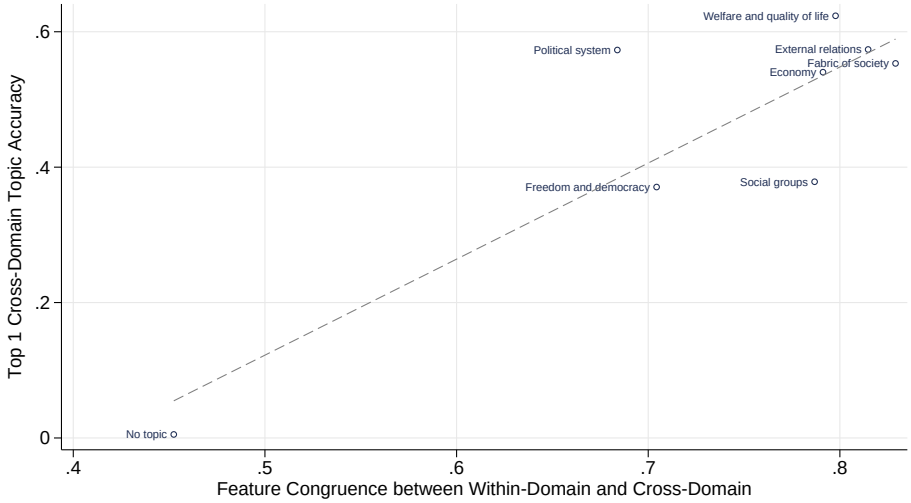
Notes. Scatter plot for the 44 topics, showing the topic's top-1 cross-domain classification accuracy (vertical axis) against the feature congruence, as defined in the text.

Figure A14: Feature congruence and cross-domain accuracy, 44-topics, top-5 accuracy



Notes. Scatter plot for the 44 topics, showing the topic's top-5 cross-domain classification accuracy (vertical axis) against the feature congruence, as defined in the text.

Figure A15: Feature congruence and cross-domain accuracy, 8-topics, top-1 accuracy



Notes. Scatter plot for the 8 topics, showing the topic's top-1 cross-domain classification accuracy (vertical axis) against the feature congruence, as defined in the text.

B.9 Inter-coder Reliability

To understand how useful the target-corpus annotations are in assessing the validity of the classifier, we would like to have some sense of the error rate in the human codings. As Mikhaylov et al. (2012) show, the coder reliability of the manifesto project is relatively low.

To check this in our context, we hired three additional coders.[6] Like the main coder, these coders also received training from the manifesto project in English-language platforms. They were not experts on New Zealand politics, however.

For this secondary annotation step, we drew a random sample of 250 speeches from the 4,165 speeches annotated by the first coder. Each of the three secondary coders annotated these same speeches, so that we had four annotations in total. We did not give the coders detailed guidelines, but asked them to code in line with their training from the manifesto project.

In Table A8, we compare the coding of these three secondary coders to the machine predictions and the annotations of the primary coder. The upper part of the table shows results for 44 topics, while the lower part shows results for 8 topics. In the first row, we find that in total the coders classified 37 percent of the speeches to the same category as the machine. If we focus on eight topics the accuracy is 50 percent, which is nearly identical to the accuracy that we obtained comparing the coding of the main coding to the machine predictions. In the next step, we compare the coding of the three coders to the coding of our main coder. In 50 percent of the speeches, the coders agreed on the same topic and this share increase to 63 percent if we limit ourselves to eight topics.

[6] We thank Pola Lehmann for providing us the contact details of the manifesto coders.

0

Table A8: Reliability of coding

	Coder 1	Coder 2	Coder 3	Total
44 Topics				
Machine Predictions	0.364	0.388	0.368	0.373
Main Coder	0.452	0.528	0.508	0.496
8 Topics				
Machine Predictions	0.484	0.520	0.484	0.496
Main Coder	0.616	0.652	0.652	0.64

Notes. Rows capture whether we are comparing human coding to the machine predictions or the main coder. Columns refer to the three coders that we hired to check the coder reliability. We report the top-1 accuracy for the 44- and 8-topic specification.

B.10 Classification of U.S. Congress Speeches

In this section, we provide further details on the U.S. speech data, which we use to assess the validity of cross-domain supervised learning in a different context. The speech data comes from Jensen et al. (2012). We focus on speeches held in the House of Representatives in the period from August 15, 1987 until July 27, 2002. In this period, the data includes 450,393 parliamentary speeches.

We started drawing a random sample of 250 speeches of speeches focusing on speeches that include more than 40 characters and hired the manifesto coder in charge of coding U.S. party manifestos to code these speeches to manifesto categories. The coder received training from the manifesto project and has significant experience in coding text documents. We did not give him any detailed instructions for the coding, but asked him to code in line with the manifesto project guidelines. The coding took in total 26 hours.

In contrast to New Zealand, most short speeches in the U.S. House of Representatives are procedural. These speeches are used, for example, to yield speaking time to parliamentarians or report voting results. Similar to other parliamentary democracies

such as United Kingdom, many short speeches in New Zealand take place in question times, which do not exist in the United States. We follow Dietrich et al. (2019) and limit our analysis of the United States to all speeches that have at least 50 words. Furthermore, we removed 12 speeches due to segmenting errors.

We used our trained classifier to predict the topics of the 150 speeches. In case we focus on 44 topics, the top-1 accuracy is 0.440, the top-3 accuracy 0.633, and the top-5 accuracy 0.693. Limiting the analysis to 8 topics leads to a top-1 accuracy of 0.520, a top-3 accuracy of 0.827 and a top-5 accuracy of 0.927. These results are similar to the results that we achieved using the New Zealand parliamentary speech data.

B.11 Bootstrapping Procedure

To assess robustness, we run the training procedure in multiple subsamples of the training set. Bootstrapped linear regressions often use hundreds or thousands of samples, but that is computationally infeasible for text-based machine learning. Antoniak and Mimno (2018) show that for language models, stability can be assessed using smaller samples, such as ten or twenty. Thus we run the classifier with 16 different re-sampled train-test splits.

We then compute the mean, median and standard deviation of the classification metrics. Table A9 illustrates the within-domain predictions focusing on the 44-topic specification. Table A10 summarizes the classification metrics based on the 8-topic specification. Tables A11 and A12 summarizes the recalls and precisions by topic of the 44-topic specification.

We do not observe notable differences in the mean/median recall or precision relative to the baseline reported values. The standard deviation in the metrics is sensitive to the frequency of the topics, however: infrequent topics have quite high standard

Table A9: Statistics on bootstrapped classifier performance in test set (44 topics)

	Mean	Median	Std Dev
Top-1 accuracy / F1 micro	0.5386	0.5379	0.0029
Balanced accuracy	0.3847	0.3849	0.0030
F1 macro	0.4135	0.4138	0.0034
F1 weighted	0.5230	0.5220	0.0030

Table A10: Statistics on bootstrapped classifier performance in test set (8 topics)

	Mean	Median	Std Dev
Top-1 accuracy / F1 micro	0.6451	0.6451	0.0028
Balanced accuracy	0.5077	0.5078	0.0030
F1 macro	0.5264	0.5274	0.0029
F1 weighted	0.6356	0.6357	0.0029

deviations, especially in topic precision. Again, this finding cautions against using infrequent source-corpus topics for empirical analysis using cross-domain learning.

Table A11: Statistics on bootstrapped class-level recall and precision (44 topics)

	Recall			Precision		
	Mean	Median	Std Dev	Mean	Median	Std Dev
Agriculture and farmers	0.5832	0.5778	0.0127	0.6314	0.6344	0.0224
Anti-growth economy	0.2522	0.2519	0.0164	0.4427	0.4511	0.0261
Anti-imperialism	0.0279	0.0000	0.0345	0.4375	0.0000	0.4961
Centralization	0.4395	0.4387	0.0185	0.4655	0.4635	0.0171
Civic mindedness	0.2592	0.2613	0.0163	0.4806	0.4708	0.0341
Constitutionalism	0.2665	0.2646	0.0212	0.5264	0.5303	0.0363
Controlled economy	0.2618	0.2599	0.0281	0.6116	0.6019	0.0821
Corporatism	0.0406	0.0430	0.0162	0.6083	0.5833	0.2655
Culture	0.5888	0.5904	0.0170	0.6900	0.6908	0.0245
Democracy	0.4864	0.4874	0.0161	0.4672	0.4647	0.0150
Economic goals	0.0368	0.0388	0.0109	0.3162	0.2990	0.0906
Economic growth	0.4759	0.4777	0.0162	0.4641	0.4606	0.0164
Economic orthodoxy	0.4954	0.4938	0.0178	0.5665	0.5694	0.0214
Economic planning	0.1111	0.1114	0.0226	0.4453	0.4580	0.0656
Education	0.7646	0.7650	0.0094	0.6870	0.6877	0.0107
Environmental protection	0.6897	0.6874	0.0113	0.5924	0.5921	0.0088
Equality	0.4599	0.4624	0.0087	0.4496	0.4509	0.0116
European union	0.5560	0.5516	0.0189	0.6045	0.6093	0.0264
Foreign special relationships	0.2438	0.2417	0.0273	0.4843	0.4694	0.0486
Free market economy	0.2300	0.2303	0.0179	0.4288	0.4298	0.0290
Freedom and human rights	0.4073	0.4076	0.0204	0.4928	0.4907	0.0251
Governmental and administrative efficiency	0.4456	0.4445	0.0140	0.4058	0.4093	0.0150
Incentives	0.4960	0.4947	0.0114	0.4982	0.4972	0.0179
Internationalism	0.5241	0.5244	0.0251	0.5248	0.5273	0.0194
Keynesian demand management	0.0464	0.0513	0.0242	0.7812	1.0000	0.3158
Labour groups	0.5849	0.5902	0.0173	0.5651	0.5641	0.0159
Law and order	0.6974	0.6974	0.0109	0.6187	0.6171	0.0103
Market regulation	0.4226	0.4238	0.0108	0.4623	0.4592	0.0145
Marxist analysis	0.0706	0.0682	0.0313	0.5962	0.6333	0.2406
Middle class and professional groups	0.2347	0.2288	0.0350	0.7001	0.6970	0.0758
Military	0.6011	0.5970	0.0153	0.6456	0.6398	0.0221
Multiculturalism	0.4576	0.4602	0.0192	0.5561	0.5577	0.0242
National way of life	0.3503	0.3516	0.0129	0.3553	0.3567	0.0113
Nationalisation	0.4139	0.4090	0.0181	0.6295	0.6249	0.0483
No topic	0.0389	0.0398	0.0082	0.2985	0.2806	0.0771
Non-economic demographic groups	0.2572	0.2590	0.0150	0.4383	0.4411	0.0226
Peace	0.3860	0.3835	0.0439	0.5788	0.5615	0.0630
Political authority	0.4521	0.4512	0.0164	0.3888	0.3874	0.0099
Political corruption	0.3043	0.3068	0.0302	0.6612	0.6710	0.0333
Protectionism	0.3776	0.3830	0.0283	0.5282	0.5222	0.0337
Technology and infrastructure	0.7023	0.7044	0.0113	0.5783	0.5783	0.0109
Traditional morality	0.3933	0.3927	0.0122	0.5827	0.5855	0.0215
Underprivileged minority groups	0.2070	0.2136	0.0181	0.4582	0.4567	0.0361
Welfare state expansion	0.7854	0.7853	0.0050	0.5507	0.5501	0.0068

Table A12: Statistics on bootstrapped class-level recall and precision (8 topics)

	Recall			Precision		
	Mean	Median	Std Dev	Mean	Median	Std Dev
Economy	0.7376	0.7370	0.0050	0.6662	0.6663	0.0044
External relations	0.6405	0.6398	0.0097	0.7295	0.7304	0.0069
Fabric of society	0.5684	0.5702	0.0064	0.6217	0.6212	0.0070
Freedom and democracy	0.4607	0.4593	0.0098	0.5974	0.5946	0.0097
No topic	0.0164	0.0152	0.0055	0.4199	0.3875	0.1253
Political system	0.4408	0.4415	0.0064	0.5216	0.5241	0.0090
Social groups	0.4164	0.4171	0.0079	0.5870	0.5874	0.0089
Welfare and quality of life	0.7809	0.7798	0.0047	0.6667	0.6653	0.0054

C Additional Material on Case Studies

C.1 Details on the Electoral Reform

The reform process began in 1986 when the Labour government appointed a Royal Commission to study a potential electoral reform in New Zealand. The Commission examined multiple electoral systems on the basis of ten criteria and finally recommended a reform toward a mixed-member electoral system. In 1993, 54 percent of the population decided in favor of the reform in a referendum and in 1996 the new electoral rules were applied for the first time. Several factors are important to understand the introduction of the reform. First, the population increasingly perceived election results and the resulting democratic representation as unfair. In 1978 and 1981 the Labour party received a majority of votes, but a minority of seats. Second, small parties and ethnic minorities were underrepresented in the parliament. Third, citizens were unsatisfied with the economic situation and the adversarial style of majoritarian systems. Lastly, citizens wanted to reduce 'electoral dictatorship' of majority parties (Lamare and Vowles, 1996; Vowles, 1995).

The reform changed the electoral system from a first-past-the-post to a mixed-member proportional electoral system. Table A13 summarizes the main differences between the electoral systems before and after the reform. In the first-past-the-post system, voters have one vote to select a candidate in their electorate. In the mixed-member proportional system, voters have two votes. Citizens can use one vote to select a candidate in the electorate and one vote to select a party at the national level. The reform increased the size of the parliament from 99 to 120 members.[7] The number of parliamentarians elected via district decreased from 95 to 60. The number of Maori districts increased in from four to five with the option of further increases. Under the

[7]Notice that the size of the New Zealand varied in the period before the reform.

mixed-electoral system, the remaining (list) seats are allocated using the Saint Laguë formula. Similar to the electoral system in Germany, parties have to achieve at least five percent of the party votes or win an electorate (Barker et al., 2003; Vowles et al., 2002; McGee, 2005).[8]

Table A13: Comparison of first-past-the-post and mixed-member electoral system.

	first-past-the-post	mixed-electoral system
number of votes	1	2
number of MPs	99	120 (+overhang)
number of general districts	95	60
number of list MPs	-	55
number of Maori district	4	5
electoral rule (districts)	relative majority	relative majority
minimum entry criteria	win an electorate	5% party votes or win an electorate
formula for list seats	-	Saint Laguë formula

Notes. The rows refer to features and the columns to the electoral systems. The table refers of the electoral systems in the periods 1993-1996 and 1996-1999.

For related work on the consequences of electoral reforms see Catalinac (2018) and Høyland and Søyland (2019).

C.2 Regression Models and Robustness

We assess the statistical significance of the graphical evidence using a fixed effects linear regression model. The dependent variable is the log probability that speech i focuses on political authority. The main explanatory variable is an indicator variable equaling one for speeches held after the electoral reform.

We run five OLS regression models with different specifications. We include in all models quadratic time trends. Standard errors are clustered by speaker (although statistical tests are robust to two-way clustering by speaker and year). Model 1 includes the baseline model. Model 2 adds speaker fixed effects, and Model 3 includes

[8]See http://www.elections.org.nz/ (accessed on July 30, 2017).

speaker-time trends. To provide some robustness, Model 4 includes in the regression the speech-level and speaker-level covariates. The covariates include indicator variables for gender, list MP, prime minister position and committee chair. They also include speech-level indicator variables for questions (e.g., questions to ministers, questions to private members), general debate, administrative speech, and committee of the whole House stage, and a party-level indicator variable for opposition party. Finally, Model 5 is the same as Model 2 (speaker fixed effects), but the observations are weighted by length of the speech (in number of words).

The full regression results are reported in Table A14. We see a significant positive effect across specifications, consistent with the graphical results. According to the preferred specification, Model 3, the probability that a specific speech corresponds to the topic political authority increases by about 30 percent after the reform. In Model 5, the effect of the reform is estimated to be smaller (17 percent) but still significant at the 0.05 level. Further, these results are robust to interacting these controls with the reform dummy.

In Table A15 we re-run models of Table A14, but include an indicator variable that is equal to 1 if parliamentarians were elected for the first time after the electoral reform was implemented, and zero otherwise. The results suggest that the increasing importance of the political authority topic is related to the behavior of parliamentarians that were elected before the reform was implemented.

Furthermore, we use as a robustness test alternative dependent variables and statistical models. Table A16 replicates the main results of our paper using a binary dependent variable and a logistic regression model. As the table shows, our results are robust. We also used multilevel regression models, which confirm the robustness of our findings.

Table A14: Regression analysis of political authority (main results)

	(1)	(2)	(3)	(4)	(5)
Post electoral reform	0.268*	0.243*	0.297*	0.247*	0.174*
	(0.109)	(0.098)	(0.117)	(0.117)	(0.074)
List MP					-0.076
					(0.094)
Woman					-0.397**
					(0.094)
Ethnic minority					0.273
					(0.223)
Prime Minister					0.790**
					(0.124)
Committee chair					-0.016
					(0.070)
Question					-0.560**
					(0.034)
General debate					0.839**
					(0.050)
Administrative speech					0.072
					(0.051)
Committee stage					0.013
					(0.030)
Government party					-0.540**
					(0.044)
Constant	-3.119**	-3.186**	-3.448**	-3.328**	-2.556**
	(0.086)	(0.096)	(0.180)	(0.133)	(0.076)
Quadratic Trend	X	X	X	X	X
Speaker Fixed Effects		X	X	X	
Speaker Trends			X		
Weighting by Speech Length				X	
Controls					X
N	290456	290456	290456	290456	290456
R^2	0.008	0.092	0.110	0.138	0.098
ll	-5.659e+05	-5.530e+05	-5.502e+05	-6.246e+05	-5.520e+05

Standard errors, clustered by speaker, in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A15: Regression analysis of political authority (new versus old MPs)

	(1)	(2)
Post electoral reform	0.320**	0.206**
	(0.107)	(0.070)
Entered parliament after reform	-0.326**	-0.399**
	(0.119)	(0.098)
List MP		0.050
		(0.082)
Woman		-0.362**
		(0.089)
Ethnic minority		0.323
		(0.203)
Prime Minister		0.766**
		(0.120)
Committee chair		-0.016
		(0.069)
Question		-0.566**
		(0.034)
General debate		0.844**
		(0.049)
Administrative speech		0.043
		(0.054)
Committee stage		0.019
		(0.028)
Government party		-0.537**
		(0.044)
Constant	-3.109**	-2.548**
	(0.086)	(0.074)
Quadratic Trend	X	X
Controls		X
N	290456	290456
R^2	0.011	0.097
ll	-5.654e+05	-5.514e+05

Standard errors, clustered by speaker, in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

Table A16: Robustness: Logistic regression analysis of political authority (binary dependent variable)

	(1)	(2)	(3)	(4)
Post electoral reform	0.286**	0.241**	0.256**	0.219**
	(0.095)	(0.084)	(0.094)	(0.067)
List MP				-0.085
				(0.106)
Woman				-0.468**
				(0.101)
Ethnic minority				0.366+
				(0.216)
Prime Minister				0.895**
				(0.060)
Committee chair				-0.181**
				(0.069)
Question				-0.699**
				(0.039)
General debate				0.973**
				(0.056)
Administrative speech				-0.177**
				(0.067)
Committee stage				-0.099**
				(0.031)
Government party				-0.433**
				(0.045)
Constant	-1.356**	-1.600**	-3.357**	-0.858**
	(0.083)	(0.098)	(0.411)	(0.084)
Quadratic Trend	X	X	X	X
Speaker Fixed Effects		X	X	
Speaker Trends			X	
Controls				X
N	290456	290456	290456	290456
Pseudo R^2	0.005	0.046	0.052	0.052
ll	-1.636e+05	-1.570e+05	-1.559e+05	-1.554e+05

Standard errors, clustered by speaker, in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

C.3 Debate Participation

Table A17 summarizes the average topic probabilities by gender focusing on the legislative period 1996-1999. Table A18 illustrate the average topic probabilities by gender focusing on all legislative periods.

Table A17: Average Probability by Topic (Period 1996-1999)

	Economy	External relations	Fabric of society	Freedom and democracy	No topic	Political system	Social groups	Welfare
Men	0.147	0.042	0.156	0.126	0.020	0.260	0.095	0.154
	(0.008)	(0.002)	(0.009)	(0.005)	(0.001)	(0.007)	(0.006)	(0.008)
Women	0.123	0.028	0.156	0.122	0.019	0.251	0.097	0.179
	(0.012)	(0.003)	(0.008)	(0.004)	(0.001)	(0.006)	(0.004)	(0.008)

Notes. The table illustrates the average of the probability that a speech focuses on a topic by gender. The table focuses on the legislative period 1996-1999 to keep the composition of the parliament constant. Standard errors are in parentheses.

Table A18: Average Probability by Gender

	Economy	External relations	Fabric of society	Freedom and democracy	No topic	Political system	Social groups	Welfare
Men	0.169	0.042	0.139	0.117	0.019	0.244	0.099	0.171
	(0.005)	(0.002)	(0.005)	(0.003)	0.001	(0.004)	(0.004)	(0.005)
Women	0.142	0.029	0.131	0.095	0.015	0.219	0.115	0.252
	(0.009)	(0.002)	(0.008)	(0.004)	(0.001)	(0.007)	(0.006)	(0.011)

References

Antoniak, M. and D. Mimno (2018). Evaluating the Stability of Embedding-based Word Similarities. *Transactions of the Association for Computational Liguistics 6*, 107–119.

Barker, F., J. Boston, S. Levine, E. McLeah, and N. S. Roberts (2003). *An Initial Assessment of the Consequenes of MMP in New Zealand*, Chapter 14, pp. 297–322. Oxford: Oxford University Press.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, and E. Tanenbaum (2001). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1998*. Oxford: Oxford University Press.

Catalinac, A. (2018). Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos. *American Political Science Review 112*(1), 31–48.

Chen, T. and C. Guestrin (2016). XGBoost: A Scalable Tree Bosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, San Francisco, California, USA, pp. 785–794.

Curran, B., K. Higham, E. Ortiz, and D. Vasques Filho (2018). Look who's Talking: Two-Mode Network as Representation of a Topic Model of New Zealand Parliamentary Speeches. *PLOS ONE 13*(6), 1–16.

Denny, M. J. and A. Spirling (2018). Text Preprocessing for Unsupervised Learning: Why it Matters, when it Misleads, and what to Do about it. *Political Analysis 26*(2), 168–189.

Dietrich, B. J., M. Hayes, and D. Z. O'Brien (2019). Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech. *American Political Science Review 113*(4), 941–962.

Edwards, C. (2015). Hansard - the True Mirror of Parliament? Key Principles in its Editorial Development.

Handler, A., M. J. Denny, H. Wallach, and B. O'Connor (2016). Bag of What? Simple Noun Phrase Extraction for Text Analysis. Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing.

Hayward, J. (Ed.) (2015). *New Zealand Government and Politics*. South Melbourne: Oxford University Press.

Høyland, B. and M. G. Søyland (2019). Electoral Reform and Parliamentary Debates. *Legislative Studies Quarterly 44*(4), 593–615.

Jensen, J., E. Kaplan, S. Naidu, and L. Wilse-Samson (2012). Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. *Brookings Papers on Economic Activity* (Fall 2012), 1–60.

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge, and M. D. McDonald (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD 1990-2003*. Oxford: Oxford University Press.

Krause, W., P. Lehmann, J. Lewandowski, T. Matthieß, N. Merz, and S. Regel (2018). Manifesto Corpus. Version: 2018-2. Berlin: WZB Berlin Social Science Center.

Lamare, J. W. and J. Vowles (1996). Party Interests, Public Opinion and Institutional Preferences: Electoral System Change in New Zealand. *Australian Journal of Political Science 31*(3), 321–346.

McGee, D. G. (2005). *Parliamentary Practice in New Zealand*. Wellington: Dunmore.

Mikhaylov, S., M. Laver, and K. Benoit (2012). Coder Reliability and Misclassification in the Human Coding of Party Manifestos. *Political Analysis 20*(1), 78–91.

Miller, R. (2005). *Party Politics in New Zealand*. Oxford: Oxford University Press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Peterson, A. and A. Spirling (2018). Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems. *Political Analysis 26*(1), 120–128.

Ralphs, K. (2009). Recording Parliamentary Debates: A Brief History with References to England and New Zealand. *Australasian Parliamentary Review 24*(2), 151–163.

Spirling, A. (2016). Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832-1915. *The Journal of Politics 78*(1), 120–136.

Vowles, J. (1995). The Politics of Electoral Reform in New Zealand. *International Political Science Review 16*(1), 95–115.

Vowles, J., P. Aimer, J. Karp, S. Banducci, R. Miller, and A. Sullivan (2002). *Proportional Representation on Trial*. Auckland: Auckland University Press.