Nobody expects the Spanish Inquisition: Which events drive topic change in Australian parliaments?

Monica Alexander[*]
Rohan Alexander[†]

12 September 2018

Abstract

We use a structural text model to explore the impact of various events on what was said in Australian state and Commonwealth parliaments from the mid-1800s through to 2017. We find that 1) changes of government where the party in power also changes, result in significant changes in topics discussed in parliament, but when the party does not change the topics are unchanged; 2) polling results do not have a significant effect; 3) economic changes, such as financial crises have a significant effect; and 4) other events, such as an unexpected attack tend not to have a prolonged change.

1 Introduction

New governments often go out of their way to be different to the governments they replace. For instance, consider Kevin Rudd's apology to Indigenous Australians which supported by all living Prime Ministers apart from his immediate predecessor. Similarly, significant events can alter the course of a government. For instance, consider the change in the Howard government after the 9/11 attacks. However it is not so clear which events drive changes in topics, for instance, do they change when the government is replaced by another of its own party? And which events are temporary, for instance when an economic crisis abates, do the topics return to pre-crisis levels?

In this paper we use the Structural Topic Model (STM) of Roberts, Stewart, and Airoldi (2016) to model the topics of discussion in Australian parliaments. The advantage of this model is that it allows for topics to be correlated between sitting days, which then allows us to test for changes in topics at various events. The events that we focus on are changes in: 1) government; 2) the political environment (as defined by polling or other results); 3) economic conditions; and 4) the significant events (such as the 9/11 attacks or the Bali Bombings).

We find [INSERT RESULTS]. We also explored the other direction (the impact of what was said in parliaments on events) but it was difficult to find significant effects.

[*]University of Toronto, Statistical Sciences and Sociology Departments, `monica.alexander@utoronto.ca`.
[†]Australian National University, Research School of Economics, `rohan.alexander@anu.edu.au`.

Our work fits into [WHATEVER IT FITS INTO]. While using text as data has well-known shortcomings, it allows larger-scale analysis that would not be viable using less-automated approaches and hence can identify patterns that may otherwise be overlooked.

2 Data

Following the example of the UK, a text-based record of what was said in Australian parliaments called Hansard has been regularly made available since their establishment. While Hansard is not necessarily verbatim, it is considered close enough for text-as-data purposes. For instance, Mollin (2008) found that in the case of the UK Hansard the differences would only affect specialised linguistic analysis. Edwards (2016) examined Australia, New Zealand and the UK, and found that changes were usually made by those responsible for creating the Hansard record, instead of the parliamentarians. Both these findings provide reassurance that differences between Hansard and a verbatim record would not be meaningful for this paper.

The parliaments generally make their hansard records available online in PDF formats that can be processed by text readers such as Ooms (2018a). Some records such as the early days of the New South Wales parliament have yet to be put through a professional OCR process, although the Google Tesseract engine as implemented by Ooms (2018b) can be used to provide a useful first-pass if necessary. Hansard records that are still images are not used in this analysis (see the Appendix for more details).

Hansard records are available for all six Australian states and the Commonwealth although the period varies (Table 1). Here we focus on the lower houses although a similar analysis is possible in a bicameral setting.

Parliament	House	Years used	Notes
Commonwealth	House of Representatives	1901 - 2017	-
Queensland	?	1861 - 2017	-
New South Wales	?	? - 2017	-
Victoria	?	? - 2017	-
Tasmania	?	? - 2017	-
South Australia	?	? - 2017	-
Western Australia	?	? - 2017	-

We start with the PDFs that have been through an OCR process (or compiled more recently using software that makes them readable by default) as supplied by various parliaments (see the Appendix for specific sources). We then use the `PDFtools` R package of Ooms (2018a) to extract these as text-based CSV records, and clean the records using functions from the `Tidyverse` R package of Wickham (2017), the `Tidytext` R package of Silge and Robinson (2016) and the [INSERT OTHERS]. The result is a

3 Model

The primary model that we use in this paper is the Structural Topic Model (STM) as implemented by the `STM` R package of Roberts, Stewart, and Tingley (2018). The basis of this type of topic modelling is the Latent Dirichlet Allocation (LDA) model of Blei, Ng, and Jordan (2003). In this section a brief overview of both the LDA model and the STM approach is provided and then the specifics of how we consider events in this setting are discussed.

3.1 Latent Dirichlet Allocation

Each day's Hansard record needs to be classified by its topic. Sometimes Hansard includes titles that make the topic clear. But not every statement has a title and the titles do not always define topics in a well-defined and consistent way, especially over longer time periods. One way to get consistent estimates of the topics of each statement in Hansard is to use the latent Dirichlet allocation (LDA) method of Blei, Ng, and Jordan (2003), for instance as implemented by the R package 'topicmodels' by Grün and Hornik (2011).

The key assumption behind the LDA method is that each day's text, 'a document', in Hansard is made by speakers who decide the topics they would like to talk about in that document, and then chooses words, 'terms', that are appropriate to those topics. A topic could be thought of as a collection of terms, and a document as a collection of topics. The topics are not specified *ex ante*; they are an outcome of the method. Terms are not necessarily unique to a particular topic, and a document could be about more than one topic. This provides more flexibility than other approaches such as a strict word count method. The goal is to have the words found in each day's Hansard group themselves to define topics.

As applied to Hansard, the LDA method considers each statement to be a result of a process where a politician first chooses the topics they want to speak about. After choosing the topics, the speaker then chooses appropriate words to use for each of those topics.

More generally, the LDA topic model works by considering each document as having been generated by some probability distribution over topics. For instance, if there were five topics and two documents, then the first document may be comprised mostly of the first few topics; the other document may be mostly about the final few topics (Figure @ref(fig:topicsoverdocuments)).

Similarly, each topic could be considered a probability distribution over terms. To choose the terms used in each document the speaker picks terms from each topic in the appropriate proportion. For instance, if there were ten terms, then one topic could be defined by giving more weight to terms related to immigration; and some other topic may give more weight to terms related to the economy (Figure @ref(fig:topicsoverterms)).

Following Blei and Lafferty (2009), Blei (2012) and Griffiths and Steyvers (2004), the process by which a document is generated is more formally considered to be:

1. There are $1, 2, \dots, k, \dots, K$ topics and the vocabulary consists of $1, 2, \dots, V$ terms. For

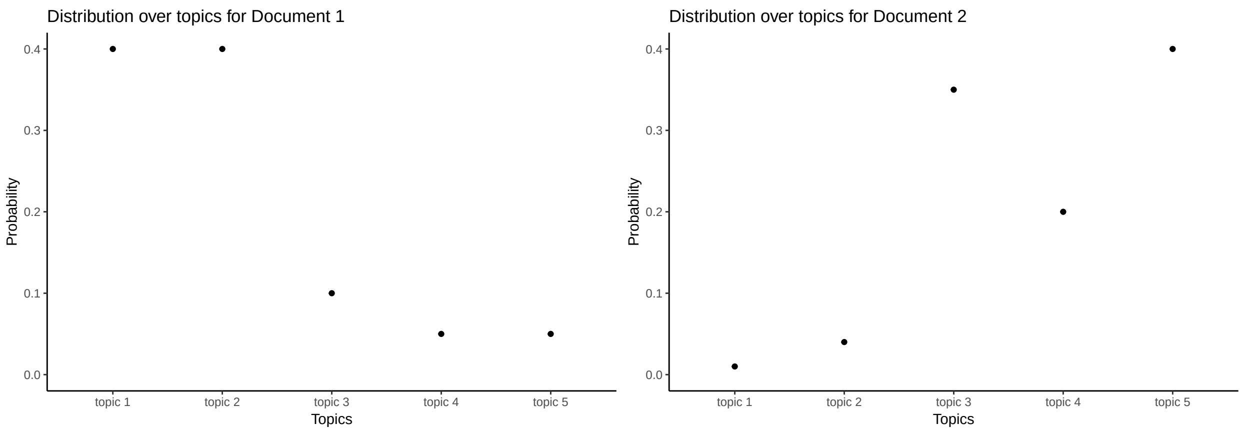

Figure 1: Probability distributions over topics

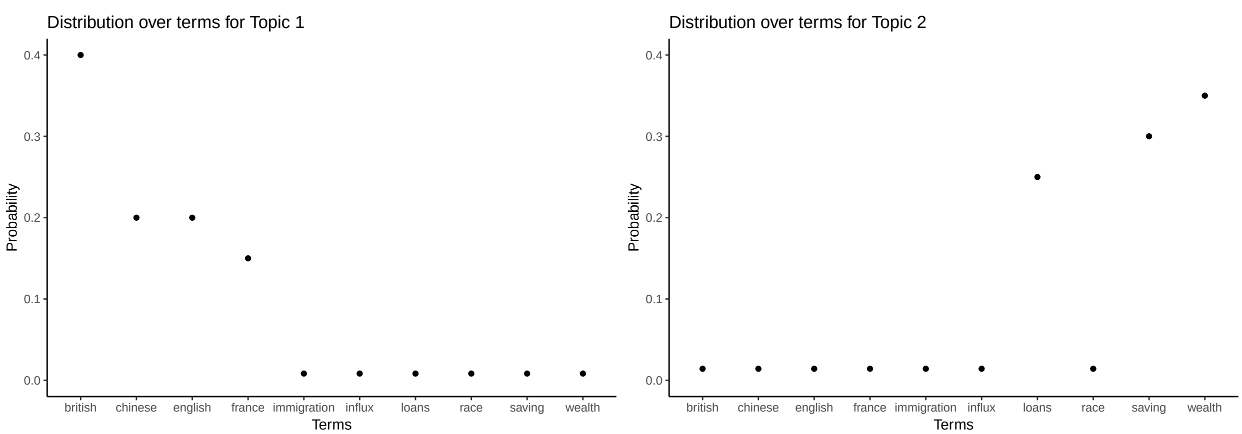

Figure 2: Probability distributions over terms

each topic, decide the terms that the topic uses by randomly drawing distributions over the terms. The distribution over the terms for the kth topic is β_k. Typically a topic would be a small number of terms and so the Dirichlet distribution with hyperparameter $0 < \eta < 1$ is used: $\beta_k \sim \text{Dirichlet}(\eta)$.[1] Strictly, η is actually a vector of hyperparameters, one for each K, but in practice they all tend to be the same value.

2. Decide the topics that each document will cover by randomly drawing distributions over the K topics for each of the $1, 2, \dots, d, \dots, D$ documents. The topic distributions for the dth document are θ_d, and $\theta_{d,k}$ is the topic distribution for topic k in document d. Again, the Dirichlet distribution with the hyperparameter $0 < \alpha < 1$ is used here because usually a document would only cover a handful of topics: $\theta_d \sim \text{Dirichlet}(\alpha)$. Again, strictly α is vector of length K of hyperparameters, but in practice each is usually the same value.
3. If there are $1, 2, \dots, n, \dots, N$ terms in the dth document, then to choose the nth term, $w_{d,n}$:
   a. Randomly choose a topic for that term n, in that document d, $z_{d,n}$, from the multinomial distribution over topics in that document, $z_{d,n} \sim \text{Multinomial}(\theta_d)$.
   b. Randomly choose a term from the relevant multinomial distribution over the terms for that topic, $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$.

Given this set-up, the joint distribution for the variables is (Blei (2012), p.6):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p\left(w_{d,n}|\beta_{1:K}, z_{d,n}\right) \right).$$

Based on this document generation process the analysis problem, discussed next, is to compute a posterior over $\beta_{1:K}$ and $\theta_{1:D}$, given $w_{1:D,1:N}$. This is intractable directly, but can be approximated (Griffiths and Steyvers (2004) and Blei (2012)).

After the documents are created, they are all that we have to analyse. The term usage in each document, $w_{1:D,1:N}$, is observed, but the topics are hidden, or 'latent'. We do not know the topics of each document, nor how terms defined the topics. That is, we do not know the probability distributions of Figures @ref(fig:topicsoverdocuments) or @ref(fig:topicsoverterms). In a sense we are trying to reverse the document generation process – we have the terms and we would like to discover the topics.

If the earlier process around how the documents were generated is assumed and we observe the terms in each document, then we can obtain estimates of the topics (Steyvers and Griffiths (2006)). The outcomes of the LDA process are probability distributions and these define the topics. Each term will be given a probability of being a member of a particular topic, and each document will be given a probability of being about a particular topic. That is, we are trying to calculate the posterior distribution of the topics given the terms observed in each

[1] The Dirichlet distribution is a variation of the beta distribution that is commonly used as a prior for categorical and multinomial variables. If there are just two categories, then the Dirichlet and the beta distributions are the same. In the special case of a symmetric Dirichlet distribution, $\eta = 1$, it is equivalent to a uniform distribution. If $\eta < 1$, then the distribution is sparse and concentrated on a smaller number of the values, and this number decreases as η decreases. A hyperparameter is a parameter of a prior distribution.

document (Blei (2012), p. 7):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N}) = \frac{p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}\right)}{p(w_{1:D,1:N})}.$$

The initial practical step when implementing LDA given a collection of documents is to remove 'stop words'. These are words that are common, but that don't typically help to define topics. There is a general list of stop words such as: "a"; "a's"; "able"; "about"; "above"... An additional list of words that are commonly found in Hansard, but likely don't help define topics is added to the general list. These additions include words such as: "act"; "amendment"; "amount"; "australia"; "australian"; "bill"... A full list can be found in Appendix @ref(hansard-stop-word). We also remove punctuation and capitalisation. The documents need to then be transformed into a document-term-matrix. This is essentially a table with a column of the number of times each term appears in each document.

After the dataset is ready, the R package 'topicmodels' by Grün and Hornik (2011) can be used to implement LDA and approximate the posterior. It does this using Gibbs sampling or the variational expectation-maximization algorithm. Following Steyvers and Griffiths (2006) and Darling (2011), the Gibbs sampling process attempts to find a topic for a particular term in a particular document, given the topics of all other terms for all other documents. Broadly, it does this by first assigning every term in every document to a random topic, specified by Dirichlet priors with $\alpha = \frac{50}{K}$ and $\eta = 0.1$ (Steyvers and Griffiths (2006) recommends $\eta = 0.01$), where α refers to the distribution over topics and η refers to the distribution over terms (Grün and Hornik (2011), p. 7). It then selects a particular term in a particular document and assigns it to a new topic based on the conditional distribution where the topics for all other terms in all documents are taken as given (Grün and Hornik (2011), p. 6):

$$p(z_{d,n} = k|w_{1:D,1:N}, z'_{d,n}) \propto \frac{\lambda'_{n\to k} + \eta}{\lambda'_{.\to k} + V\eta} \frac{\lambda'^{(d)}_{n\to k} + \alpha}{\lambda'^{(d)}_{-i} + K\alpha}$$

where $z'_{d,n}$ refers to all other topic assignments; $\lambda'_{n\to k}$ is a count of how many other times that term has been assigned to topic k; $\lambda'_{.\to k}$ is a count of how many other times that any term has been assigned to topic k; $\lambda'^{(d)}_{n\to k}$ is a count of how many other times that term has been assigned to topic k in that particular document; and $\lambda'^{(d)}_{-i}$ is a count of how many other times that term has been assigned in that document. Once $z_{d,n}$ has been estimated, then estimates for the distribution of words into topics and topics into documents can be backed out.

This conditional distribution assigns topics depending on how often a term has been assigned to that topic previously, and how common the topic is in that document (Steyvers and Griffiths (2006)). The initial random allocation of topics means that the results of early passes through the corpus of document are poor, but given enough time the algorithm converges to an appropriate estimate.

The choice of the number of topics, k, affects the results, and must be specified *a priori*. If there is a strong reason for a particular number, then this can be used. Otherwise, one way to choose an appropriate number is to use a test and training set process. Essentially, this means running the process on a variety of possible values for k and then picking an appropriate value that performs well.

One weakness of the LDA method is that it considers a 'bag of words' where the order of those words does not matter (Blei (2012)). It is possible to extend the model to reduce the impact of the bag-of-words assumption and add conditionality to word order. Additionally, alternatives to the Dirichlet distribution can be used to extend the model to allow for correlation. For instance, in Hansard topics related the army may be expected to be more commonly found with topics related to the navy, but less commonly with topics related to banking. This motivates the use of the Structural Topic Model, described in the next section.

3.2 Structural Topic Model

3.2.1 Overview and example

[TBD]

Note that each of the states and the Commonwealth are treated independently here. Future work could expand the model to better understand, and allow, for correlation between them.

3.2.2 Considering events

[TBD]

4 Results

4.1 Political events

When you change the government, you change the country. Paul Keating.

Change of government.

4.2 Polling events

The only poll that matters is the one on election day. John Howard.

[TBD]

4.3 Economic events

Major economic changes.

[TBD]

4.4 External events

Events, dear boy, events. Attributed to Harold Macmillan.

Major event such as 9/11 attacks, or economic change.

5 Summary and conclusions

What could happen if we had longer terms. Eg GST needed multiple generations of politicians but carbon tax couldn't because it was one generation.

Text analysis has well-known biases and weaknesses and is a complement to more detailed analysis such as qualitative methods and case studies. We consider the results presented in this paper, as well as many of those results of the larger text-as-data research program, as fitting within findings based on other methods.

6 Appendix

6.1 Selection of number of topics

6.2 Robustness of results

Here we change the number of sitting days considered either side of an event. The results in the main section of the paper are for the nearest ten days either side of an event. Here are show that the results are essentially the same if the nearest one, two, five, and twenty days either side of an event.

6.3 Document sources

Where from?

Which years are being used (not non-OCRd)

References

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.

Blei, David M, and John D Lafferty. 2009. "Topic Models." In *Text Mining*, 101–24. Chapman; Hall/CRC.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Darling, William M. 2011. "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling." In.

Edwards, Cecilia. 2016. "The Political Consequences of Hansard Editorial Policies: The Case for Greater Transparency." *Australasian Parliamentary Review* 31 (2): 145–60.

Griffiths, Thomas, and Mark Steyvers. 2004. "Finding Scientific Topics." *PNAS* 101: 5228–35.

Grün, Bettina, and Kurt Hornik. 2011. "Topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software, Articles* 40 (13): 1–30. doi:10.18637/jss.v040.i13.

Mollin, Sandra. 2008. "The Hansard Hazard: Gauging the Accuracy of British Parliamentary Transcripts." *Corpora* 2 (2): 187–210.

Ooms, Jeroen. 2018a. *Pdftools: Text Extraction, Rendering and Converting of Pdf Documents*. https://CRAN.R-project.org/package=pdftools.

———. 2018b. *Tesseract: Open Source Ocr Engine*. https://CRAN.R-project.org/package=tesseract.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2018. *Stm: R Package for Structural Topic Models*. http://www.structuraltopicmodel.com.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1 (3). The Open Journal. doi:10.21105/joss.00037.

Steyvers, Mark, and Tom Griffiths. 2006. "Probabilistic Topic Models." In *Latent Semantic Analysis: A Road to Meaning*, edited by T. Landauer, D McNamara, S. Dennis, and W. Kintsch.

Wickham, Hadley. 2017. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.