An Unsupervised Text-As-Data Approach to Detecting the Effects of Electoral Constraints on Online Communication Strategies

Musashi Harukawa, University of Oxford

Abstract

A wealth of political science literature notes the connection between the constraints created by electoral systems, the incentive for re-election, and legislator priorities. This study measures the effect of these constraints on legislators' online communication strategies. Using the MMP representational electoral system in New Zealand, I compare a variety of text-as-data models to find systematic differences between List and SMD MPs in a dataset of 15,381 tweets by New Zealand legislators. A number of patterns emerge: List MPs are more likely to engage in rhetoric critical of spending, whereas SMD MPs are more likely to advocate reform. List MPs are also more likely to use Twitter to celebrate or congratulate, whereas SMD MPs are more likely to respond directly to other users.

Introduction

In this paper, I present a variety of methods showing how text-as-data approaches and Twitter data can be used to analyse the effect of constraints created by electoral context on legislator role prioritisation. Politicians' actions, while driven by multiple competing motivations (Strøm 1997), are normally constrained by the desire to be re-elected and continue exerting political power. The nature of this constraint is in turn affected by the electoral system they operate within; Carey and Shugart (1995) provide a framework in which they rank electoral systems by their propensity to produce personal vote cultivation, and thus intraparty competition, or party unity.

Earlier empirical works in this area (Heitshusen, Young, and Wood 2005; Martin 2011) find mixed results for the strength of the constraining effect of electoral systems. I aim to identify areas in which this constraint has a substantial effect on legislator behaviour. My substantive contribution examines whether this constraint has an effect on online communication strategy in particular, and especially on the social media platform Twitter.

The role of Twitter in modern politics is hard to ignore; the majority of elected officials in advanced English-speaking democracies have an account which sees regular use[1], and one estimate[2] counts more than 5 million tweets written by world leaders alone as of May 2017. It has become an important channel of communication between elected officials and the populace, whether by direct usage on the Twitter platform, or via an intermediary such as a newspaper reporting on and providing analysis of politicians' tweets (such as Donald Trump). Being already digitised, it also offers advantages as a data source whose collection is highly scalable.

Developments in computational linguistics/natural language processing (hereafter NLP) provide researchers for extracting useful insights from these large quantities of data. I propose to use NLP methods measure the effect of these constraints on the online communication strategies via Twitter. In particular, I look at the case of New Zealand, where the mixed-member proportional (MMP) representation electoral system provides for the comparison of legislators from the same party seeking re-election via different electoral systems. Previous literature (Carey and Shugart 1995) theorises that the incentive structures under single-member district plurality and closed list proportional representation should lead to high levels of personal vote cultivation in the former, and high levels of party loyalty in the latter. Given that New Zealand's electoral system contains legislators within both systems, I expect this constraint to have measurable effects on communication strategy.

[1] Based on my own data collection and analysis of Twitter usage by politicians in Australia, Canada, Ireland, New Zealand, the United Kingdom and the United States. This data is included in the replication materials.

[2] https://twiplomacy.com/blog/twiplomacy-study-2017/

I also take the opportunity to compare and discuss four different approaches to addressing the problem of extracting useful information from large text datasets. Linguistic data is inherently high-dimensional, leading to a dimensionality reduction/information retrieval problem. The four models I use to address this problem include: Wordfish, Principal Component Analysis, Topic Modelling, and Document Embedding. I aim to clarify the strengths and pitfalls of each approach for the task at hand, and provide guidance for future political science researchers interested in using these methods in their own research.

Theory and Literature

As noted by Searing (1994), legislators occupy multiple roles, from ministerial positions, to drafting legislation, to exercising oversight, to pushing their constituents' interest at a national level. Strøm defines *legislative roles* "as behavioural strategies conditioned by the institutional framework in which parliamentarians operate… *[p]arliamentary roles are routines, driven by reasons (preferences) and constrained by rules*" (Strøm 1997, pgs. 157-158, italics original). Variation in behavioural strategies by parliamentarians therefore has individual and institutional components. Preferences for particular activities or objectives, such as office-seeking, policy-seeking and/or executive oversight can be understood to be exogenous and given, whereas "rules," or institutional constraints are dependent on the political system.

Strøm (1997) notes that limited resources of time and effort create a "hierarchy" for these objectives, such that legislators must prioritise certain goals over others. While their innate preferences are not measurable, the constraints generated by the institutional configuration the legislator operates within are observable. In particular, given that achieving any of the aforementioned goals depends on (re-) election, legislators must account for the ways in which their actions affects their prospects for retaining their seats.

For instance, in the United Kingdom and Canada (among others), the re-election context of parliamentarians is determined by single-member district plurality (SMD) electoral rules. This means that parliamentarians need to gain a plurality of votes in the district that they represent, and that votes cast for that parliamentarian have no effect on the election prospect of those outside of the district. Within this context, legislators wishing to be re-elected want to maintain their popularity amongst those who keep them in power, meaning they must be seen to be acting in a way that satisfies the interests of their constituents. In contrast, in Israel, legislators are voted in with a closed-list proportional representation (PR) system, where voters cast a single vote for a party within a single nation-wide district, then parties allocate seats to their legislators equal to the share of the total vote they received. In this context, re-election depends on creating a popular image for their party, and ensuring that the party elite favours placing them high on the list to increase the likelihood that they will have a seat.

Carey and Shugart (1995) provide a framework for ranking electoral systems by their propensity to produce incentives for personal vote cultivation and therefore intraparty competition, or party unity. According to the authors, the above example of closed-list PR is conducive to producing the highest levels of party unity and the lowest levels of personal vote cultivation because re-election for a legislator depends on the party elite. In this context, efforts to cultivate personal vote is not rewarded, and may even be punished (Carey and Shugart 1995). On the other end of the spectrum, re-election in SMD is determined solely (in the sense that there is no pooling of votes) by an MP's ability to gather votes in their own constituency, and therefore rewards a constituency focus by the legislator.

A wide variety of methods and data have been employed in empirical works studying the effect of electoral constraints on strategic behaviour by legislators, showing that minimally, we cannot reject Carey and Shugart (1995)'s framework. The first study I examine by Heitshusen, Young, and Wood (2005) surveys 254 parliamentarians, asking them to rank their legislative priorities, and finds that re-election context has the strongest effect on prioritising constituency focus. The strength of this approach is that it is a direct measure of the variable of interest: role prioritisation by legislators. Although Heitshusen, Young, and Wood (2005) gather an impressively large cross-national dataset, I argue that it wrongly places an emphasis on beliefs over strategies. Asking legislators what they believe they prioritise can be distinct what they actually do. Moreover, legislators electorally beholden to their own constituents are more likely to answer that they prioritise constituents, regardless of their actual priorities, in order to improve their personal reputation among those who determine that they continue exercise political power. On the other hand, given that I am interested in political outcomes„ I argue that measurements of legislator prioritisation should be based on realised activities because the strategies of politicians and the actions that these entail are more important than what they “actually” believe. Their beliefs are important only insofar as they affect their strategies and actions.

Martin (2011)‘s approach is better in this regard; he compares parliamentary questions (PQs) brought forward by Dáil Deputies with varying district magnitudes. 123,762 PQs brought forward between 1997 and 2002 are manually coded to measure personal vote-earning strategies. In contrast to Heitshusen, Young, and Wood (2005), this study finds that the effect of district magnitude is negligble, but that distance from Dublin to their constituency has a significant effect. Given that this behaviour is an integral part of a legislators’ role, asking constituent-focused PQs both signals a focus on constituency service and simultaneously performs it. My main criticism of this approach is the reliability and affordability of having a panel of experts to manually code 123,762 questions. This is unavailable to many researchers, and as I will argue, unnecessary. I also note that the null result on the effect of the electoral system in his study is reconcilable with Carey and Shugart’s framework, as legislators within the same electoral system but different district magnitudes are argued to only have a minimal difference in incentive and behaviour. Therefore a better test of the framework would compare systems at opposite ends of the ranking, which is what I do.

I argue that politicians’ tweets indicate to us whether a politician wishes to be viewed as holding a particular position on a given issue. In other words, I believe that tweets are deliberate and strategic attempts to cultivate particular types of reputation, with the target audience being minimally the politically literate (e.g. Crick and Porter 1978). In the sense that politicians have incentives to be disingenous in their online communications in order to portray a particular image of themselves, my approach suffers from similar issues to Heitshusen, Young, and Wood (2005). However, tweets can instead inform us of the strategies of politicians, and therefore allow us to infer what aspects of their role they believe to be important.[3]

Jackson and Lilleker (2011) present a typology for the strategic functions of tweets. A given tweet may fall into one or more of the following categories: *ingratiation* (making oneself likeable), *self-promotion* (making oneself appear competent), *supplication* (making oneself appear in need of assistance), *exemplification* (making oneself appear exemplary and worthy), or *intimidation* (making oneself appear dangerous or threatening). I expect that SMD MPs will engage in a higher degree of ingratiation, self-promotion and supplication, whereas List MPs are more likely to exemplify their own party. Intimidation is likely idiosyncratic, but to the extent that it can successfully build a personal reputation as a maverick, I believe that SMD MPs will engage in this kind of behaviour more.

The extent to which Twitter matters for political issues has been debated (Wilson and Dunn 2011; Morozov 2011; Lynch 2011), and Tucker et al. (2016) provide a persuasive account of why it is premature to entirely dismiss it. One aim of this research is to assess the utility of Twitter data by analysing the information that can be extracted from legislators’ discourse, considering how it can help us understand intraparty dynamics, and measuring how it systematically differs from other forms of communication by the same actors.

A number of recent studies of political communication introduce applications for text-as-data approaches. Broadly, text-as-data approaches are methods for textual analysis that combine machine-learning algorithms, statistical models and the numerical representation of text data. For instance, Grimmer (2013; 2016) provides multiple notable studies of legislator communication strategies. In the latter paper, Grimmer (2016) uses a hierarchical topic model to extract credit-claiming behaviour from the press releases of US legislators.

Another important example is Peterson and Spirling (2018), who use the accuracy of classification algorithms on parliamentary debate transcripts to measure polarisation. Using 78 years of transcripts from Westminster and comparing four different classifiers, they find that periods where the classifier is better at correctly labelling a given speech as Labour or Conservative correlate with historic periods associated with a higher degree of polarisation within parliament. I extend this approach in my own work by combining it with document embeddings.

The strength of these approaches is that they allow researchers to operationalise large bodies of text and

[3] A related question one might ask is how a given tweet affects voters’ perceptions of the legislator who authored it. While interesting, this question is beyond the scope of this study.

recover quantities of interest (also known as Information Retrieval). The benefit of these methods is not limited to their scalability and speed either; text-as-data methods are capable of detecting overall patterns within large texts that humans are unlikely to notice, such as identifying 1000+ line meters in poetry (Agirrezabal, Alegria, and Hulden 2016).

Text-as-data methods are not without their pitfalls, either. Standard pre-processing procedures for converting text to numeric data often lose a great deal of the syntactic and semantic information, and tend to struggle with polysemy (where individual words have multiple meanings) (Grimmer and Stewart 2013). Interpreting the patterns detected by unsupervised approaches can be a subjective process (Egami et al. 2018) and do not necessarily have a relevant substantive meaning. Validation is therefore extremely important (Peterson and Spirling 2018), and while no "gold standard" of validation exists (although Rodman (2019) offers some candidates), being clear about when and where researchers have made these subjective interpretations improves visibility and reproducibility. For each of the models that I present, I will discuss the validity of the results.

Model and Hypotheses

I present a model of the effect of the electoral constraints on legislators' tweets using a framework and jargon from formal laguage theory (see e.g. Wintner 2010). This model assumes that tweets (documents, d), are probabilistically generated by a function, called the document-generating process (DGP, M), which chooses words (*tokens*[4], w_n) from a finite set of possible tokens (the vocabulary, V). For each draw from the vocabulary by the DGP, there is a non-zero probability of selecting the special terminating string, w_ω, which denotes the end of a document.

Therefore, document i by legislator j at time t[5], d_{ijt}, can be defined as an ordered set of tokens $\langle w_1, w_2, ..., w_\omega \rangle$:

$$d_{ijt} := \langle w_1, w_2, ..., w_n, w_\omega \rangle; \; \forall n : w_n \in V$$

The DGP, $M(\cdot)$, is a probability distribution over n draws from V parameterised by, in addition to the syntactic rules defining the language $\mathcal{L}$, the preferences and constraints of a legislator. It returns an ordered set of n tokens, i.e. the document as defined above. I propose the following document-generating process for legislators' tweets:

$$d_{ijt} \leftarrow M(\lambda(e_j, m_j, u_j, g_p, c_j), \mathcal{L})$$

where:

- d_{ijt}: document i for legislator j at time t, formed of a sequence of tokens $\langle w_n, ..., w_\omega \rangle$
- $\lambda(u_j, e_j, g_p, c_j, m_j)$ the re-election constraint function, which is a linear combination of g_p, c and m.
- e_j: the electoral context of legislator j.
- m_j: the competitiveness of the re-election bid for legislator j. For example, this could be measured by the margin by which legislator j was elected in the previous election.
- u_j: the policy preference (utility) function for legislator j.
- g_p: the policy preference (utility) function for party p and legislator j.
- c_j: the policy preference (utility) function for the constituency of legislator j.
- $\mathcal{L}$ is the language of document d, and therefore a syntactic/semantic constraint on the probability of tokens.

The re-election constraint function $\lambda(\cdot)$ represents the relative weights a legislator gives to the preferences of various groups when producing statements. It assigns weights $\sum\{\lambda_u \lambda_g, \lambda_c\} = 1$ to its arguments u_j, g_p and c depending on the electoral context of legislator j, e_j, and the expected competitiveness of the re-election bid.

[4] To be more consistent with the conventions of formal language theory, *strings* would be the more appropriate terminology in this case. I use the term *tokens* for consistency with the natural language processing methods, and to reduce the amount of jargon used in this paper.

[5] The subscript i only applies for cases where time t is measured in discrete periods, as otherwise the maximum number of documents authored by legislator j at time t would be 1.

λ_u is decreasing in m_j, λ_g is larger when the re-election is more highly dependent on the party elite, and λ_c is larger when the re-election is dependent on constituents.

The core of my argument is that controlling for competitiveness, party and personal preference, the re-election constraint has a sufficient effect on the document-generating process such that the tweets of New Zealander legislators in SMD seats should systematically differ from the tweets of legislators in Closed List PR (hereafter List) seats. In other words, *I expect List MPs place a greater weight on party preference relative to constituency preference, when compared to SMD MPs.* Because I assume that party preferences are not always identical to constituent preferences ($\exists j : g_p \neq c_j$), there will be some SMD MPs whose DGP will place higher likelihood on producing documents not identical to the ones produced by List MPs of the same party.

The systematic difference between SMD and List MPs communication strategies may appear in a number of ways. Firstly, and somewhat obviously, SMD MPs will display greater *constituent focus* in their tweets. This will take the form of tweets endorsing events in their own constituency, announcing policies or measures that will benefit their own constituents, or directly engaging with their constituents online. This hypothesis is somewhat trivial, as List MPs do not have a constituency to focus on, but experience has shown that List MPs may make locally-focused tweets for the sake of their party, or if they want to compete for said constituency in a future election.

- `Hypothesis 1` (*Constituency Focus*):
  SMD MPs are more likely to write tweets focused on particular places, especially their own constituency, than List MPs.

Another systematic difference I expect to find is a greater degree of ideological language in List MPs tweets, conditional on the extent to which their party's brand is dependent on ideological and not programmatic linkages (Kitschelt 2000; Lupu 2013). Where the party's brand is ideological, List MPs likelihood of re-selection onto a high position on the party list is dependent on their success in "marketing" this party brand well. In Jackson and Lilleker (2011)'s framework, I would expect a higher frequency of "exemplification" tweets.

- `Hypothesis 2` (*Ideological Language*):
  Where a party's brand is based on ideological linkages, List MPs are more likely to write ideologically-focused tweets than SMD MPs from the same party.

Note that it is difficult to provide hypotheses that I *a priori* know that I am able to test with the output of an unsupervised model. Assessing either of these hypotheses depends on how successful I am in detecting these particular patterns of constituent-focused or ideologically-focused tweets, but there is no guarantee that any of the models that I use will be able to find these particular patterns. Where I am able to operationalise these patterns, I can use statistical hypothesis testing to determine whether these hypotheses are compatible with my data.

Methods

The data that I have for testing the above hypotheses is inherently high-dimensional, meaning that I have an information retrieval (IR) and dimensionality reduction (DR) problem. I therefore propose to use four different unsupervised models to detect patterns in the data attributable to the constraint function λ.

It is not always obvious that a given IR/DR method has captured a dimension of interest, and identifying this dimension of interest depends on the method. I offer ways to interpret the output of the model on a per-model basis below. I note however that labelling the latent dimensions identified by the model should not be done on the basis of how well these latent dimensions predict some outcome. In other words, I should not identify the dimension that best sorts between List and SMD candidates as the "constituency focus" dimension _just because it provides the best accuracy for classifying List and SMD candidates and I have a theoretical prior for believing that constituency focus is the dimension that separates List and SMD candidates. Rather, I should lable the dimension based on what tokens it places high weights on, and then subsequently interpret the effectiveness of this label in differentiating between legislators.

Special attention needs to be paid to the hierarchical sources of variance in text data. Lauderdale and Herzog (2016) identify these, "in roughly descending order[...] (1) language, (2) style, (3) topic and only then (4) position, preference or sentiment." The electoral constraint can affect several levels of this variance; it may affect language such as when a MP represents a Maori constituency; it may affect style; it certainly affects topic. However, in order to measure the cases where it affects position, preference or sentiment, which is a quantity of interest, I need to use models that can account for and measure the variance created by the higher-order sources of variance: language, style and topic.

Method 1: Wordfish

The first approach uses the scaling algorithm Wordfish (Slapin and Proksch 2008). This unsupervised approach, originally created to estimate ideological distances between party manifestos, estimates the relative positions θ[6] of parties in a latent space by assuming that word frequencies y_{ijt} are drawn from a Poisson distribution over λ_{ijt} (not the same λ as the constraint function in the DGP), where λ_{ijt} is estimated with the following fixed effects model:

$$y_{ijt} \sim Poisson(\lambda_{ijt}) \, : \, \lambda_{ijt} = e^{\alpha_{it} + \psi_j + \beta_j \theta_{it}}$$

where y_{ijt} is the word frequency of word j in legislator i's tweet at time t, α is a set of legislator-time fixed effects, ψ is a set of word fixed effects, β captures the word-specific importance for word j in discriminating between positions, and θ is the position of tweet i at time t.

This model explicitly describes the DGP; in this case it is a Poisson distribution parameterised by the legislator fixed effects, word fixed effects, and a unidimensional distribution of documents in a latent space. That Wordfish uses the Poisson distribution is based on an empirical reality–given a sufficiently large random text, word frequencies can be reasonably well-approximated by the Poisson distribution (Slapin and Proksch 2008). Unlike other similar distributions, the Poisson has a single parameter (λ) which is endogenous to the model. A major advantage of the Wordfish model is that it does not require any hyperparameter optimisation; the only parameters that need to be passed initially are two reference documents to determine "left" and "right" on the resulting dimension.

The major downside of this model is that the standard variant of Wordfish reduces variation in position within the corpus (collection of all documents, D) to a single dimension θ, and conflates all sources of non-ideological variance in word probabilities into a pair of legislator-specific and word fixed effects α and ψ, whereas our DGP entails hierarchical sources of variance, i.e. party effects, constituency effects, and electoral context effects. Reduction of position to a single dimension is problematic because movement along θ could represent a difference in ideology, audience focus, or topic. Therefore although Wordfish provides the means to operationalise discursive distance between politicians, it does not provide us with a clear meaning of what these distances mean.

One way to partially circumvent this issue would be to use the Wordfish model to conduct out-of-sample (OOS) prediction. This functionality does not exist in the `R` library `quanteda` (or any other implementation of the Wordfish model), but I outline how it could be implemented.[7] By estimating positions of documents in a prediction dataset with parameters fitted from a training dataset, it is possible to provide a more specific meaning to the scale provided by θ. For example, I fit a Wordfish model on the subset of documents belonging to SMD MPs, D_{SMD}, I get fitted values $\hat{\alpha}_{SMD}$, $\hat{\psi}_{SMD}$, $\hat{\beta}_{SMD}$ and $\hat{\theta}_{SMD}$. These fitted values could then be used to derive θ for documents in D_{List}. These positions $\hat{\theta}^{List}_{SMD}$ tell us how List politicians vary in the dimension that distinguishes SMD politicians. Given that we believe SMD politicians vary based on their parties and the preferences of their constituencies, we may find that $\hat{\theta}^{List}_{SMD}$ varies less than $\hat{\theta}^{SMD}_{SMD}$, indicating that List MPs demonstrate a higher degree of party unity than their SMD counterparts. This methos accounts for higher-order variation between datasets; given that language is constant between the two dimensions, and fixed effects could be included for individual legislators, the only conflated variation will be topic. The drawback of this approach is that the vocabulary V will have to be reduced to the intersection of the training

[6] Note: the original paper uses ω to denote the position of a document, whereas later implementations of the algorithm use θ. For consistency with the `R` implementation of Wordfish in `quanteda` (Benoit et al 2018), I refer to the position as θ.

[7] Although I have discussed this method with the maintainers of the `quanteda` library, implementing this code was beyond the scope of this thesis. Similarly, there are no multi-dimensional variants of the Wordfish model currently implemented.

and prediction corpora. The utility of this approach therefore depends on the similarity of the two corpora. The step of removing all terms that do not occur in both corpora may also bias the effect of any difference downwards by removing terms associated with one of the groups.

In absence of this OOS prediction alternative, I present several alternatives for determining systematic differences between List and SMD MPs. The first is to limit the corpora to the subset of parties which have both List and SMD MPs (Labour and National), and then conduct two-sample Kolmogorov-Smirnov tests[8] to compare the distribution of $\hat{\theta}$ for List and SMD candidates within each party. The next is to fit separate θs for List and SMD candidates and graphically compare the distributions per-party. The final is to fit a single Wordfish model to the full corpus, and compare the distribution of every party-electoral context subtype with Kolmogorov-Smirnov tests and graphs.

Method 2: PCA and Sparse PCA

Princpal Component Analysis (PCA) is a standard matrix decomposition technique that given a $D \times V$ matrix of values, finds $K : \; K \leq V$ components to produce $D \times K$ and $K \times V$ matrices of linearly uncorrelated vectors called principal components. The first principal component k_1 captures the highest degree of variance within the dataset, and the second captures the second highest, and so on, until the case where $K = V$ and the full variance of $D \times V$ is captured.

Sparse PCA (Zou, Hastie, and Tibshirani 2006) is a variant of this technique that relaxes the constaint on each principal component being a linear combination of all variables V, thus allowing the resulting components to be a product of a subset of the variables. This is similar to the LASSO/Elastic Net regression framework. I include this variant because the document-feature matrix is highly sparse (i.e. most elements of the matrix are zero), meaning I can expect components to reflect more commonly occurring terms versus less common ones.

By finding components $k' \in K$ that weight terms that I *a priori* believe should have discriminating potential between List and SMD documents (such as mentions of a legislators' own constituency, or localised funding projects), I test the null hypothesis that the distribution of List and SMD documents onto this component is identical using a two-sample Kolmogorov-Smirnov test. A low p-value will indicate that there is something substantively different in the way that these legislators tweet along this latent dimension.

The second approach uses the Sparse Principal Components Analysis algorithm (Zou, Hastie, and Tibshirani 2006) as implemented in the Python library `scikit-learn` (Pedregosa et al. 2011), a variant on the more common PCA technique that better handles the sparse nature of text data. The components calculated by PCA capture the maximum degree of variance in the data in a lower dimension. This algorithm provides the advantage of being more computationally efficient than topic models (below), and can be interpreted similarly.

I am not aware of any applications of PCA or Sparse PCA to text data in a political science context. I suspect that this is due to the fact that political scientists often deal with relatively small corpora, on the order of tens or hundreds of thousands of documents, whereas computational linguists train their models with corpora tens of millions of documents in length (Rodman 2019). Given that my dataset is relatively small (15,381 documents), I expect that the model will be especially sensitive to outliers, which will drive a large proportion of the variation in the dataset. Although the pre-processing steps will remove some of the outliers, lexical sets that occur in high frequency in a small number of documents will remain in the data, and PCA techniques are likely to pick up foremost on these sources of variation.

Method 3: Topic Models

Topic models are a class of matrix decomposition methods that have gained a great deal of traction as a dimensionality reduction and information retrieval technique for political text data (Egami et al. 2018). In addition to the study of credit claiming in press releases by US Congress members (Grimmer 2016), notable applications include the classification of 7920 Russian-language public statements by Russian military

[8] The Kolmogorov-Smirnov test gives the probability that a sample has been drawn from a given distribution. The two-sampled variant provides the probability that two samples have been drawn from the same underlying distribution, without specifying the shape of this distribution. I use this to test the null hypothesis that two samples (fitted values) were drawn from the same distribution, and present the resulting p-values.

elites and civilians on foreign policy as being either *Restrained*, *Activist* or *Neutral* (Stewart and Zhukov 2009), the classification of legislation as being about environment or other issues (congressionalbills.org), and operationalising xenophobic attitudes in open-ended survey responses (Margaret E. Roberts et al. 2014a).

At a high level, topic models function similarly to PCA, in that they provide a $D \times K$ (the topical prevalence matrix) and $K \times V$ (the topical content matrix) decomposition of a $D \times V$ matrix (the corpus) where $K \leq V$. However, topic models are designed to detect hidden semantic structures within the text, and the exact way they do so depends on the variant.

In this paper, I use the structural topic model (STM) introduced in Margaret E. Roberts et al. (2014b) and implemented in the `R` library `stm`. STM differs from other topic-modelling techniques such as Latent Dirichlet Allocation (LDA) in that allows for the inclusion of document-level and lexical covariates that effect either topical prevalence, topical content, neither or both (Lucas et al. 2015). These can include author, author's age or gender, date of publication, or specific terms within the vocabulary.

The result of STM, in contrast to PCA and Sparse PCA, is a set of topics K in which the highest-weighted terms in the topical content matrix often reflect substantive "topics" within the text. Although the process of labelling topics is subjective, by presenting the highest probability terms per-topic in a table, I allow for the reader to contest my interpretation of these labels.

Once I have the labels, I fit a multinomial logit model to predict topical prevalence matrix weightings as a function of document covariates including electoral context, party, date, and other controls. The difference in estimated topic weight for List and SMD MPs with all other controls held equal can be understood as a measure of the systematic difference between the two groups along the latent dimension identified by the topical content vector.

Method 4: Document Embeddings

Document embeddings are a variant of word embeddings, which are the vector representation of text in a latent space. This model, proposed in Mikolov et al. (2013), is an implementation of a linguistic theory called the distrbutional hypothesis (Firth 1957), which argues that "independent of any other context or even grammatical order, the systematic collection of word collocations can allow us to make semantic 'sense' out of words" (Rodman 2019, 5). The model uses a shallow neural network to predict the missing term in a moving window of word collocations within documents, and the resultant vector representation of individual words has many attractive features. One is that semantically similar terms are proximate in the latent space. Another is that semantic differences "obey" simple vector arithmetic. The example given in the `word2vec` paper is *vector("King") - vector("Man") + vector("Woman")* results in a vector that is closest to the vector representation of the word "Queen" (Mikolov et al. 2013, 2). This achieves a degree of semantic recognition that other models are unable to capture.[9]

Document embeddings are an extension of this model, introduced in Le and Mikolov (2014) and implemented in the Python library `gensim` (Řehůřek and Sojka 2010). The design is similar, except that in addition to a moving window of proximate terms, the entire document is included as a predictor of individual terms within the document, and the result is a vector representation of the entire document within a latent space.

This model is exceptionally good at detecting sentiment and semantic relationships between documents (Le and Mikolov 2014). Agglomerative clustering by cosine distance (Pedregosa et al. 2011) over the document vectors can give us groups of documents that resemble each other in meaning. I attempt to classify the document vectors as being generated by tweets written by List or SMD MPs, and interpret the accuracy of the classifier (Peterson and Spirling 2018) as a systematic difference between List and SMD MPs. I believe that the extent of clustering here can also tell us about relative levels of party unity in communication strategy, but this discussion is beyond the scope of this paper.

However, similarly to PCA, `doc2vec` requires a large number of documents to achieve these levels of sentiment and semantic recognition. Because my dataset is small compared to standard applications for this algorithm, I will implement validation checks to test whether it has captured substantively meaningful links between

[9] For a concise and helpful explanation of word vectors, see Rodman (2019), pgs 5-9.

documents. In the absence of these checks, I cannot state whether classifier accuracy itself is measuring polarisation in the corpus, or meaningless noise.

Data

New Zealand

I use New Zealand as a case study for this paper because their MMP electoral system allows me to compare politicians from the same party and different electoral contexts. The current system, passed by referendum in 1993 and implemented in 1995, has a unicameral legislature with 120 MPs. 64 MPs are elected in single-member districts by plurality rule. The remaining 56 seats are divided between a 49-seat nation-wide closed list PR constituency, and seven SMD Māori electorates. The number of Māori seats varies between elections, and is proportional to the number of indigenous Māori people signing up to vote on the Māori voter role instead of the general one[10]. Given that Māori seats are voted in using SMD, I treat them as having the same electoral context and incentives as the other SMD seats.

Assignment to being a List or SMD MP is not randomised. Heitshusen et al (2005) note "*in interviews with list MPs [in New Zealand], it became clear that many would prefer to be electorate MPs. Indeed, MMP contracted the number of districts; therefore, some sitting MPs had to 'settle' for getting elected via the list. While some of these MPs were resigned to (if also unsettled by) their new list role, a few were trying to regain a district seat.*" Although this may be less applicable 14 years on, it is likely that politicians that become List MPs are inherently more ideological, or that politicians that become SMD MPs are inherently more constituency-focused. Ironically, as this choice is absent in non-mixed electoral systems, this self-selection effect will not exist in other contexts, but if we look only at non-mixed systems then it would be impossible to isolate country fixed effects from electoral ones. As such, I do not interpret my models as causal.

Currently, there are five parties and one independent MP in parliament. A minority coalition of the Labour Party and New Zealand First are in government, with a confidence-and-supply agreement with the Green Party. Jacinda Ardern (Labour, Mt. Albert) is Prime Minister.

Notable events in New Zealand politics within the time period captured by my data (since 1 Jan 2019) include the Christchurch Mosque Shooting (15 March 2019), a subsequent major reform of New Zealand's gun ownership laws, including a ban on semi-automatic weapons, and the failure of a flagship housing development policy KiwiBuild to create 1,000 homes by 1 July. I expect to see one or more of these events appear in the results of the models.

Collecting Tweets

In order to obtain the tweets of MPs, I downloaded a list of the MPs and their electorates[11], then found their corresponding Twitter usernames, or "handles," manually via a web search. Where it was not clear whether a handle was the official account of a politician, or access to the account required permission of the owner, the handle was discarded. A full list of the handles can be found in the Appendix.

Not all politicians use Twitter; prior research shows that particular types of politicians are more likely to to incorporate Twitter into their communication strategy. Jungherr (2016) provides a literature review of 127 studies looking at the use of Twitter in election campaigns, noting that the majority of studies find that the following factors increase the likelihood of a politician using Twitter: belonging to the opposition, belonging to a major or established party, having a large campaign budget, being young, representing an urban constituency, or having strong ideological views. Because these factors could bias my data collection, I check whether certain parties are over-represented in my dataset. Fortunately in New Zealand 87.5% (105 out of 120) MPs have official Twitter accounts, with all five parties (and the one independent MP) represented, allowing me to collect a balanced dataset.

[10] The Māori constituencies for 2017 are Hauraki-Waikato, Ikaroa-Rāwhiti, Tāmaki Makaurau, Te Tai Hauāuru, Te Tai Tokerau, Te Tai Tonga and Waiariki.

[11] From https://www.parliament.nz/en/mps-and-electorates/members-of-parliament/.

These handles were then used with the Twitter Search API to collect all of the tweets on that account from 1 January 2019 to 18 July 2019. The raw data has been saved in `json` format and is included in the replication materials.

The 35,945 tweets were then classified as retweet (a verbatim re-posting of another user's tweet), quote (a response with a framed box including the tweet to which the quote is responding to), or regular. Retweets and quotes were omitted from the dataset because my model does not have a clear way to deal with interactions, or multiple authors within a single document. Barberá et al. (2015) provide ways for measuring these interactions.

The remaining 15,381 tweets were then converted to a tabular data format with columns for meta-data including Party, Electorate, name, time and date of postng, and the full text of the tweet.

The corpus contains mostly English texts, but also a number of Māori phrases. The use of Māori phrases in English documents can be understood as a politically relevant signal, i.e. an appeal to an ethnic minority constituency or the endorsement of inclusive policies supported by non-Maori voters.

Pre-Processing

In text-as-data models, pre-processing steps refer to the preparation of the text data for its conversion to a numeric format compatible with NLP models. The numeric representation of text almost always entails the loss of information, whether it be syntactic, semantic or otherwise. Although there are procedures considered to be standard practice (Lucas et al. 2015), decisions taken during these steps can drastically affect outcomes. I therefore explain all of my coding choices, especially where they differ from decisions taken by other scholars in similar contexts, in detail. Given that my documents were universally very short (280 characters or less), I made conservative choices when it came to removing tokens from the documents. Finally, none of the choices I made were perfect, and in each case I describe the shortcomings of my approach and a way that this may be circumvented.

First, whereas standard practice consists of removing all non-alphabetical characters, I instead converted all emoji to their unicode descriptions, e.g. the unicode character `U+1F60A` was converted to `smiling_face_with_smiling_eyes`. I contend that emoji contain relevant sentiment information for text models beyond what is contained in the non-emoji text portion of the document. A smarter version of this step could provide a dictionary classifying each emoji according to their sentiment and replace them instead with a token.

In a related paper (Grimmer 2016), the author discards all mentions of geographic locations. I believe that this loses crucial information that can help identify constituency focus in tweets. I therefore instead replace all mentions of an MP's own constituency with the token `OWN_CONSTITUENCY`. Given that we are interested patterns of MPs tweeting about their own constituency, and not specifically which constituency that is, I believe this to be a better solution. This step does introduce a degree of artificial heterogeneity between List and SMD MPs because I have introduced a token that can only occur in tweets by SMD MPs.

Thirdly, all URLs were removed. Twitter's API replaces all websites with a shortened URL of the form `https://t.co/xxxxxxxxxx`, thus creating unique tokens with no clear meaning on their own. In a larger project, this step could have resolved the URL each short URL pointed to and then used the base of that URL (e.g. https://**wikipedia**.org) in order to see what websites various MPs are referencing.

Fourth, all "@" mentions were removed because they introduced a very high level of sparsity and would have been removed at a later step. As such, I do not believe that this step had a significant effect on the result.

Whereas many text-as-data applications remove words based on a list of "stopwords" (Lucas et al. 2015), I used the part-of-speech (POS) tagger implemented in `spaCy` (Honnibal and Montani 2017) to remove all punctuation, symbols, word-parts, whitespace and determiners. This has the added benefit of retaining polysemic words such as "will" when they are used as a noun (e.g. "the *will* of the people"), but dropping it when it is used as an auxiliary verb indicating the future tense. In an alternative version of this step, I tried retaining only nouns, verbs and adjectives, but this removed too much of the data, reducing the power of my statistical tests.

Sixth, I stemmed the remaining tokens using the Porter algorithm as implemented in the nltk library (Bird, Klein, and Loper 2009). This consists of algorithmically removing the end of a word to reduce it to its base form, e.g. "computational" $\implies$ "comput." An alternative to stemming is lemmatisation, which converts words to their dictionary form, or *lemma*, using morphological rules based on an understanding of the relevant grammar.

The final step consisted of converting the array of pre-processed tweets to a document-feature matrix (alternatively known as the document-term, document-token, term-document, or document-word matrix, abbreviated here as DFM). The DFM records each of the documents, $d \in D$, as rows and each of the unique tokens, $w \in V$, as columns. The element (d, w) of the $D \times V$ matrix is the number of occurrences of token w in document d. This is also known as the bag-of-words model. After conversion, the lowest 1% of tokens by frequency were discarded. At future steps, where necessary to ensure that algorithm converged, I also removed all tokens with corpus frequency of less than 8 or 10.

Four different subsets of the data are used in the fitting of the Wordfish (Method 1) portion of this study. The first two are with the List and SMD subsets of the corpus (D_{List} and D_{SMD} respectively), the third includes the Labour and National subsets of the corpus ($D_{Labour} \cup D_{National}$), and the final with the full dataset. D contains 15,381 tweets, $D_{Labour} \cup D_{National}$ contains 10,447 tweets, D_{List} contains 7,292 tweets, and D_{SMD} contains 8,089 tweets. There are a total of 2,612 unique tokens in V, 1,959 in $V_{Labour \cup National}$, 1,669 in V_{List}, 1,604 in V_{SMD}, and 1,290 in $V_{List} \cap V_{SMD}$. For all other models, the full $D \times V$ DFM was used, with a total of 15,381 tweets and 2,612 unique tokens.

All of the above steps were carried out using in Python using the pandas (McKinney and others 2010), spaCy, and nltk libraries. The resulting DFM was converted to csv format for compatibility with various R libraries.

Two alternative datasets were prepared but are not presented in the results. One concatenates all documents by author, and another by author and week. This was done to try and increase the length of documents, but ultimately did not provide clear or helpful results.

Results

Pre-Analysis

Before running the models, I used the random forest algorithm to classify documents as either belonging to List or SMD MPs, and used the mean GINI decrease of each feature to understand which individual tokens were the most helpful in classifying a document as either List or SMD. The mean GINI decrease essentially tells us the extent to which splits in individual decision trees made on that feature lead to correct predictions. Thus features with high mean GINI decreases are, ceteris paribus, good indicators of a token belonging to one group or another.

The top 30 features by mean GINI decrease are listed in Table 1. It can be seen that for instance, the word stem "human" (thus including words such as humans, humane, humanity) had the greatest differentiating power, occurring mostly in List MPs' tweets. This may be related to discussions of human rights, an inherently ideological issue; confirming is possible by re-preparing the data with entity recognition.

Similarly, the word stems "women," "democraci" and "race" are all more likely to occur in documents produced by List MPs; these all suggest ideological tweets. In contrast, the word stem "question" is more likely to occur in an SMD tweet, suggesting that SMD MPs are more likely to question government policy. This is in line with the expectations produced by Carey and Shugart (1995) and the results in Martin (2011).

Table 1: Top Terms by Random Forest Classifier Mean GINI Decrease Score.

Feature	Importance	Feature Freq.	Document Freq.	SMD Feat. Freq.
human	0.045534	108	99	18
red_heart	0.027466	186	161	34
ora	0.018778	125	122	31

Feature	Importance	Feature Freq.	Document Freq.	SMD Feat. Freq.
grinning_face_with_big_ey	0.018435	50	49	2
green_heart	0.017719	175	142	5
women	0.017270	209	177	61
democraci	0.016733	59	57	9
dark_skin_ton	0.016635	138	111	6
race	0.016384	83	71	19
transport	0.015904	215	184	68
ye	0.015167	389	381	255
voic	0.014688	97	91	17
platform	0.014290	45	44	5
crime	0.014095	68	56	13
phil	0.013867	100	93	89
regul	0.013576	66	58	11
need	0.012964	661	609	252
vehicl	0.012113	88	70	14
super	0.012071	37	36	3
mell	0.011522	51	39	45
thumbs_up	0.011098	129	120	104
live	0.011085	249	233	89
zerocarbonbil	0.010649	31	31	2
also	0.010437	415	406	152
clean	0.010206	66	66	14
safe	0.009917	155	142	44
youth	0.009797	134	97	39
question	0.009673	311	294	208
tbh	0.009503	39	39	5
nzqt	0.009297	28	28	27

Results: Wordfish

As mentioned, four different subsets of the data are used to fit the Wordfish mdoels. The first contained the full dataset, and is used for maximal information and comparison (Figures 1 and 5). The second, which contains only tweets from the two main parties Labour and National, is used to compare the distribution of $\hat{\theta}_{List}$ and $\hat{\theta}_{SMD}$ for MPs of the same party (Figure 4). The third and fourth are fit with the List and SMD subsets of the data, and have been fit in order to calculate a $\hat{\theta}$ that only measures variance along a dimension which only one of List or SMD MPs vary along (Figures 2, 3 and 5). In other words, $\hat{\theta}_{List}$ measures a dimension that captures variance within List MPs, and $\hat{\theta}_{SMD}$ measures a dimension that captures variance within SMD MPs.

Figures 1-3, colloquially known as Eiffel tower plots, plot the individual terms on β against ψ. Words higher on the y-axis, ψ, occur with higher frequencies. Words with higher values of β discriminate the words as being further "right" on θ, whereas words with lower values discriminate the words as being further "left" on θ. The direction of left-right in this model is somewhat arbitrary; the values could be flipped and the output of the model would be fundamentally the same. When fitting the model, I selected a Labour Party tweet as a reference point for the "left" direction, and a National Party tweet as a reference point for the "right" direction.

In the first figure, all emoji are highlighted. In the second and third, the terms "muslim," "terror," "christchurch," "crime," "vehicl," "regul," "kiwibuild," "clean" and "youth" are highlighted.

Figure 1 presents an interesting result: politicians on the left-hand side of the political spectrum are far more likely to use emoji in their tweets. Although this seems trivial, this may reflect an attempt by Green Party voters to appeal to a younger constituency, or to have younger, "media-savvy" users amongst their

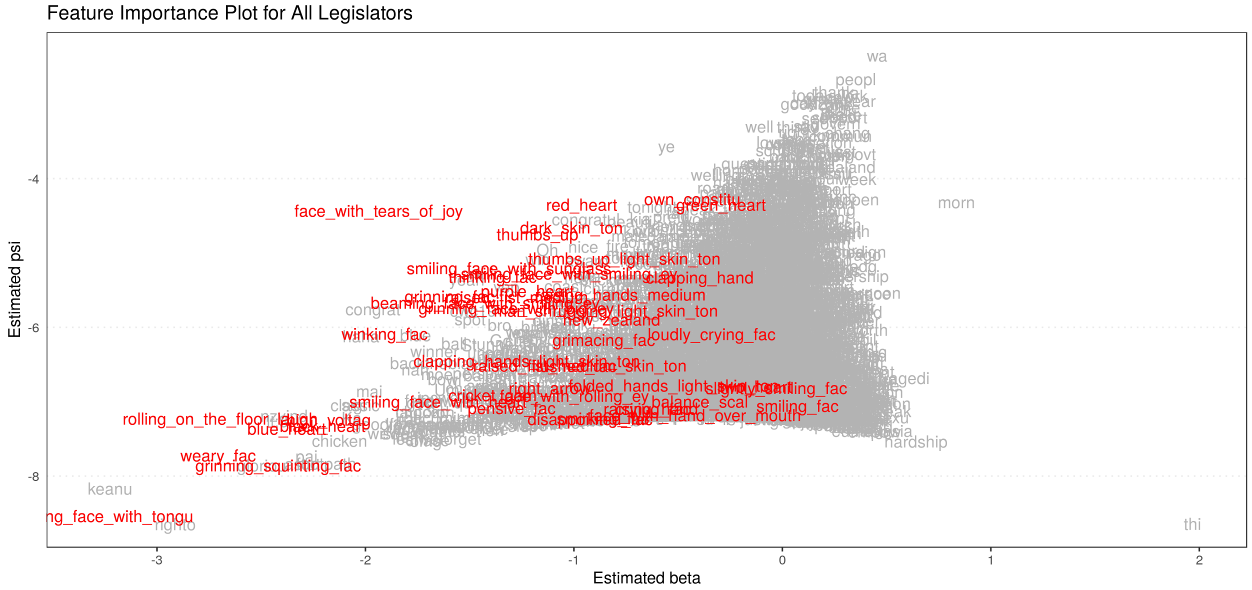

Figure 1: Feature Importance Plot for all Legislators

ranks—within the dataset, there are a large number of tweets by Green Party MP Golriz Gharaman, who also uses many emoji.

A comparison of figures 2 and 3 reveals interesting patterns in word usage. Firstly, the term terror is close to the right-hand edge of the distribution amongst List politicians whereas it is further left amongst SMD politicians, and lower in frequency. This likely reflects that the nationalist right-wing party New Zealand First has only List MPs, but also reveals that amongst the parties that have SMD seats (Labour, National, ACT, and an independent), the issue is more associated with the political left (i.e. Labour).

The tokens "clean" and "vehicl" are all higher-frequency amongst List MPs, indicating that proportionally, List MPs discuss transportation and environmental policy more than SMD ones. "Christchurch," referring to either the largest city in New Zealand, or more likely, the mass shooting that occurred at a mosque on 15 March 2019, is used roughly equally by both groups and does not appear to be more associated with the political left or right. It is interesting to note that "kiwibuild," a housing development scheme pursued by the Labour Party, does not occur in List tweets.

The final point to note is the token `own_constitu` occurs with a high frequency and is not particularly associated with the left or right.

Figures 4 and 5 compare the distributions of the fitted thetas (document positions) between List and SMD MPs of the same party. Figure 4 uses thetas fit with the Labour and National Party subset of the data, as these are the only parties to have both List and SMD MPs. The upper plot shows a histogram and kernel density estimate of $\hat{\theta}$ for Labour Party MPs, and the lower plot shows the same for National Party MPs. SMD MPs are shown in blue and List MPs in orange.

The main result to note in figure 4 is the difference in the distributions of the fitted values of theta for List ($\hat{\theta}^{List}$) and SMD ($\hat{\theta}^{SMD}$) for each party. I test the hypothesis that the two samples are drawn from identical distributions using a two-sample two-sided Kolmogorov-Smirnov test, and find that whereas for Labour $p \approx 0.61$, and therefore we cannot reject the hypothesis that the samples are drawn from the same distribution, this is not the case for the National Party ($p \approx 0.000$). It is also notable that the distributions of Labour and National are located in roughly equal positions on the left-right dimension captured by $\hat{\theta}$. This is surprising, given that the National Party is generally considered to be to the right of the Labour Party.

In figure 5, I compare the distributions of three different fits of the Wordfish model, denoted by a *subscript* on θ, between parties (y-axis), and between List and SMD MPs, denoted by a *superscript* on theta. The objects

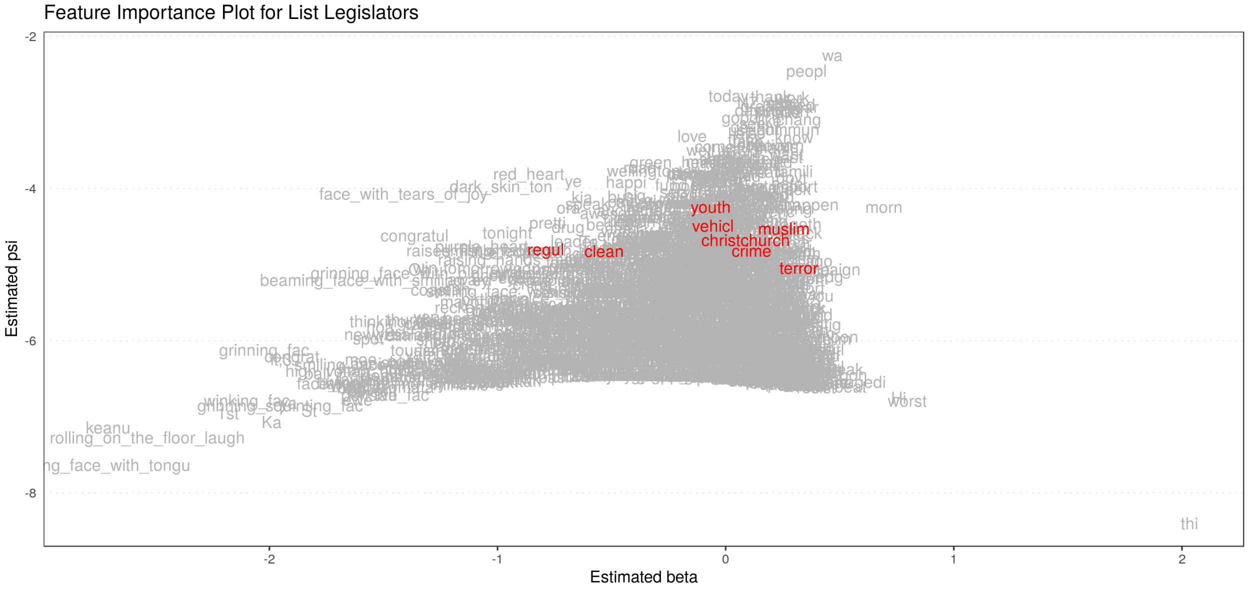

Figure 2: Feature Importance Plot for List Legislators

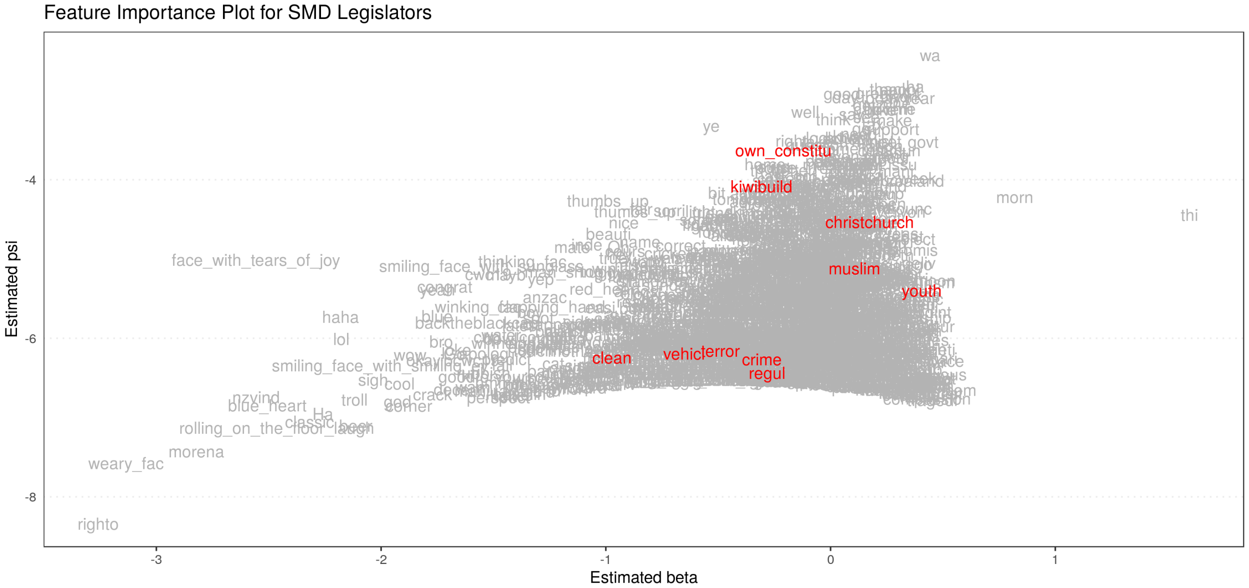

Figure 3: Feature Importance Plot for SMD Legislators

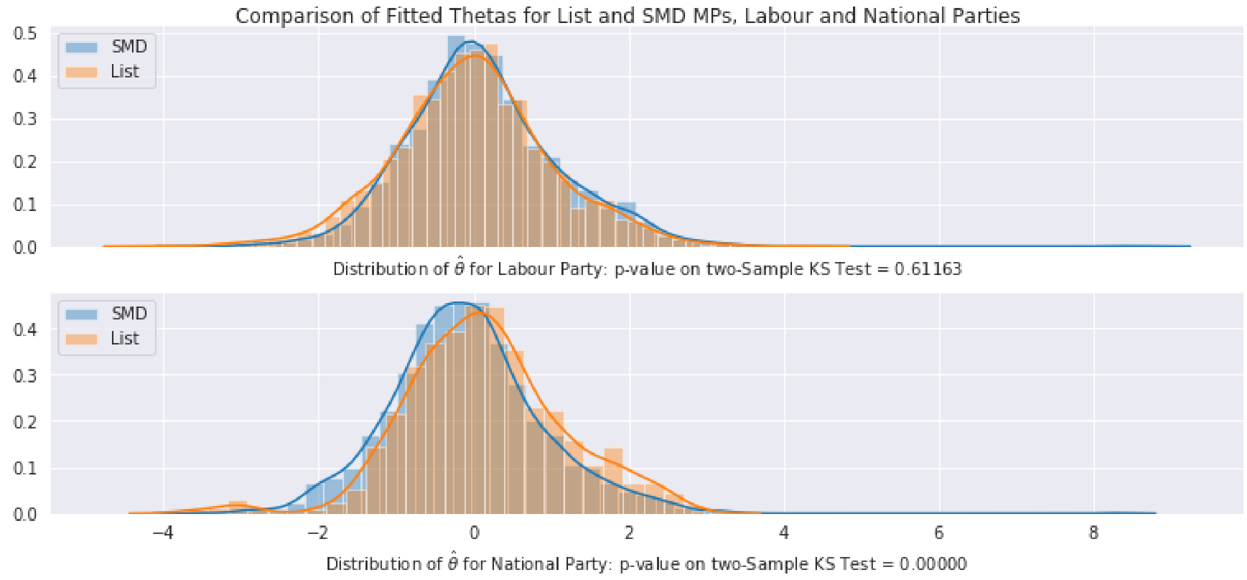

Figure 4: Comparison of Fitted Thetas for List and SMD MPs, Labour and National Parties

in the graph are violin plots, which show the kernel density estimate of a sample of a random variable. The blue violins show the distribution of theta for SMD MPs in the model fit only with SMD MPs, the orange violins show the distribution of theta for List MPs in the model fit only with List MPs, the green violins show the distribution of theta for SMD MPs in the model fit with all MPs, and the red violins show the distribution of theta for List MPs in the model fit with all MPs.

There are several things to note from this graph. The first is that for parties with both List and SMD MPs, the distribution of theta extends further to the right for SMD MPs than List ones (blue vs orange, green vs red). The second is that for all parties, the distribution of theta for the model with with all MPs (green and red) displays a higher variance than the distribution of theta for the model fit with only SMD or List MPs (blue and orange).

Testing the same hypothesis as above with the model fit for all MPs $\hat{\theta}_{Total}$ shows similar results to figure 4. We can reject the null hypothesis that $\hat{\theta}^{SMD}_{Total}$ and $\hat{\theta}^{List}_{Total}$ are drawn from the same distribution for the National Party ($p \approx= 0.000$), but not for Labour ($p \approx 0.653$). Note that it is not appropriate to conduct KS tests for $\hat{\theta}_{SMD}$ against $\hat{\theta}_{List}$ because the two samples are generated by different models, different datasets, and therefore different random variables.

In summary, the Wordfish model shows us that (1) emoji usage is associated with left-wing parties, (2) the terms "terror" and "crime" are more associated with the right for List MPs but not SMD ones, (3) the distribution of estimated document positions is statistically indistinguishable for List and SMD MPs in the Labour Party, but not for the National Party and (4) we cannot compare $\hat{\theta}$ between fits of Wordfish without an OOS implementation.

Results: Principal Component Analysis

I use the PCA and Sparse PCA algorithms to decompose the DFM, a 15,381 by 11,182 matrix, to 100 components (K). In comparing the two algorithms, the first thing to note is processing time: whereas the regular PCA components fit in under 10 minutes, the Sparse PCA component took well over twelve hours[12]. Thus if time is a concern, I strongly recommend the regular PCA algorithm.

Figure 6 shows the distribution of document-component weights for List and SMD politicians from the Labour

[12] Models were run on a laptop with a 8th-gen Intel i7-8550u core and 32GB of RAM, with multiprocessing on 8 threads.

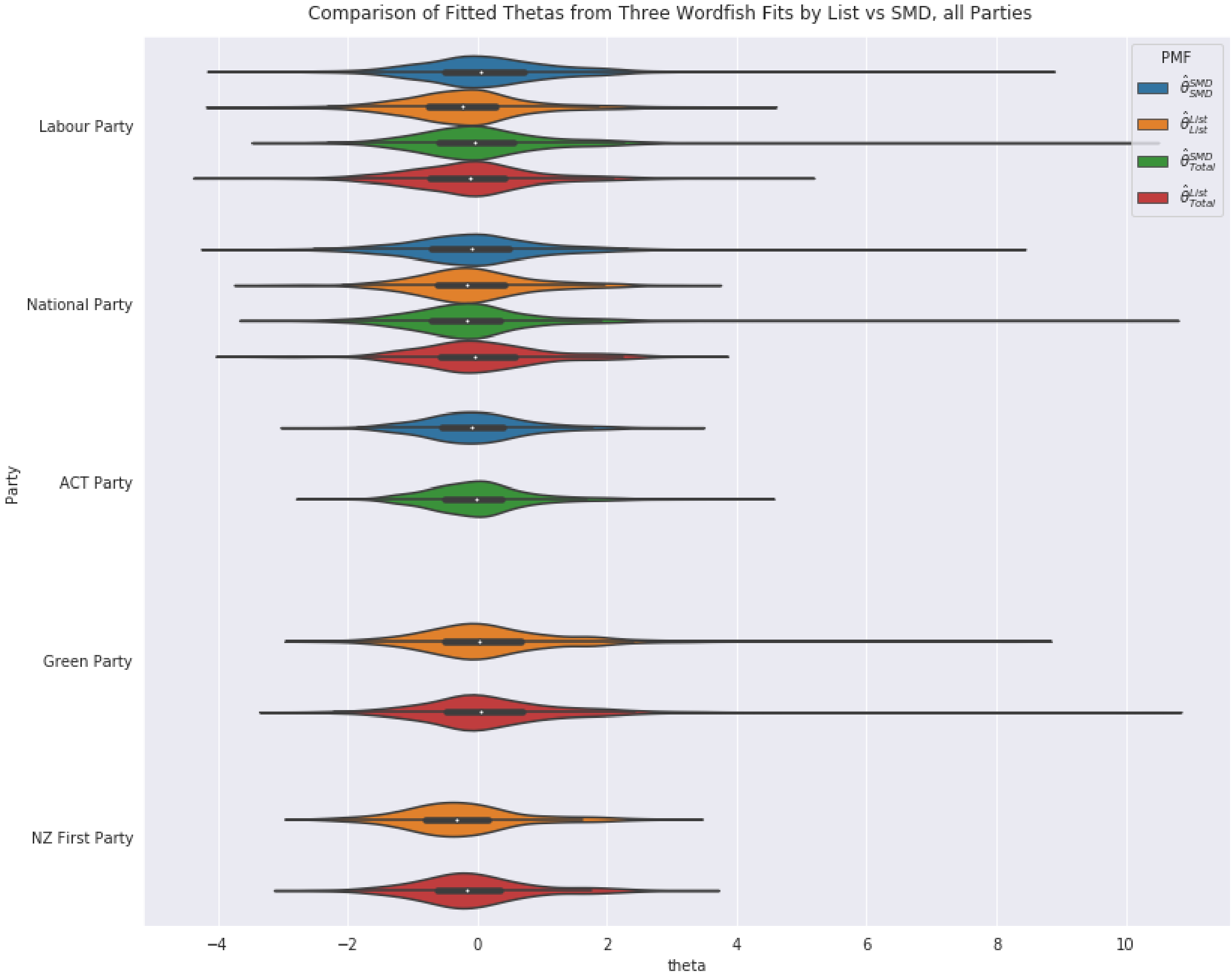

Figure 5: Comparison of Fitted Thetas from Three Wordfish Fits by List vs SMD, all Parties

and National parties, along with the p-value for the associated Kolmogorov-Smirnov test of the null hypothesis that the distribution of document-component weights are drawn from identical distributions. The subplots on the left use values computed by the PCA algorithm, and the subplots on the right show values computed by the Sparse PCA algorithm. For each, the five components plotted are the ones with the highest ranking of the `own_constituency` token in the $K \times V$ component-feature matrix. I use this approach for selecting components because intuitively, components that place higher weight on variance generated by the mention of an author's own constituency are detecting sources of variance associated with a constituency focus. This approach avoids selecting on the outcome, and instead tests *a priori* the theoretical framework of Strøm (1997), Carey and Shugart (1995), and others.

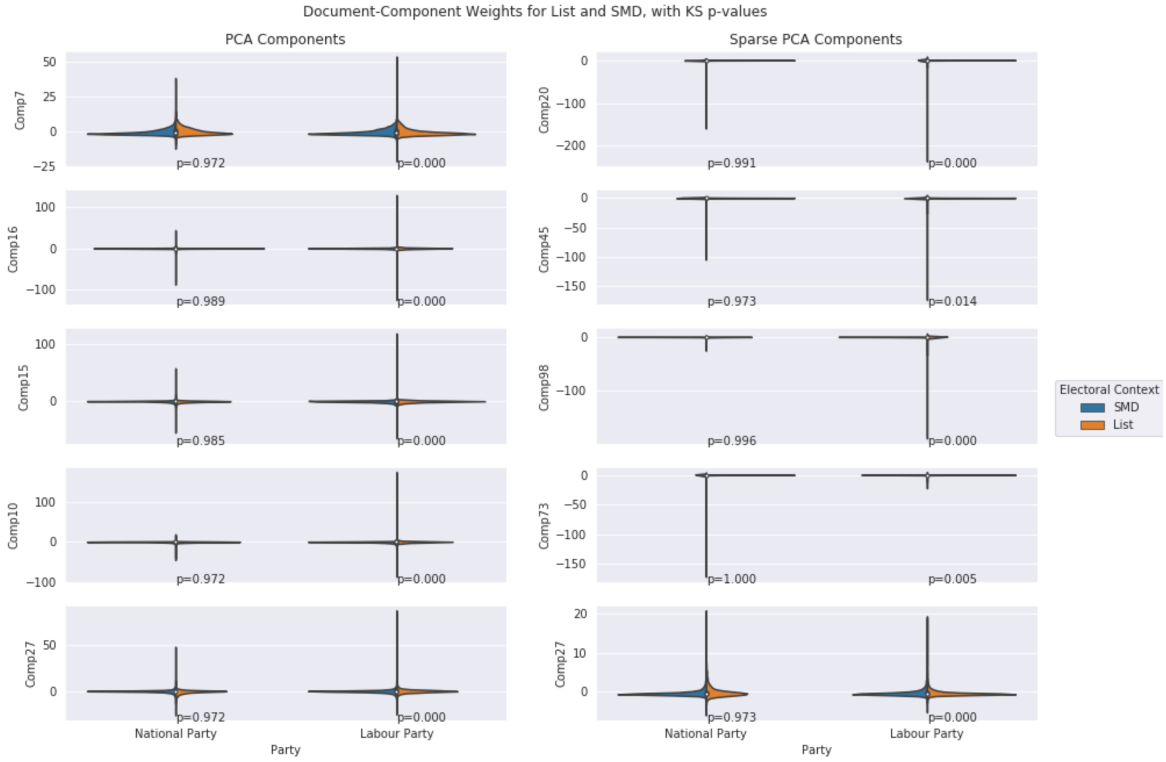

Figure 6: Document-Component Weights comparing List and SMD for Labour and National Parties, PCA vs Sparse PCA, for the Five Components Placing Highest Weight on `OWN_CONSTITUENCY` token, with Kolmogorov-Smirnov Test Statistic on List and SMD Distributions Being Equal. See Table 2 for Top Features.

The p-values indicate that there the distribution of document-component weights is significantly different for National Party MPs, but not for Labour Party. The only exception to this pattern is component 45 of Sparse PCA model. Thus along the latent dimensions captured by these components, National Party List and SMD MPs tweeted in a systematically different manner, but Labour Party List and SMD MPs did not.

Table 2: Top Weighted Terms for PCA and Sparse PCA

Rank	PCA	Sparse PCA
1	Comp7 - dorat france woolhandling sheering cha...	Comp20 - karapiro 1700 lake 61 tenei kahui spr...
2	Comp16 - tekau iwa okura lbp blog marine fu qi...	Comp45 - n't pounamu glimpse flew overh...
3	Comp15 - haggis hogget duds nbhs agrikids some...	Comp98 - ramblers puhoi browse performing back...
4	Comp10 - kemp slates hongi hika signature unes...	Comp73 - stomach firmly dorp bobby domi...
5	Comp27 - 140th ipu140 assembly ipu womenmps ca...	Comp27 - roads speeds infrastructure cycling k...

Greater difficulty comes from attempting to interpret these components. I present some of the top terms for each component in table 2. Looking at the top weighted terms of each component, it is not clear how to label the variance that many these components are identifying. Even where the label is relatively clear, such as component 27 of PCA, which is identifying tweets relating to the 140th Inter-Parliamentary Union assembly, how this then relates back to constraints generated by electoral context is less clear. The one exception may be component 27 of Sparse PCA, where the terms "roads," "speeds," "infrastructure" and "cycling" are ranked highly. For this component, there were statistically significant difference in usage patterns for National Party List and SMD MPs, with SMD MPs demonstrtaing a higher average usage frequency. Intuitively this makes some sense, given that those terms relate to an issue dimension that affects constituents, but may or may not be a priority of the National Party.

Figure 7 shows the per-week average document-component weight for each party for the first three components for both parties. This is computed by taking the conditional mean of the document-component weight provided by the $D \times K$ document-component matrix for each week and party.

$$\forall k \leq 3, T:$$

$$\mu_{\mathbf{P}} = \{\mu_{P,t=1}, \mu_{P,t=2}, ..., \mu_{P,t=T}\}$$

$$\mu_{P,t=t} = E[D \times V_k | d_P = Party, d_t = t]$$

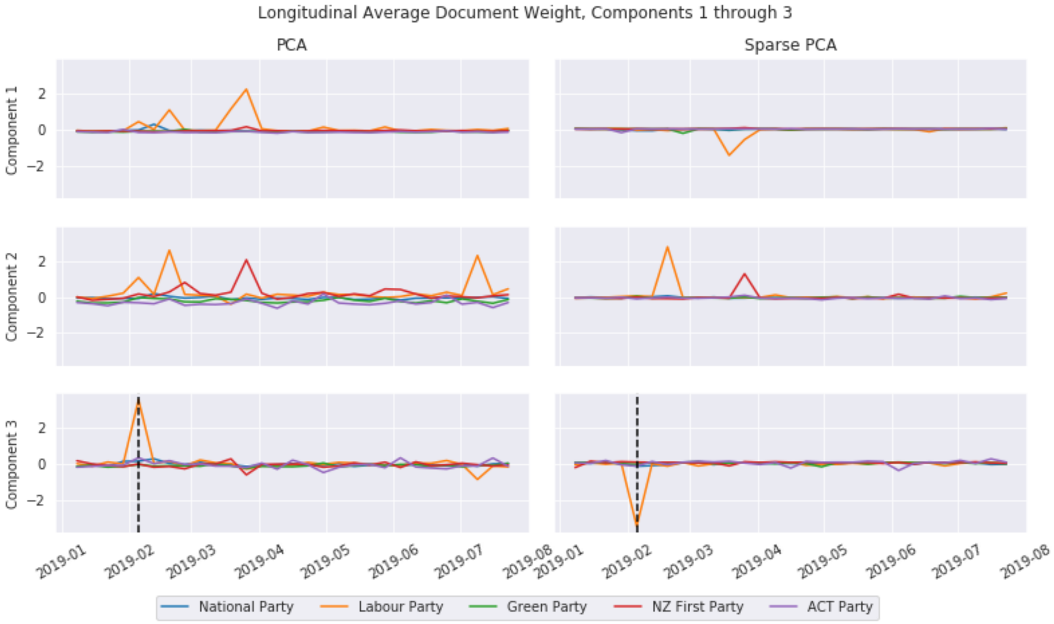

Figure 7: Longitudinal Average Document Weight, Components 1 through 3

Figure 7 shows the key issue with applying PCA or PCA-like models to the text data. Given that PCA finds the K uncorrelated components that explain the greatest proportion of the variance in the data, it is very sensitive to higher-level sources of variance in the data. For example, component 3 is driven by *language*. Around Lunar New Year (indicated by the vertical dotted black line in figure 7), several MPs included Chinese characters in their tweets. Other components may be being driven by *style*, but looking at the highest-weighted components gives little insight into how this may be interpreted. Thus, even if I find components along which List and SMD MPs tweet in a systematically different way, interpreting this difference is difficult.

This difficulty in interpreting the components also limits my ability to argue for the benefits of the Sparse PCA model over the regular one. Without a better method for linking these components to their substantive meaning, the only significant difference between the two models is processing time, which is usually not a major concern in a research setting.

Results: Topic Models

I fit two structural topic models (STMs) (Margaret E. Roberts et al. 2014b) with ten and twenty topics respectively ($k = 10$, $k = 20$). However, because fits of STM with $k > 10$ with my data generated topics with no clear meaning, I present the results for $k = 10$ only. During fitting, I used spectral clustering to find the globally "optimal" set of topics. Topic models suffer from multimodality, whereby depending on the random initialisation, one can get different results for the same data and number of topics. Although there is no empirical reason to prefer one mode over another, running the model with the parameter ensures a greater degree of consistency within my results. (For an extended discussion of the issue of multimodality in topics and the "optimal" number of topics, see Margaret E. Roberts, Stewart, and Tingley (2016)).

I present the ten topics, the labels I assigned them, and their top terms by four criteria[13], in table 3. Unlike the results in Grimmer (2016), I was unable to find a topic corresponding to credit-claiming or constituency focus. Nevertheless, the topics reveal a number of interesting patterns.

Table 3: Ten Topics with Assigned Labels and Top Terms by Four Metrics; Highest Probability, FREX (Frequent and Exclusive Terms), LDA Score and Lift Score

Topic 1	**Responding Directly**
	Highest Prob: `good, thi, thank, like, know, come, us, ye, help, realli`
	FREX: `Oh, feel, read, ye, realli, tweet, know, hope, thumbs_up_light_skin_ton, lol`
	Lift: `god, alarm, aw, bugger, content, DM, email, Ha, hero, lt;3,`
	Score: `thank, know, ye, us, good, read, like, feel, realli, come`
Topic 2	**Legislative Reform**
	Highest Prob: `peopl, can, make, need, say, right, want, thing, polit, better`
	FREX: `law, recommend, protect, system, drug, make, believ, polit, can, democraci`
	Lift: `argu, civil, consist, crimin, incit, intent, law, neutral, privaci, recommend`
	Score: `backtheblackcap, can, need, make, peopl, law, hate, right, system, say`
Topic 3	**Maori/Environment**
	Highest Prob: `one, take, also, parti, call, chang, vote, climat, keep, maori`
	FREX: `maori, kia, ora, action, te, word, keep, climat, voic, parti`
	Lift: `nga, zerocarbonbil, kia, medic, pai, represent, schoolstrike4clim, tau, te, action`
	Score: `maori, one, climat, kia, parti, vote, ora, te, also, kri`
Topic 4	**Commemoration**
	Highest Prob: `wa, go, well, last, hi, even, got, home, week, hear`
	FREX: `got, wa, thought, well, bit, photo, dog, went, done, told`
	Lift: `uncl, badli, dog, edit, humour, op, radio, sharp, sleep, tenni`
	Score: `wa, well, hi, got, last, week, sharp, go, thought, bit`
Topic 5	**Transport Infrastructure**
	Highest Prob: `think, use, said, point, road, car, way, may, wrong, differ`
	FREX: `car, vehicl, road, use, emiss, less, altern, benefit, reduc, speed`
	Lift: `frequent, mainli, stuart, user, altern, charg, cheaper, congest, effici, electr`
	Score: `think, car, use, road, vehicl, point, cost, reduc, said, cycl`
Topic 6	**Fiscal Restraint**
	Highest Prob: `govern, tax, new, govt, labour, school, fund, ha, public, budget`
	FREX: `govern, fund, budget, teacher, bu, spend, govt, project, growth, privat`
	Lift: `failur, partnership, pet, provinci, ptom, upgrad, valuabl, wage, agreement, billion`
	Score: `govern, tax, govt, fund, budget, smiling_face_with_heart, increas, labour, polici`

[13] These four criteria provided with the `stm` package are: "Highest Prob," the terms with the highest weights in the topical content matrix; "FREX," a method for finding frequent and exclusive topics; "Score," provided from the `LDA` package; and "Lift," provided from the `maptpx` package. Precise definitions and formulae are provided in the documentation for the `stm` package.

Table 3: Ten Topics with Assigned Labels and Top Terms by Four Metrics; Highest Probability, FREX (Frequent and Exclusive Terms), LDA Score and Lift Score

Topic 7	**Congratulating/Celebrating**
	Highest Prob: great, day, today, new, love, meet, morn, thank, see, MP
	FREX: team, own_constitu, celebr, visit, congratul, youth, thumbs_up, enjoy, host, award
	Lift: superb, air, ambassador, anniversari, ashburton, balloon, black_heart, breakfast, cafe
	Score: great, day, today, meet, celebr, thumbs_up, morn, love, congratul, team
Topic 8	**Events in Parliament**
	Highest Prob: thi, year, get, ha, time, nation, back, issu, first, sure
	FREX: committe, select, news, later, back, amend, month, year, sinc, ago
	Lift: automat, bottl, box, committe, expenditur, gcsb, later, octob, Rd, search
	Score: year, thi, get, committe, ha, bill, nation, Rd, back, sure
Topic 9	**Housing (Kiwibuild)**
	Highest Prob: look, minist, hous, face_with_tears_of_joy, PM, countri, question, next, red_hear
	FREX: face_with_tears_of_joy, red_heart, dark_skin_ton, wait, phil, twyford, look, PM
	Lift: copi, damn, decriminalis, herald, judith, ladyhawk, marijuana, newshub, raised_fist_mediu
	Score: hous, minist, dark_skin_ton, look, red_heart, green_heart, kiwibuild, PM
Topic 10	**Christchurch/Muslim Solidarity**
	Highest Prob: work, support, commun, mani, NZ, famili, peopl, live, part, announc
	FREX: health, christchurch, togeth, muslim, commun, mental, victim, famili, heart, terror
	Lift: islam, balance_scal, christchurch, condol, dialogu, heart, mental, porirua, prayer, resto
	Score: commun, christchurch, muslim, support, health, mental, eg, work, famili, violenc

Of the ten topics, the two I am most uncertain about are the fourth (labelled "commemoration") and first (labelled "Responding Directly"). In order to get a better idea of the kinds of tweets they were referring to, I looked at the top 100 tweets in the topical prevalence matrix for each of these topics. Of these, topic 4 ("commemoration") had a lot of tweets paying tribute to a New Zealander political journalist Rob Hosking, who passed away in late 2018. Topic 1 was characterised by tweets in which the MP seemed to be responding directly to (often hostile) tweets from non-MPs[14].

I first discuss topic-level trends. In order to validate my labels, I show, as do Grimmer (2016), Margaret E. Roberts et al. (2014a), Lucas et al. (2015), that certain longitudinal trends that we expect do appear in the data. In the figure 8, I show that the Christchurch/Muslims topic spikes just after the Christchurch Mosque Shooting[15] (indicated by a vertical black dotted line), which gives me greater confidence that this is the correct label. Other topics are harder to confirm in a longitudinal plot, especially where they remain relatively constant in prevalence over time.

Unlike LDA or Hierarchical Dirichlet Process (HDP, used in Grimmer 2016), STM allows us to visualise between-topic correlations (Lucas et al. 2015). I present these correlations in figure 9. In this figure, the size of the label for each node represents the overall prevalence of the topic in the corpus, and the thickness of the edge indicates the correlation between the nodes.

It can be seen that the Celebrating/Congratulating topic has the highest overall corpus proportion. Specifically, this implies that on aggregate, documents in the corpus are generated more by the celebrating/congratulating function than anything else. In plainer terms (but also committing an ecological fallacy), this means a plurality of tweets congratulate or celebrate. This result is not surprising, given that many politicians use Twitter to celebrate a local event, national holiday, the passage of a bill, and so on.

An interesting cluster in figure 9 is the transport infrastructure-fiscal restraint group, which have a pairwise correlation of 0.31. This suggests that where tweets urge fiscal restraint, it is often in conjunction with

[14] For space, I do not list these tweets here, but they can be viewed in the reproduction materials using the `get_nhighest_docs` function in the script `stm.R`.

[15] The jump appears to begin prior to the shooting, which is due to the use of a smoothing function for the graphing of the data.

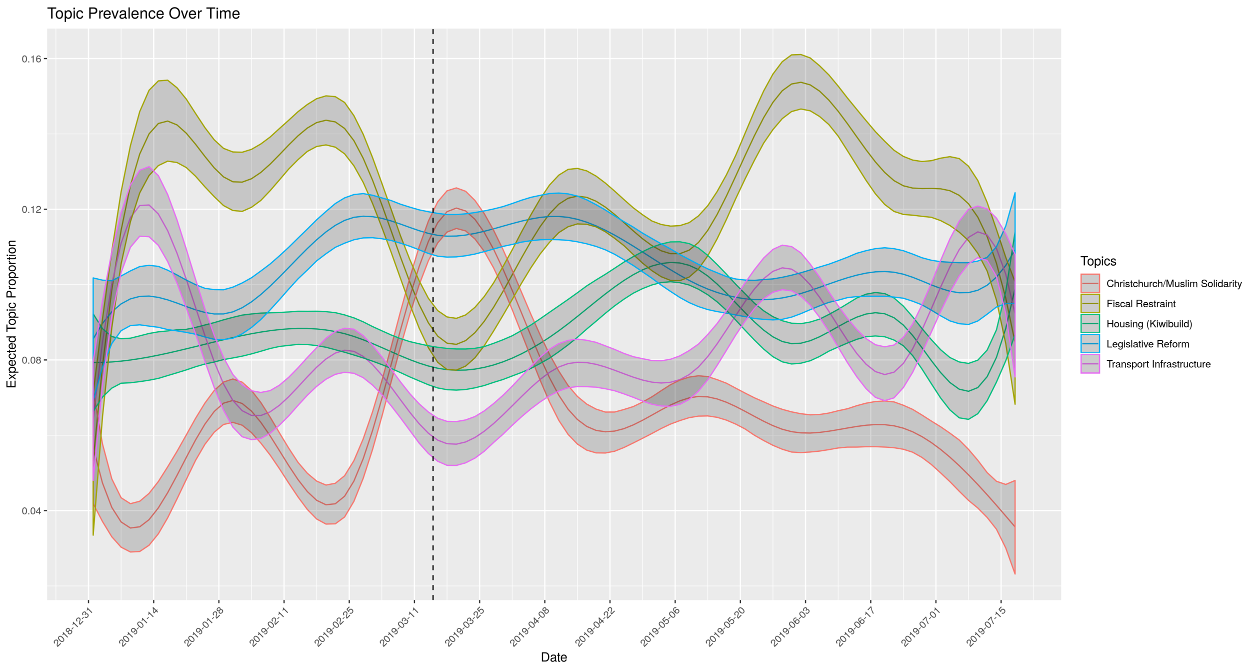

Figure 8: Topic Prevalence Over Time for Five Topics

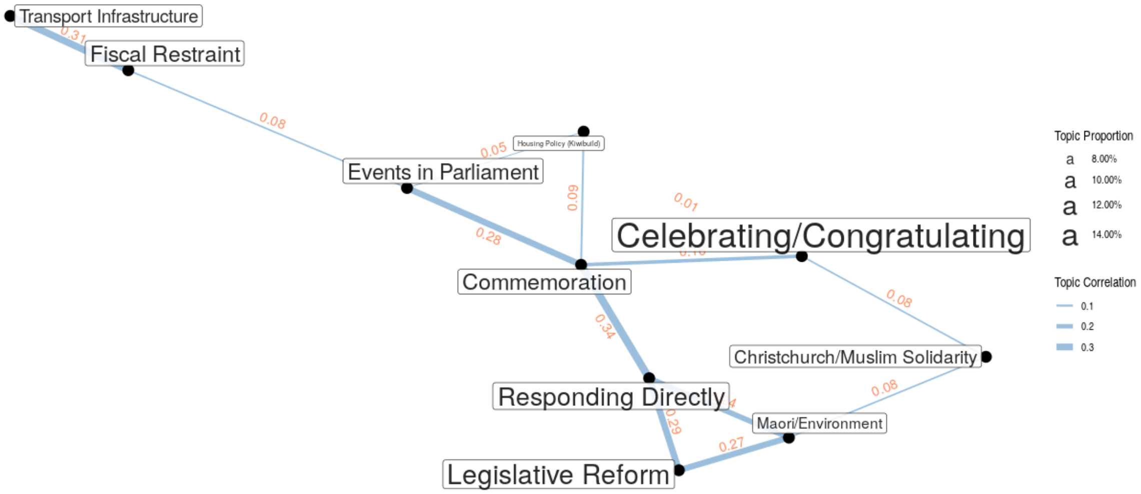

Figure 9: Topic Correlations

discussing transportation infrastructure policy. Given this, it is surprising that the Housing Policy/Kiwibuild topic is not correlated with the fiscal restraint cluster, as this topic is most prevalent in tweets by the opposition National Party.

I now discuss how these topics relate to the DGP. For a given document, the document-topic weights (referred to as θ herafter, where $\theta = \{\theta_1, ..., \theta_K\}$) add up to one, and indicate the "proportion" of the document that was "generated" by each topic. To get an intuition for what this means in practice, the following tweet is associated with the following values of θ, shown in figure 10.

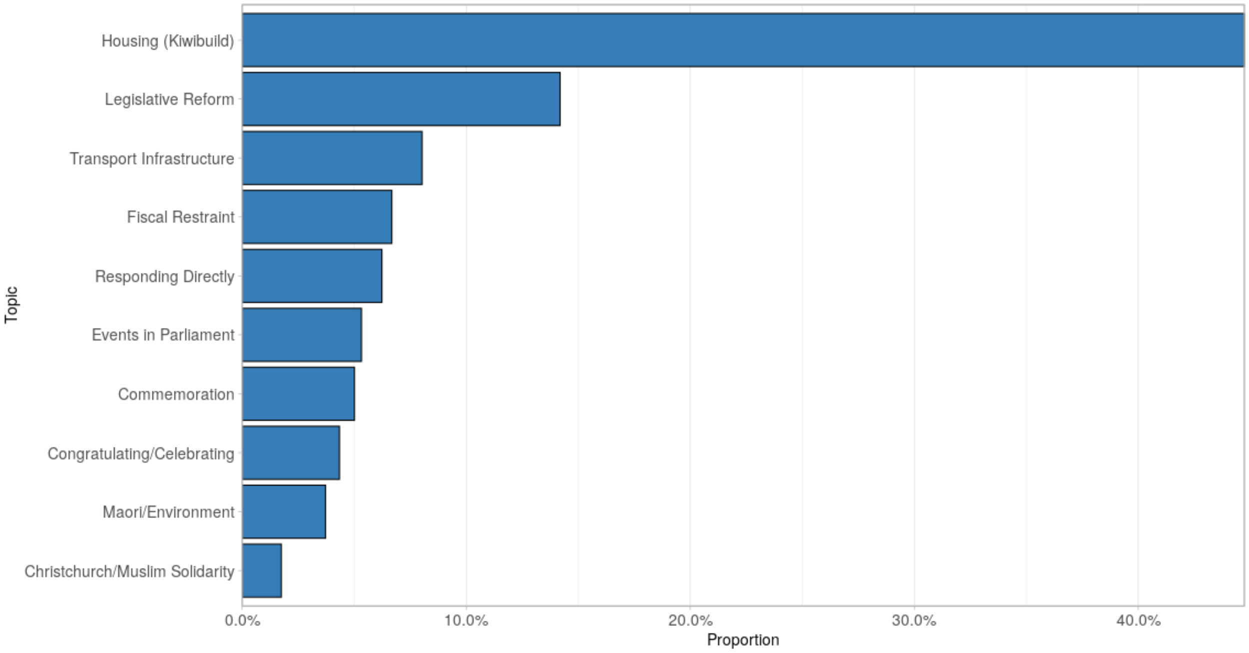

Figure 10: Example Topic Prevalence Proportions for Tweet: *"Can't see how guaranteeing prices for houses already built, can possibly be 'adding to housing stock.' Auditor-General investigators to look into Judith Collins' KiwiBuild concerns | https://t.co/ACiJunOW9j https://t.co/6O0z99mnMJ"*

In order to isolate the variation in communication strategy controlling for party and topic, I estimate a multinomial logit regression with document-covariates including electoral context as predictors and θ as the outcomes. I then calculate the difference in predicted topic weight ($\hat{\theta}_k^{List} - \hat{\theta}_k^{SMD}$) for List and SMD MPs with all other controls held at their median. The difference, with the associated 95% confidence interval, is plotted in figure 11.

Whereas in other models I was unable to easily control for date and party, here the fitting document-topic prevalence covariates and regression controls include party and date, meaning this difference is independent of party and date-specific trends. It may, however, be affected by an omitted variable that affects electoral context, party and θ.

Table 4 presents the covariate on the dummy variable for SMD for each of the ten topics. The stars in the column "Sig" indicate significance level; "***" indicates significance at the 0.1% level, "**" indicates significance at the 1% level, and "*" indicates signifiance at the 5% level. It can be seen in the table and figure that the effect of electoral context is significant at the 5% level for the topics "Responding Directly," "Legislative Reform," "Fiscal Restraint" "Maori/Environment," "Christchurch/Muslim Solidarity" and "Congratulating/Celebrating."

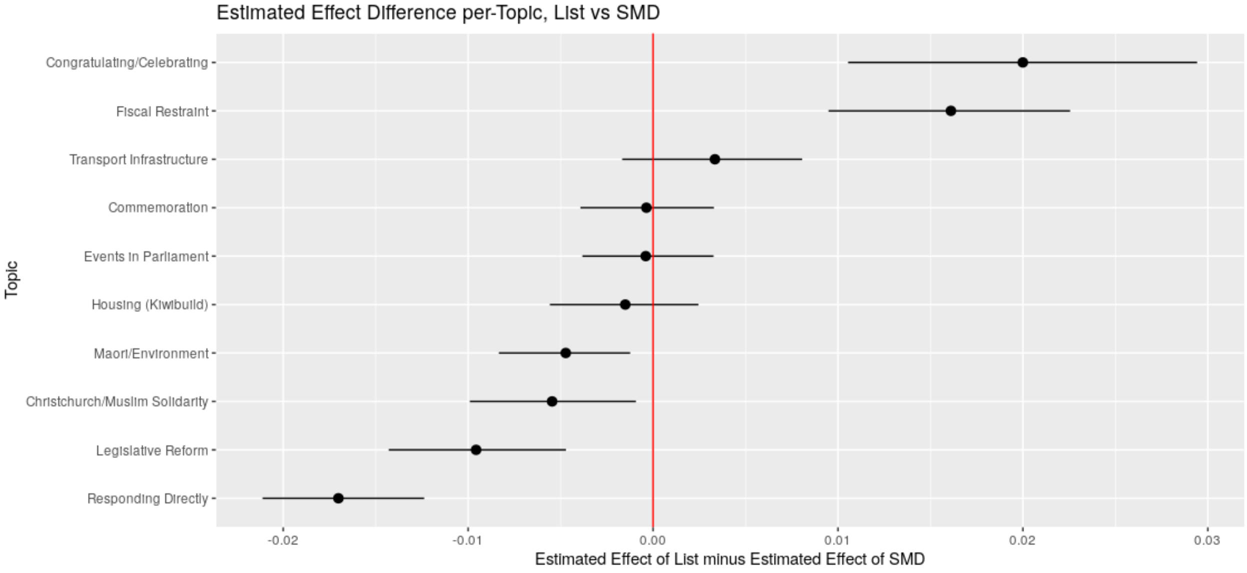

Figure 11: Estimated Difference in Topic Prevalence for List and SMD Candidates with 95% c.i.

Table 4: Estimated Effect of SMD Dummy in Multinomial Logit

Topic	Estimate	Std. Error	t value	Pr(>|t|)	Sig
Responding Directly	0.017	0.002	7.796	0.000	***
Legislative Reform	0.010	0.003	3.852	0.000	***
Maori/Environment	0.005	0.002	2.608	0.009	**
Commemoration	0.000	0.002	0.203	0.839	
Transport Infrastructure	-0.003	0.002	-1.347	0.178	
Fiscal Restraint	-0.016	0.003	-4.671	0.000	***
Congratulating/Celebrating	-0.020	0.005	-4.099	0.000	***
Events in Parliament	0.000	0.002	0.217	0.828	
Housing (Kiwibuild)	0.002	0.002	0.745	0.456	
Christchurch/Muslim Solidarity	0.005	0.002	2.334	0.020	*

The statistical significance indicates that for these topics, it is unlikely that the way that List and SMD MPs tweet is not connected to their electoral context (again, the direction of causality is unclear). A number of these results match intuitions we may have: SMD MPs are more likely to engage directly with individuals on Twitter, whereas List MPs are more likely to write tweets congratulating or celebrating events such as national holidays, which are not specific to any constituency. Commemoration tweets, especially when commemorating individuals such as journalist Rob Hosking, are unrelated to electoral constraints and therefore the effect of electoral context has no effect.

Other results are not so intuitive: why are List MPs more likely to criticise spending ("Fiscal Restraint"), and why are SMD MPs more likely to advocate legislative reform? I discuss this further in the conclusion.

Results: Document Embeddings

I fit the `doc2vec` model (Le and Mikolov 2014) as implemented in the Python library `gensim` (Řehůřek and Sojka 2010). The pre-processing steps differed slightly for this model because the `doc2vec` model does not take a DFM as its input. I therefore used the texts prepared in the manner described in the *Pre-Processing* section, but skipped the final step where word frequencies were used to construct the DFM, and instead passed these texts to the default pre-processing methods implemented in `gensim`. In accordance with standard practice, I set the length of the document vectors to the rounded fourth root of the number of features—fifteen. During training, I fit the model with a variable number of epochs (i.e. passes over the data), but found that the result did not vary significantly past 800 epochs.

I tried several methods for analysing the resulting document vectors. Firstly, I applied a hierarchical clustering algorithm by cosine distance, and then measured the extent to which these clusters corresponded to List and SMD MPs. Having tried with every number of clusters between 2 and 1000, I can state that I was unable to classify documents according to whether they were written by a List or SMD MP based on their document vectors.

I visualise this clustering in figure 12. Figure 12 has two parts: the dendrogram on the top shows the hierarchical clustering of documents by cosine distance, and the coloured bars on the bottom show how these clusters correspond to electoral context and party labels. If cosine distances in the latent vector space generated by `doc2vec` had substantive meaning that was affected by electoral context or political party, then these coloured bars would be sorted into solid bands of colour. It is clear from looking at the bars that they have not been sorted in any meaningful or helpful way.

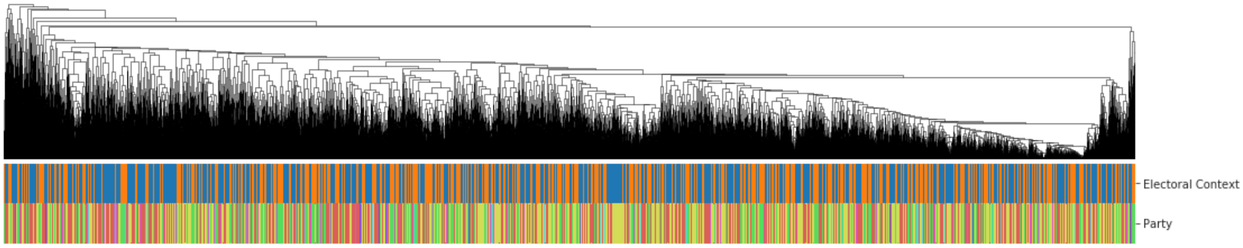

Figure 12: Hierarchical Clustering of Documents on Cosine Distance, with Electoral Context and Party Labels

As a final check on whether cosine distances in the document vector space have any substantive meaning whatsoever, and in order to validate my inconclusive results as being due to the model, I calculate a pairwise-matrix of cosine distances between document vectors, and for each vector that uses the special token `own_constituency`, I find the document corresponding to the closest vector. I present some of these results below:

```
---------------------------
EXAMPLE #1:
ORIGINAL TWEET:

Wishing  OWN_CONSTITUENCY  boy Kane and the  team all the best ahead of the game of their lives,
you've got the whole nation behind you and Kiwis couldn't be more proud.
_crossed_fingers_ #BACKTHEBLACKCAPS #CWC19

CLOSEST MATCH:

Twyford deliberately asked for this meeting to be "political with no officials" so he could discuss
the highly contentious Rural-Urban Boundary.

I don't buy the argument it was a mistake

---------------------------
---------------------------
EXAMPLE #2:
ORIGINAL TWEET:

Gong Xi Fa Cai! A big crowd turned out to celebrate the Chinese New Year at the Botany Town Centre today
with some wonderful performances by the  OWN_CONSTITUENCY  Chinese Association!
Xin Nian Kuai Le!

CLOSEST MATCH:
```

```
Yeah I include that in our bollocks tax haven conditions so to speak.

---------------------------
```

There is no clear relationship between the original tweet and its closest match in either of these cases (or any of the 171 others). As such, I am fairly confident that I was unsucessful in capturing a dimension of interest using `doc2vec`, and this is due to the fact that my dataset was too small to be used with document embeddings.

Although Peterson and Spirling (2018) use classifier accuracy as a measure of polarisation, I do not interpret the low accuracy of this classifier as indicating high polarisation. The manual validation of the fit indicates that it is not clustering on a dimension of interest.

Discussion and Further Research

I had two main aims when conducting this research. The first is to contribute to the literature in understanding the effect of electoral system on legislator behaviour and within-party dynamics. The second is to evaluate the effectiveness of a combination various increasingly popular text-as-data methods with Twitter data to exploring these relationships. I assess how successful I am in achieving each of these aims in turn.

My results show four ways in which the communication strategies of New Zealander parliamentarians is related in a systematic way to electoral context. The first result, shown by several models (Wordfish, STM), is that although the content and focus of List and SMD politicians differs in a statistically significant manner, the effect is more consistent for the opposition National Party than for the governing Labour Party (figures 4 and 5), and more pronounced for particular topics (figure 11 and table 4). It is possible that the effect of electoral system on communication strategy is therefore conditional on whether the legislator's party is in government or the opposition. This may be because MPs in government are competing for cabinet positions, and thus have a greater incentive to maintain party unity, but it may also just be specific to the Labour and National Parties of New Zealand. Testing this hypothesis requires data from other time periods and countries, to see whether the conditionality of this relationship applies in other cases.

The second result relates to how function differs between List and SMD politicians. I show that this difference in Twitter usage is strongest for tweets that are congratulating or celebrating, where List MPs are more likely to engage in this behaviour on Twitter than SMD MPs. On the reverse, SMD MPs are more likely to write tweets directly responding to individuals (figure 11 and table 4). Neither of these results are explicitly predicted by past authors, but both are consistent with the mainstream frameworks of Strøm (1997) and Carey and Shugart (1995).

The third result is less obviously explained by the existing literature. The STM showed that List MPs are more likely to criticise spending, whereas SMD MPs are more likely to advocate legislative reform. I suggest that this is because parties tend to have fewer policies they advocate than policy proposals that they are against. Specifically, parties tend to present a single solution to a given issue, and reject the solutions presented by other parties unless they are in a coalition with them. Even where parties are ideologically aligned, they will attempt to distinguish their policy proposals in order to compete for votes. As a result, while List MPs are limited in the number of policies they can advocate to the ones officially advocated by the parties, they are more able to criticise any policy not advocated by their party. On the contrary, SMD MPs may have greater incentive to build a personal reputation as an individual that presents solutions to issues, and face less risk from angering the party elite by advocating reforms that are not strictly in line with the official policies of the party.

The fourth result relates to image and emulation. Word frequencies and Wordfish find that emoji usage on Twitter is associated with left-wing parties in New Zealand. Understanding whether this pattern is widespread may give insight into the relationship between ideology and image management by elected representatives. A related research agenda is to understand whether these differences are consistent across different types of media, such as press releases, floor speeches and televised addresses. While obviously none of these will have emoji, it may be the case that left wing parties are more consistent in their attempts to emulate younger voters in rhetoric.

The models I used had varying levels of success. The greatest challenge in extracting theoretically useful dimensions from the text data was controlling for higher-order sources of variance in language [language, style and topic; Lauderdale and Herzog (2016)]. While negative and inconclusive results may be difficult to publish in a refereed journal, I present them in this dissertation in order to provide insight for other researchers into what kinds of questions each model is best for, what requirements each model has of the data, and the limitations of the output of each model.

Wordfish (Slapin and Proksch 2008) provides insights into how word frequencies can be associated with particular positions along a single ideological spectrum, but suffers from multiple limitations. The first is that this single dimension conflates ideology, topic, and intended audience, making it impossible to discern which of these was driving differences between List and SMD candidates. This issue can be mitigated by an OOS implementation of Wordfish, which I outline how to do, and will develop in the near future.

PCA and Sparse PCA are both highly sensitive to "outliers." In this case, these outliers were lexical fields that were longitudinally highly localised, such as the example of Lunar New Year (figure 7). Although this result was unhelpful for understanding broader patterns and trends within the data, it may be a useful tool for researchers who are specifically focused on outliers. It may also be possible to utilise this method for larger datasets, where stricter pre-processing rules can be implemented to smooth longitudinally localised variance.

The structural topic model (Margaret E. Roberts et al. 2014b) was the most successful in helping us understand the effect of the electoral constraint on the DGP. By combining the topical prevalence matrix with a multinomial logit, I was able to isolate the variance due to electoral system from the higher-order variance in a way that was not possible with other models. As a result, the majority of the substantively interesting results in this conclusion are based on results found using STM. For future research, a hierarchical variant of the structural topic model that can include document- and feature-prevalence would be of great use. There were topics that I wanted to break down further, such as the style topics commemoration, congratulating and responding; knowing the substance of what these documents are commemorating, congratulating or responding to would allow me to be more precise about the nature of the differences in communication strategies between List and SMD MPs.

Document embedding was disappointing, and its failure to provide meaningful results is likely due to the small-corpus nature of this research project. Given, however, that much of political science research is small-corpus by computational linguistics standards, I suggest a number of approaches for overcoming this problem that I will test in future research. The first improvement is to increase the amount of document meta-data used as features; this will help provide contextual information that is otherwise absent. The second is to combine `doc2vec` with topic models, and include the topical prevalence vector $\hat{\theta}_k$ in the features. This gives additional information on variance due to topic, and may increase the predictive and descriptive power of this approach. To my knowledge, `doc2vec` as been combined with LDA (`lda2vec`, https://github.com/cemoody/lda2vec), but nobody has implemented the a combination of the more useful `stm` with `doc2vec`.

All of these methods are ultimately best-suited to discovering patterns in high-dimensional text data that would otherwise be undetectable. In future work, I would recommend a mixed-methods approach combining in-depth case study to understand the direction of causality and manual content analysis to gain a more intuitive understanding of particular patterns that the models highlight. I argue that unsupervised approaches are fundamentally descriptive, and can provide evidence within a causal framework as a causal process observation, but none of the methods discussed here satisfactorily prove causality on their own.

Appendix: Additional Tables

Table 5: List of New Zealander MPs, their Electorate, and Twitter Username

Surname, Firstname	Party	Electorate	Username
Adams, Amy	National Party	Selwyn	amyadamsMP
Allan, Kiritapu	Labour Party	List	KiriAllan
Andersen, Ginny	Labour Party	List	ginnyandersen
Ardern, Jacinda	Labour Party	Mt Albert	jacindaardern
Bakshi, Kanwaljit Singh	National Party	List	Bakshiks
Ball, Darroch	NZ First Party	List	darrochball
Barry, Maggie	National Party	North Shore	maggiebarrynz
Bayly, Andrew	National Party	Hunua	—
Bennett, David	National Party	Hamilton East	DavidBennettMP
Bennett, Paula	National Party	Upper Harbour	paulabennettmp
Bidois, Dan	National Party	Northcote	—
Bishop, Chris	National Party	Hutt South	cjsbishop
Bridges, Simon	National Party	Tauranga	simonjbridges
Brown, Simeon	National Party	Pakuranga	SimeonBrownMP
Brownlee, Gerry	National Party	Ilam	GerryBrownleeMP
Carter, David	National Party	List	DavidCarterMP
Clark, David	Labour Party	Dunedin North	DavidClarkNZ
Coffey, Tamati	Labour Party	Waiariki	tamaticoffey
Collins, Judith	National Party	Papakura	JudithCollinsMP
Craig, Liz	Labour Party	List	—
Curran, Clare	Labour Party	Dunedin South	ClareCurranMP
Davidson, Marama	Green Party	List	MaramaDavidson
Davis, Kelvin	Labour Party	Te Tai Tokerau	NgatiBird
Dean, Jacqui	National Party	Waitaki	—
Doocey, Matt	National Party	Waimakariri	—
Dowie, Sarah	National Party	Invercargill	nzsarahdowie
Dyson, Ruth	Labour Party	Port Hills	ruthdysonmp
Eagle, Paul	Labour Party	Rongotai	pauleaglenz
Faafoi, Kris	Labour Party	Mana	KrisinMana
Falloon, Andrew	National Party	Rangitata	andrewfalloon
Garcia, Paulo	National Party	List	—
Genter, Julie Anne	Green Party	List	JulieAnneGenter
Ghahraman, Golriz	Green Party	List	golrizghahraman
Goldsmith, Paul	National Party	List	PaulGoldsmithMP
Guy, Nathan	National Party	Ōtaki	NathanGuyOtaki
Hayes, Joanne	National Party	List	jo_hayes1
Henare, Peeni	Labour Party	Tāmaki Makaurau	PeeniHenare
Hipango, Harete	National Party	Whanganui	—
Hipkins, Chris	Labour Party	Rimutaka	chrishipkins
Hudson, Brett	National Party	List	bhudson_nz
Hughes, Gareth	Green Party	List	GarethMP
Huo, Raymond	Labour Party	List	RaymondHuo
Jackson, Willie	Labour Party	List	WillieJLabour
Jones, Shane	NZ First Party	List	—
Kanongata'a-Suisuiki, Anahila	Labour Party	List	AAnahila
Kaye, Nikki	National Party	Auckland Central	nikkikaye
King, Matt	National Party	Northland	MattKingMP
Kuriger, Barbara	National Party	Taranaki-King Country	BarbaraKuriger

Surname, Firstname	Party	Electorate	Username
Lee, Denise	National Party	Maungakiekie	DeniseLeeMP
Lee, Melissa	National Party	List	melissaleemp
Lees-Galloway, Iain	Labour Party	Palmerston North	IainLG
Little, Andrew	Labour Party	List	AndrewLittleMP
Logie, Jan	Green Party	List	janlogie
Loheni, Agnes	National Party	List	—
Lubeck, Marja	Labour Party	List	MarjaLubeck
Luxton, Jo	Labour Party	List	joluxx
Macindoe, Tim	National Party	Hamilton West	timmacindoe
Mahuta, Nanaia	Labour Party	Hauraki-Waikato	NanaiaMahuta
Mallard, Trevor	Labour Party	List	SpeakerTrevor
Marcroft, Jenny	NZ First Party	List	jennymarcroft
Mark, Ron	NZ First Party	List	RonMarkNZF
Martin, Tracey	NZ First Party	List	TraceyMartinMP
McAnulty, Kieran	Labour Party	List	Kieran_McAnulty
McClay, Todd	National Party	Rotorua	toddmcclaymp
McKelvie, Ian	National Party	Rangitīkei	ianmckelviemp
Mitchell, Clayton	NZ First Party	List	—
Mitchell, Mark	National Party	Rodney	MarkMitchellMP
Muller, Todd	National Party	Bay of Plenty	toddmullerBoP
Nash, Stuart	Labour Party	Napier	Stuart_NashMP
Ngaro, Alfred	National Party	List	AlfredNgaroMP
O’Connor, Damien	Labour Party	West Coast-Tasman	DamienOConnorMP
O’Connor, Greg	Labour Party	Ōhāriu	GregOhariu
O’Connor, Simon	National Party	Tāmaki	—
Parker, David	Labour Party	List	DavidParkerMP
Parmar, Parmjeet	National Party	List	Parmjeet_Parmar
Patterson, Mark	NZ First Party	List	markpattersonmp
Penk, Chris	National Party	Helensville	ChrisPenknz
Peters, Winston	NZ First Party	List	winstonpeters
Prime, Willow-Jean	Labour Party	List	WillowPrime
Pugh, Maureen	National Party	List	MaureenPughNat
Radhakrishnan, Priyanca	Labour Party	List	priyancanzlp
Reti, Shane	National Party	Whangarei	DrShaneRetiMP
Robertson, Grant	Labour Party	Wellington Central	grantrobertson1
Ross, Jami-Lee	Independent - not party affiliated	Botany	jamileeross
Rurawhe, Adrian	Labour Party	Te Tai Hauāuru	adrianrurawhe
Russell, Deborah	Labour Party	New Lynn	BeeFaerie
Sage, Eugenie	Green Party	List	EugenieSage
Salesa, Jenny	Labour Party	Manukau East	JennySalesa
Scott, Alastair	National Party	Wairarapa	ascottwairarapa
Sepuloni, Carmel	Labour Party	Kelston	carmelsepuloni
Seymour, David	ACT Party	Epsom	dbseymour
Shaw, James	Green Party	List	jamespeshaw
Simpson, Scott	National Party	Coromandel	ScottSimpsonMP
Sio, Aupito William	Labour Party	Māngere	AupitoWSio_MP
Smith, Nick	National Party	Nelson	—
Smith, Stuart	National Party	Kaikōura	stuartsmithmp
Stanford, Erica	National Party	East Coast Bays	—
Strange, Jamie	Labour Party	List	jamiestrangenz
Swarbrick, Chlöe	Green Party	List	_chloeswarbrick
Tabuteau, Fletcher	NZ First Party	List	FletcherNZFirst

Surname, Firstname	Party	Electorate	Username
Tinetti, Jan	Labour Party	List	jantinetti
Tirikatene, Rino	Labour Party	Te Tai Tonga	RinoTirikatene
Tolley, Anne	National Party	East Coast	AnneTolleyMP
Twyford, Phil	Labour Party	Te Atatū	PhilTwyford
Upston, Louise	National Party	Taupō	LouiseUpston
van de Molen, Tim	National Party	Waikato	timvandemolen
Wagner, Nicky	National Party	List	nickywagner
Walker, Hamish	National Party	Clutha-Southland	HamishWalkerMP
Wall, Louisa	Labour Party	Manurewa	—
Warren-Clark, Angie	Labour Party	List	angewarrenclark
Webb, Duncan	Labour Party	Christchurch Central	Duncan_Webb_
Whaitiri, Meka	Labour Party	Ikaroa-Rāwhiti	mekawhaitiri
Williams, Poto	Labour Party	Christchurch East	PotoChchEast
Willis, Nicola	National Party	List	NicolaWillisMP
Wood, Michael	Labour Party	Mt Roskill	michaelwoodnz
Woodhouse, Michael	National Party	List	WoodhouseMP
Woods, Megan	Labour Party	Wigram	Megan_Woods
Yang, Jian	National Party	List	—
Young, Jonathan	National Party	New Plymouth	JonathanYoungMP
Yule, Lawrence	National Party	Tukituki	LawrenceYuleMP

References

Agirrezabal, Manex, Inaki Alegria, and Mans Hulden. 2016. "Machine Learning for Metrical Analysis of English Poetry." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 772–81.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26 (10): 1531–42.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O'Reilly Media, Inc.".

Carey, John M, and Matthew Soberg Shugart. 1995. "Incentives to Cultivate a Personal Vote: A Rank Ordering of Electoral Formulas." *Electoral Studies* 14 (4): 417–39. https://doi.org/https://doi.org/10.1016/0261-3794(94)00035-2.

Crick, Bernard, and Alex Porter. 1978. *Political Education and Political Literacy : The Report and Papers of, and the Evidence Submitted to, the Working Party of the Hansard Society's 'Programme for Political Education'*. London: Longman.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. "How to Make Causal Inferences Using Texts." *arXiv Preprint arXiv:1802.02163*.

Firth, John R. 1957. "A Synopsis of Linguistic Theory, 1930-1955." *Studies in Linguistic Analysis*.

Grimmer, Justin. 2016. "Measuring Representational Style in the House: The Tea Party, Obama, and Legislators' Changing Expressed Priorities." In *Computational Social Science: Discovery and Prediction*, edited by R. Michael Alvarez, 225–45. Analytical Methods for Social Research. Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.010.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Heitshusen, Valerie, Garry Young, and David M Wood. 2005. "Electoral Context and MP Constituency Focus in Australia, Canada, Ireland, New Zealand, and the United Kingdom." *American Journal of Political Science* 49 (1): 32–45.

Honnibal, Matthew, and Ines Montani. 2017. "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing." *In Review.*

Jackson, Nigel, and Darren Lilleker. 2011. "Microblogging, Constituency Service and Impression Management: UK MPs and the Use of Twitter." *The Journal of Legislative Studies* 17 (1): 86–105.

Jungherr, Andreas. 2016. "Twitter Use in Election Campaigns: A Systematic Literature Review." *Journal of Information Technology & Politics* 13 (1): 72–91.

Kitschelt, Herbert. 2000. "Linkages Between Citizens and Politicians in Democratic Polities." *Comparative Political Studies* 33 (6-7): 845–79.

Lauderdale, Benjamin E, and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (3): 374–94.

Le, Quoc V., and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *CoRR* abs/1405.4053. http://arxiv.org/abs/1405.4053.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–77.

Lupu, Noam. 2013. "Party Brands and Partisanship: Theory with Evidence from a Survey Experiment in Argentina." *American Journal of Political Science* 57 (1): 49–64.

Lynch, Marc. 2011. "After Egypt: The Limits and Promise of Online Challenges to the Authoritarian Arab State." *Perspectives on Politics* 9 (2): 301–10.

Martin, Shane. 2011. "Using Parliamentary Questions to Measure Constituency Focus: An Application to the Irish Case." *Political Studies* 59 (2): 472–88.

McKinney, Wes, and others. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, 445:51–56. Austin, TX.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *CoRR* abs/1301.3781. http://arxiv.org/abs/1301.3781.

Morozov, Evgeny. 2011. "Whither Internet Control?" *Journal of Democracy* 22 (April): 62–74. https://doi.org/10.1353/jod.2011.0022.

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: 2825–30.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26 (January): 120–28. https://doi.org/10.1017/pan.2017.39.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Computational Social Science: Discovery and Prediction*, edited by R. Michael Alvarez, 51–97. Analytical Methods for Social Research. Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.004.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014a. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Roberts, Margaret E, Brandon M Stewart, Edoardo M Airoldi, K Benoit, D Blei, P Brandt, and A Spirling. 2014b. "Structural Topic Models." *Retrieved May* 30: 2014.

Rodman, Emma. 2019. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis*, 1–25.

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Searing, Donald. 1994. *Westminster's World: Understanding Political Roles.* Harvard University Press.

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

Stewart, Brandon M., and Yuri M. Zhukov. 2009. "Use of Force and Civil–Military Relations in Russia: An Automated Content Analysis." *Small Wars & Insurgencies* 20 (2): 319–43. https://doi.org/10.1080/09592310902975455.

Strøm, Kaare. 1997. "Rules, Reasons and Routines: Legislative Roles in Parliamentary Democracies." *The Journal of Legislative Studies* 3 (1): 155–74.

Tucker, Joshua A., Jonathan Nagler, Megan Macduffee Metzger, Pablo Barberá, Duncan Penfold-Brown, and Richard Bonneau. 2016. "Big Data, Social Media, and Protest: Foundations for a Research Agenda." In *Computational Social Science: Discovery and Prediction*, edited by R. Michael Alvarez, 199–224. Analytical Methods for Social Research. Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.009.

Wilson, Christopher, and Alexandra Dunn. 2011. "Digital Media in the Egyptian Revolution: Descriptive Analysis from the Tahrir Data Sets." *International Journal of Communication* 5 (January): 1248–72.

Wintner, Shuly. 2010. "Formal Language Theory." In *The Handbook of Computational Linguistics and Natural Language Processing*, 11–42. Wiley Online Library.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15 (2): 265–86.