QCA

15 May 2017

Data and Methodology

Data

The data for this en devour is sourced from multiple sources. The primary dataset to analyse the differences in party ideology comes from the Comparative Manifesto Project. The project addresses the collection and the comparative content analysis of parties' manifestos, covering over 1000 parties from 1945 onward in over 50 countries on five continents. The dataset is limited to the major political parties in Germany for our analysis. Furthermore, due to the lack of historical archives to media articles, as well as issues with storing and analysing the huge corpora of text, the data is further limited to elections post 2000 for comparison with trends in media.

Dataset for media is collected from scraping Der Spiegel, a daily left-leaning German news publisher. Der Speigel is unique in their free access to their archives as well as their pre-categorization of the news articles. All articles on Der Spiegel are categorized into topics such as politics, sports, culture . . . , with the topic of politics being further categorized into domestic and international political news. The pre-processing of data into broad categories, allows us to focus on collecting only the relevant dataset towards our analysis, without the risk of bias. Thus, we scrape all article which relate to domestic politics from 2000 onward.

The dataset on news is further limited to only articles which were published 1 year before each election. Thus for the 2002 elections, articles published in between the periods September 2001 and 2002 were scraped, and so forth. This is led by the assumption of *limited memory*; voters and individuals are constrained by their limited memory when making decisions. Studies have noted the consumer choice to be bounded by small term memory and processing skills [@James1986, @Dick2017]. In a similar vein, voters are limited to recent memory when making a choice. Similarly, individuals do not focus on historical events, but on the events that occur closer to elections. Thus it would be imperative to assume that the events that are closer to elections, and the media coverage of these events, should drive the narrative of political parties as well. It is however important to note that the use of 1 year before each election is an assumption and may be violated. There are no studies which look at the limits of long-term memory when it concerns towards political choice, and the use of different data intervals are left to future research.

The website is scraped using Scrapy, a web scraping framework written in Python. The dataset for news articles post 2000 onward is encompasses nearly 70000 observations, though they are limited to nearly 14000 for our analysis. The data as well as the scraping codes are available on Github.

Methodology

The analysis can be broken into two steps; the first step will look at analysing the differences withing the priorities for each major party in Germany. We will use the CMP dataset to work on the differences between the parties using both a measure of right-left slant of the topic, as well as the importance of that topic for the party, measured using the distribution of the different topics.

The next step would be to compare the individual distribution for the parties with the distributions derived from news articles. High correlations between the two would imply a convergence between the different parties on major topics, while low correlations would imply a dichotomy between the issues of the hour and the response to the parties to those issues.

A brief divergence must be made here to defend the methods followed by the paper. We use a topic model to differentiate the different topics in news across the different years, but we do not provide them with a left

or a right leaning political score. There are two reasons for this approach; Firstly there is reason to believe that the scores might be biased for our dataset. As explained earlier, Der Spiegel is a left leaning daily news publisher in Germany. As such it is not difficult to believe that the scores may be biased against right and conservative parties in Germany.

Secondly, we believe that it is a difficult task to compare the political scores between media and political parties to explain the convergence or dichotomy between the parties and the general public. The act of comparison presumes a standard to which one or the other party must reach, in our case we presume that the media is a *vox populi*, and political parties will be driven towards addressing those voices. However by using political scores such as rile, we further assume the slant, and opinions, of the media to be a gold standard which other parties must concur towards. This is a dangerous notion, both for democracy and free speech. It is a better standard to allow the different parties to differ in their response to an issue, as long as the political parties follow the urgency of a particular issue. Thus by comparing the topic distributions between parties and media, we focus only on the *importance* accorded to each topic by the political parties, allowing them to differentiate the responses according to their political leanings.

Topic Modelling News Articles

The nearly 14000 news articles were then categorised using unsupervised learning methods. In particular, we used the Latent Dirchlett Allocation to determine the topic probabilities and the term-topic distribution. Formally the model is defined as follows;

Documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for a corpus D consisting of M documents each of length N_i. The topic distribution for each documnet θ_i is chosen such that $\theta_i \sim Dir(\alpha)$ where $i \in 1, ..., M$ and $Dir(\alpha)$ is a Dirchlet distribution with the hyperparameter α measuring the *spread* or sparsity of topics.

φ_k, the word distribution for topic k, is then chosen such that $\varphi_k \sim Dir(\beta)$ for $k \in 1, ..., K$. Finally for each of the word positions i, j, where $j \in 1, \ldots, N_i$, and $i \in 1, \ldots, M$, a topic $z_{i,j}$ and word $w_{i,j}$ is chosen such that $z_{i,j} \sim Multinomial(\theta_i)$ and $w_{i,j} \sim Multinomial(\varphi_{z_{i,j}})$.

We use the simple LDA framework, noted above, sestablish by @Blei2013 without adding a time parameter to allow for dynamic topic modelling. This *muddles the water* a bit, as LDA does not diffrentiate between the timings of the different topics, and thus may bias the result. Nonetheless, the use of dynamic topics are left for future research. Finally, we set the hyperparameters α as 0.8 and K as 100. 100 topics should cover a breath of the articles, though we are left in the dark regarding the correct distribution.

Finally, before the text is categorised, all of the articles are cleaned to remove common stop words and punctuation. The words are further *lemmatised* using the Treetagger algorithims derived by [**xxxx**]. Lemmatisation of text allows us to control for tenses within the text, as well as reduce the feature space. A brief demonstration of the lemmatisation is given below.

Before Lemmatisation;

> Üblicherweise tagt der 6. Strafsenat des Oberlandesgerichts Düsseldorf in einem Hochsicherheitsbunker hinter Beton Panzerglas und direkt neben dem Landeskriminalamt. Vor der Tür stehen an den Verhandlungstagen bewaffnete Polizisten mit Maschinenpistolen im Anschlag. Demnächst

After Lemmatisation;

> üblicherweise tagen Strafsenat Oberlandesgericht Düsseldorf Hochsicherheitsbunker Beton Panzerglas direkt neben Landeskriminalamt vor Türe stehen Verhandlungstag bewaffnet Polizist Maschinenpistole Anschlag demnächst

As seen, the lemmatised text returns all words to their root form, as well as removing digits and punctuation. All of these help in reducing feature space to categorise using LDA.

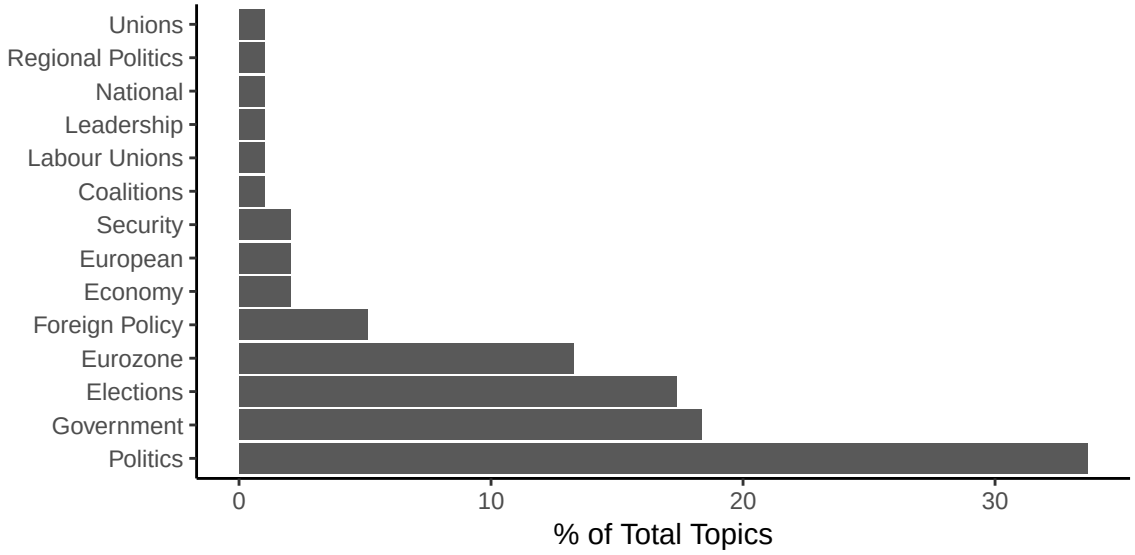
Results

Inter-Party Differences

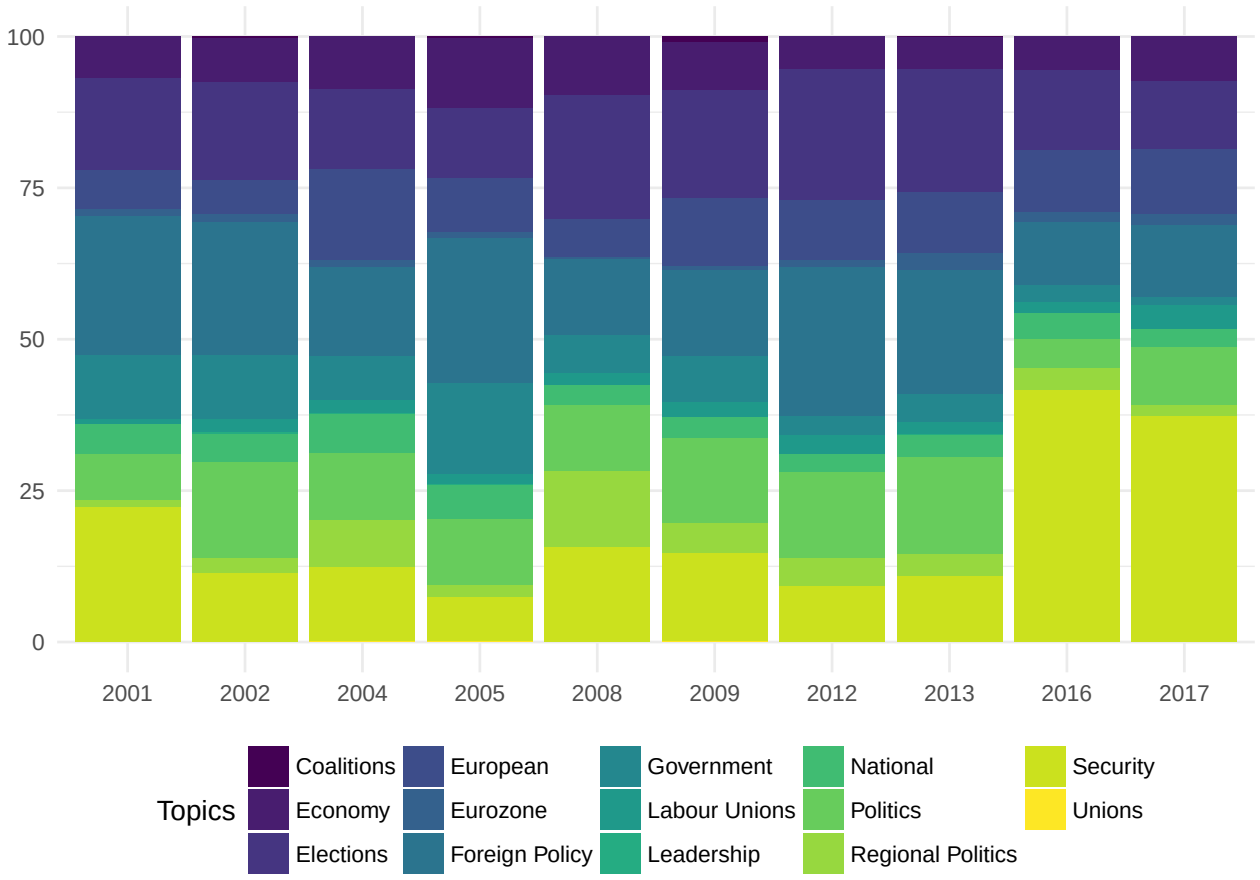
Party and Media

Before we move on to correlate the topic and party distributions, we first look at the breakdowns of the different topics. The 100 topics were labeled, and often spoke of similar topics, though we tried to be as specific as possible with regards to labelling of the different topics. A breakdown of the different topics are given below.

As evident, the topics are unfortunately too broad to compare with political party manifestos. The papers often describe Eurozone policies or security concerns which may be picked up by the different political parties. However, the bulk of the topics tend towards Politics (Wether a party is going through turmoil), Elections (which party can expect how much votes) or Government (Coalitions, fall of coalitions etc.). The topics of elections are particularly interesting, since in may cases they occur months before the actual date of election.



Nearly 30% of our topics talk of general politics, invalidating a bulk of our data. Nonetheless, it would be interesting to view the changes within these topics over time. The figure below plots out the changes in the distribution of these topics over time.



Thus while our topic models worked, the end result is too broad to be able to compare with the policies which parties should speak off.

Conclusion and Further Work

References