Multidimensional Topic Analysis in Political Texts

Cäcilia Zirn and Heiner Stuckenschmidt

Research Group Data and Web Science
University of Mannheim
B6 26
Germany
{heiner,caecilia}@informatik.uni-mannheim.de

Abstract

Automatic content analysis is more and more becoming an accepted research method in social science. In political science researchers are using party manifestos and transcripts of political speeches to analyze the positions of different actors. Existing approaches are limited to a single dimension, in particular, they cannot distinguish between the positions with respect to a specific topic. In this paper, we propose a method for analyzing and comparing documents according to a set of predefined topics that is based on an extension of Latent Dirichlet Allocation (LDA) for inducing knowledge about relevant topics. We validate the method by showing that it can guess which member of a coalition was assigned a certain ministry based on a comparison of the parties' election manifestos with the coalition contract. We apply the method to German National Elections since 1990 and show that the use of our method consistently outperforms a baseline method that simulates manual annotation of individual sentences based on keywords and standard text comparison. In our experiments, we compare two different extensions of LDA and investigate the influence of the used seed set. Finally, we give a brief illustration of how the output of our method can be interpreted to compare positions towards specific topics across several parties.

Keywords: Topic Models, Political Science

1. Motivation

Data analysis has a longstanding tradition in social science as a main driver of empirical research. Traditionally, research has focused on survey

data as a main foundation. Recently, automatic text analysis has been discovered as a promising alternative to traditional survey based analysis, especially in the political sciences [1], where policy positions that have been identified automatically based on text can for example be used as input for simulations of party competition behavior [2]. The approach to text analysis adopted by researchers in this area is still strongly influenced by statistical methods used to interpret survey data [3]. A typical application is to place parties on a left-right scale based on the content of their party manifestos [4]. While it has been shown that existing methods can be very useful for analyzing and comparing party positions over time, existing methods are limited to a single dimension, typically the left-right scale. This means that positions of a party on various topics are reduced to a single number indicating an overall party position independent of a specific policy area. In this paper, we argue that there is a need for new analysis methods that are able to discriminate between positions on different policy areas and treat them independently. We propose a new approach on multidimensional analysis of party positions with respect to different policy areas. Often, we are interested in the position of a party *with respect to a certain topic* rather than an overall position. Existing methods are only able to answer questions of that kind if the input are texts talking exclusively about the topic under consideration (e.g. [5]). In contrast, there is a good reason why party manifestos have been the primary subject of attempts to identify party positions [6], as they are independent of personal opinions and opportunistic statements that influence for instance political speeches. This means that on the one hand manifestos are an important reference point for various comparisons and party position analyses, but on the other hand are hard to analyze with existing approaches as they cover a large variety of topics and the respective party's position towards this topic. We conclude that there is a need for methods that allow for position analysis based on multi-topic documents that takes these different topics into account.

In this paper, we address the problems of current one-dimensional analyses of political positions by proposing a content analysis method based on topic models that identifies topics put forward by parties in connection with a certain policy area. The general idea is the following: To compare two documents containing several topics, we first extract the topics automatically by running a topic model on each of the documents. Then, the positions towards the topics can be analyzed by measuring the distance between the corresponding topics.

We use a variant of topic models that allows the inclusion of seed words for characterizing the respective policy areas. This approach has a number of advantages over conventional topic models where topics are solely formed based on the analysis of a corpus. For standard topic models, the construction of topics can only be influenced by specifying the number of the expected topics within the corpus and some assumptions about their distributions. However, it is not possible to influence the thematic focus of the topics. As a result, it is neither possible to analyze a set of previously specified topics, nor is it possible to directly compare topics that were created from two distinct corpora, as it cannot be inferred directly which topic corresponds to another.

As it seems to be a problem to compare the output of two separate topic models, one might wonder why we do not run one single topic model on all the documents that are to be compared. In this case, however, the different positions the documents take towards various issues cannot be distinguished, as they end up within the very same topic.

Based on these requirements, we suggest the usage of existing variants of topic models for our approach, LogicLDA and Labeled LDA:

- Each of those variants allows to define certain policy areas that the topics in the model are supposed to represent.
- This in turn makes it possible to compare party interests in a certain policy area defined by a set of seed words.
- The use of seed words provides the flexibility to adapt analyzed areas to the given question, e.g. policy areas that are of interest in a regional election will not necessarily be of interest in the context of a federal election and vice versa.

The positions towards a policy area can be analyzed by comparing the distance of the corresponding topics that were the result of the topic models run on the documents.

We test the capability of our approach in two different scenarios. In the first experiment described in section 3, we show that the method can be used to predict the distribution of ministries between the parties of a winning coalition based on the distance of the positions extracted from their manifestos to the positions in the coalition agreement. We explain the rationale of this experiment in more detail later on. We also show that although of course the result of the analysis depends on the choice of the seed words,

the general principle works independently from a specific set of keywords. We compare the method to a baseline that simulated a manual approach to the problem where individual sentences are assigned to a topic based on keywords and sentences assigned to the same topic are compared. We show that our method consistently outperforms this baseline with respect to the task of predicting the assignment of the ministry. We will further investigate the impact of specific Latent Dirichlet Allocation (LDA) extensions, the seed set and the words included in the analysis.

The paper is organized as follows. In section 2 we present our multidimensional content analysis method that uses two alternative extensions of LDA for generating a topic model according to a predefined set of policy areas. Section 3 describes the experiments we conducted to validate the method by describing the rational of the experiment as well as the data sources used and the experimental setting. An example how the methods could be actually applied by political scientists to analyze party positions is described in section 4. We conclude with a discussion of the results and the implications for computer-aided content analysis in the social sciences.

2. Multi-dimensional Analysis

The goal of our work is the creation of a method for analyzing the positions a certain document takes towards various topical areas and comparing them to those of other documents. The method follows a number of assumptions that have to be explicated before discussing the method itself. First of all, we assume that there is a well defined set of topic (or policy) areas and that the document(s) to be analyzed actually contain(s) information related to these topic areas. The second fundamental assumption is that topic areas and specific positions can be described in terms of words associated with the respective topical area. This does not only allow us to characterize a topical area in terms of a number of seed words, it also justifies the use of topic models as an adequate statistical tool for carrying out the analysis. Finally, we assume that the distance between topic descriptions in terms of distributions over words is an indicator for the actual distance between the positions of the authors of the documents analyzed, in our case the parties stating their political program. Based on these assumptions, we have designed the following method for analyzing (political) positions based on documents such as party manifestos.

2.1. Data Preparation

Data preparation is an important step for any content analysis as the quality of the raw data has high influence on the quality of the analysis. For our method, we need to carry out two basic preprocessing steps: the first one is the creation of the corpus to be analyzed, the second one is to determine the vocabulary that should be the basis for the creation and the comparison of the topics.

Text Tiling. Topic models rely on co-occurrence statistics of words within a corpus consisting of multiple documents, each covering an arbitrary mixture of topics. As we are interested in analyzing single document[1] rather than a whole collection, e.g. a party manifesto, the data preparation step has to generate a corpus of documents with meaningful co-occurrences. As a solution to create appropriate input, we split this single document into several parts, which are considered as separate documents. While this can of course be done manually by reading the document and dividing it in a thematically coherent way, we aim at automating the analysis as far as possible to be able to carry out large scale analyses with limited manpower. Please note that the documents analyzed by topic models are allowed to cover various topics and are not limited to a single one. In our approach, we use TextTiling [7], which is a popular method for automatically cutting texts into topically coherent subparts using lexical cohesion as a main criteria. TextTiling determines thematic blocks in a document in three steps. First, the document is segmented into individual tokens (roughly words) that can be compared. Further, the method splits the document into sequences of tokens with equal length called token sentences. In the second step the Cosine similarity between adjacent token-sentences is determined and plotted into a graph. In the final step, thematic boundaries between token sentences are determined based on changes in the similarity. We chose this segmentation method because of its underlying assumption that text segments always contain a number of parallel information threads ([7], end of page 3). This is very close to the underlying assumption of Latent Dirichlet Allocation, that a single document always addresses a number of different topics to a certain extent which is given by the Dirichlet distribution. A positive side effect of the TextTiling method is that it is domain independent and does not require external parameters to be set.

[1]The approach can as well be applied to several documents sharing the same positions

Part-of-speech filtering. Another decision that has to be made when preparing the data is which types of words should be taken into account when building the statistical model. Of course, all words occurring in a document can in principle be used, however, this often leads to rather meaningless topics that contain a lot of words that do not actually carry a meaning. A rather natural restriction is to only use words of a certain type. For this purpose, we determine word types in our documents using a state of the art part-of-speech tagger [8] and filter the documents based on word types. For the purpose of our experiments it turned out that using nouns only works best, as they are best suited to describe a topic. For some questions it might also be useful to include adjectives to identify how certain words are perceived by the respective party (e.g. 'unfair' vs. 'effective' tax system) or verbs to get an idea of planned actions ('raise' vs. 'lower' taxes). Regardless of the chosen word types it can make sense to exclude infrequent words or stop words from the analysis. Stop words are function words that appear with high frequency in all kinds of text and are therefore useless for content analysis. As we restrict our vocabulary to nouns only, we do not have to care about stop words. Addressing the issue of very infrequent words, we only take into account terms that occur at least twice in the corpus.

2.2. Topic Creation

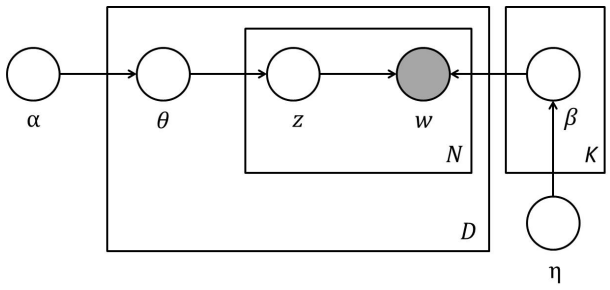

Figure 1: Graphical model for standard LDA

Topic models assume that the creation of a document resembles a generative process. The resulting document consists of a mixture of topics, with each topic consisting of a certain distribution of words. For example, a political document about the shut down of nuclear power plants could be a mixture

of the topics environment and economics, with the word energy appearing in topic economics with a high probability.

One of the most well-known topic models is Latent Dirichlet Allocation (LDA) by Blei et al. [9]. Figure 1 shows the graphical model for LDA. According to LDA, a collection D of documents is created the following way:

1. To receive the word distributions that describe the K available topics, draw each topic $\beta_i \sim \text{Dir}(\eta)$ for $i \in \{1, \ldots, K\}$, while $\text{Dir}(\eta)$ being a Dirichlet prior with parameters η.
2. Then, for each document in D draw the topic proportions $\theta \sim \text{Dir}(\alpha)$.
   (a) For each word in the document, draw the per-word topic assignment $Z_{d,n} \sim \text{Mult}(\theta_d)$ with $\text{Mult}(\theta_d)$ being a Multinomial mixture distribution depending on the topic proportions θ_d.
   (b) For each word, draw the word $W_{d,n} \sim \text{Mult}(\beta_{z_{d,n}})$ with $\text{Mult}(\beta_{z_{d,n}})$ being a Multinomial mixture distribution depending on the word distribution of the topic drawn for this word.

The gray shaded bubbles in the graphical model refer to the observed parameters. Assuming we have a collection of documents and we are interested in the topics they consist of, we need to invert this process, thus we are interested in inferring per-corpus topic distributions β_K. This can be done using state-of-the-art methods like Gibbs sampling [10].

2.2.1. Topic Creation with LabeledLDA

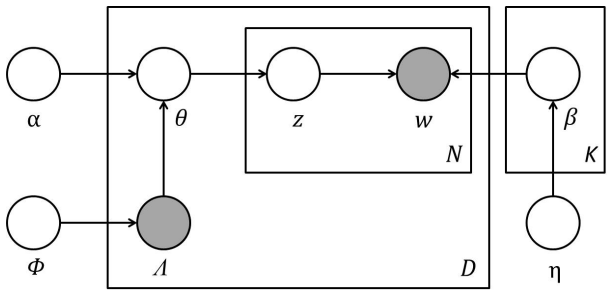

Figure 2: Graphical model for Labeled LDA

Labeled LDA by Ramage et al. [11] extends LDA by allowing for learning from multiply labeled documents, while labels correspond to topics. The

main difference to standard LDA is that when inferring the topics, the topics for a document are restricted to its given labels. Figure 2 shows the graphical model for Labeled LDA.

Λ denotes a list of binary topic presence/absence indicators. The number of topics K in this case corresponds to the amount of unique labels appearing in the documents. As mentioned before, θ is restricted to the labels Λ only. For this purpose, the document's labels Λ are generated with a Bernoulli coin toss with a labeling prior probability Φ, and θ is dependent on both α and Λ. For further details, please refer to Ramage et al..

For our approach we need to generate topics that can be compared among the output of multiple topic models, and we want to influence the content of the topics. Most important, we do not want to invest manual work into hand-coding documents manually, therefore we cannot apply Labeled LDA directly. However, we use a trick to produce labels following a simple heuristic. For each of the topics we want to extract from the documents, we have a set of seed words. As described in section 2.1, the documents we want to analyze are divided into snippets already. Now, we create labels for the snippets the following way: if a snippet contains a seed word for a topic, we add its topic as a label. Now, we can use the collection of snippets with their labels as input for Labeled LDA.

2.2.2. Topic Creation with LogicLDA

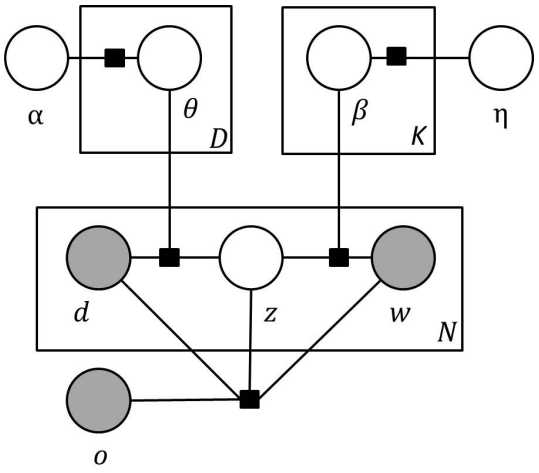

Figure 3: Factor graph for logicLDA

LogicLDA [12] by Andrzejewski et al. is an extension of LDA that offers the possibility to include first order knowledge.

The topics learned by the logicLDA model are influenced by both word-document statistics like in LDA and domain knowledge rules as in Markov Logic Networks. Figure 3 shows the standard logicLDA factor graph. It corresponds to the standard LDA model, except for the fact that there is an additional parameter o that denotes external observations. o is directly influencing the values of the topics z of a document and indirectly influencing the word-distributions that describe the topic β and the multinomial θ over topics for the document.

The type of knowledge integrated via o can be manifold. One possibility is to specify knowledge like 'The word *Euro* stems from *topic 2*', with topic 2 being e.g. 'finance'. This would be stated with the following rule:

$$W(i, Euro) \Rightarrow Z(i, 2) \tag{1}$$

More formally, Andrzejewski et al. define special predicates modeling the assignment of word tokens to documents

- $Z(i, t)$ is true iff the hidden topic $z_i = t$.
- $W(i, v)$ is true iff word $w_i = v$.
- $D(i, j)$ is true iff $d_i = j$.

We use them to link the topics to be created with seed words taken from external sources. For this purpose, we introduce a new predicate $SEED(w, t)$ that is true if a word w is a seed word for topic t. The general impact of seed words on the topic model is then described by the following knowledge base:

$$\bigwedge_{i=1}^{N} W(i, w) \wedge SEED(w, t) \implies Z(i, t)$$

Based on this general definition, we can now introduce additional rules for defining the SEED predicate, thereby defining what kind of words act as seed words for a certain topic.

The actual creation of the topic model consists of two steps. In the first step the topic structure is determined by setting the number of topics, selecting seed information for each topic and linking the seed information to the vocabulary created in the preparation step. In the second step, a topic model is generated using corpus statistics and the seed information using the LogicLDA respectively Labeled LDA system.

2.3. Measuring Topic-related Distance

The result of the topic creation is a set of multinomials over word tokens that represent the different topics in a document. According to our assumptions, these multinomials represent the position of the authors of a document with respect to the respective topic. In political science, it often is of interest how close the positions of different parties are on a certain issue. If our assumption is true, we can determine the distance of the positions of different parties with respect to a certain topic by measuring the distance between the multinomials representing the same topic.

Cosine similarity is a well established method for comparing the similarity of documents represented as sparse vectors which is defined as follows:

$$COS(q, r) = \frac{\sum_y q(y)r(y)}{\sqrt{\sum_y q(y)^2 \sum_y r(y)^2}}$$

A similar idea can be found in [13], in which Rosen-Zvi et al. present an author-topic model to determine authors and topics in a corpus. As for an application, they calculate the distance between authors using symmetric Kullback-Leibler divergence.

3. Experiments

We test the method described above in a number of experiments in the context of political science research. The purpose of these experiments is to test the ability of the proposed method to determine positions on particular topics stated in documents rather than to answer an actual research question in political science. In the following, we first provide a more detailed justification and the rational for the experiments carried out. Afterwards the data sources and the detailed experimental design are described.

3.1. Predicting ministries based on coalition contracts

As mentioned in the introduction, the goal of this work is to develop a content analysis method that is able to determine the (relative) position with respect to a certain topic stated in a document. As we have explained in the last section, we do this by creating a topic model whose topics are partially predefined by the use of seed words to make them comparable. We claim that the distribution of words in a topic of the resulting model represents

the position expressed in the document. In particular, we claim that the distance between the topic multinomials generated from different documents represent the distance of the positions stated in the two documents.

In this experiment, we test this hypothesis in an indirect way, analyzing party manifestos and coalition contracts. In particular, we determine the distances between the parties' positions stated in their manifestos and the coalition contract, and compare those distances among the two parties participating in the coalition. The underlying assumption is that the party that was to get control over the respective ministry has a stronger influence on the position stated in the coalition agreement on the topics represented by that ministry. Therefore, we can assume that the position on a topic stated in the coalition agreement is more similar to the position stated in the manifesto of the party that was assigned the ministry. In particular, we assume a data generation process, where first the ministries are assigned to parties, afterwards, the respective part of the coalition agreement is generated. We assume that the party in charge of a ministry also leads the generation of the related part of the coalition agreement, which is reflected in a stronger relation to the position of the respective party, both in terms of short term and long term positions. Further, we assume that the short term position of a party is reflected in the corresponding election manifesto while the long term position can be found in the latest basic party program available. However, our purpose is not to develop a system that predicts ministries. We intend to use this scenario to evaluate whether our system is able to determine distances between positions regarding specific topics.

We apply our method in the following way. First, we generate a separate topic model for each of the following documents:

- The party manifestos of the parties participating in a coalition.
- The coalition agreement.

For the creation of the topics, we use the policy areas provided by Seher and Pappi [14] which will be described in more detail in section 3.2. For each topic, we then measure the distance of each party to the coalition agreement. We expect the party with the lesser distance to the coalition contract to have the greater influence on the coalition contract regarding this topic. We consider our method to work as planned if our method is able to 'guess' the party that is in control of a certain ministry based on the positions generated

from the party manifestos and the coalition contract with a certain level of confidence.

3.2. Data Sources

In our previous work ([15]), we analyzed data from the last three German national elections (2002, 2005 and 2009). We extended the experiments with the previous three elections of 1990, 1994 and 1998. In all six elections, the coalition was formed by two parties. We have different variations of coalitions: in 1990, 1994 and 2009 it was a coalition between the CDU/CSU and FDP, with the FDP being the junior partner. Similarly, in 1998 and 2002 the SPD was the dominant partner in a coalition with the Greens. In contrast, in 2005 the election resulted in a grand coalition with the CDU and the SPD as (almost) equal partners.

We use plain text versions of party manifestos provided by the Manifesto Project Database[2](1990 - 1998) and the Mannheim Centre for European Social Research (MZES)[3] (2002 - 2009). As it turned out that in some cases using the manifesto from a single election only does not provide sufficient data to obtain meaningful statistics during the topic modeling process, we supplemented the election manifestos with the general programs of the respective parties[4] that we retrieved from the web and semi-automatically converted to plain text format. Finally, we used plain text versions of the coalition agreements provided by Sven-Oliver Proksch from the MZES.

In [14] Seher and Pappi investigate the topics addressed by German Parties on the level of federal states. For their analysis they use a set of 15 policy areas each characterized by a set of portfolios whose descriptions can be used as seed information[5]. We map the topics of their scheme to the German ministries having the responsibility for the respective political areas. The topics and the mappings to their corresponding ministries are the following:

- **Social Affairs** and Labour Market ('Arbeit und Soziales'):
  Federal Ministry of Labour and Social Affairs ('Bundesministerium für Arbeit und Soziales')

[2]https://manifesto-project.wzb.eu/

[3]http://www.mzes.uni-mannheim.de/projekte/polidoc_net/index_new.php?view=home

[4]The general programs originate from the following years: FDP: 1985/1997; SPD: 1997/2007; Greens: 1980/2002; CDU: 1978/1994/2007 respectively.

[5]The corresponding seed words are shown in appendix A

- **Culture** and Education ('Kultus'):
  Federal Ministry of Education and Research ('Bundesministerium für Bildung und Forschung')
- **Agriculture** ('Landwirtschaft'):
  Federal Ministry of Food, Agriculture and Consumer Protection ('Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz')
- **Finance** ('Finanzen'):
  Federal Ministry of Finance ('Bundesministerium der Finanzen')
- **Justice** ('Justiz') :
  Federal Ministry of Justice ('Bundesministerium der Justiz')
- **Internal Affairs** ('Inneres'):
  Federal Ministry of the Interior ('Bundesministerium des Innern')
- **Environment** and Regional Planning ('Umwelt und Landesplanung'):
  Federal Ministry for the Environment, Nature Conservation and Nuclear Safety ('Bundesumweltministerium')
- **Economics** and Transport ('Wirtschaft und Verkehr'):
  Federal Ministry for Economic Cooperation and Development ('Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung') / Federal Ministry of Transport, Building and Urban Development ('Bundesministerium für Verkehr, Bau und Stadtentwicklung')
- **Security** and Foreign Affairs ('Aussen- und Sicherheitspolitik')
  Federal Ministry of Defence ('Bundesministerium der Verteidigung') / Foreign Office ('Auswärtiges Amt')
- Development and Reconstruction ('Aufbau, Wiederaufbau'):
  not mapped to any ministry
- Building ('Bau'):
  not mapped to any ministry
- National and European Affairs ('Bund und Europa'):
  not mapped to any ministry
- Post War Effects ('Kriegsfolgen'):
  not mapped to any ministry

- Special Topics ('Sonderaufgaben'):
  not mapped to any ministry
- Chancellery ('Staatskanzlei'):
  not mapped to any ministry

For some ministries, there is no direct correspondence between the description of a topic and the responsibilities for a ministry. In the cases of Economics and Transport as well as Security and Foreign Affairs, we had to map the topic to two ministries each.

For better readability of the tables in the following sections, we shorten the name of the topics consisting of more than two terms to their first part, which is marked in the listing above by the bold printed terms.

3.3. Experimental Design

In the course of our experiments, we first transformed all documents into plain text format. We manually removed indexes and tables of contents. We appended the general program of a party to its party manifesto in order to extend the data. For each election, we applied the TextTiling Method to the extended manifestos of the two parties under consideration and to the coalition contract, obtaining three sets of documents. In the next step, we ran a POS-tagger on all documents and filtered for nouns, resulting in corpora whose documents consist of nouns only. According to [8], the POS-tagger has an accuracy of 97.53%.

For each corpus, we then generated the vocabulary which consists of all nouns that appear at least twice in the corpus.

The results are compared to a baseline described in section 3.4 for which we collect and compare the direct context around seed words for a topic as well as to a majority baseline.

Furthermore, we will compare the use of LogicLDA for the topic creation process with using Labeled LDA. Both systems are run using standard settings. To calculate the similarity between the output topics, we consider the 100 top ranked terms within a topic with their normalized probabilities. The resulting information is stored in a vector representation and the similarity of the vectors is computed using the Stanford OpenNLP API.

Finally, we will justify the decision to use nouns only and discuss the used seed set.

3.4. Baseline

We compare the results of the described method to a baseline. The purpose of our method is to analyze multiple dimensions, i.e. various topics, being within one single document. A straight forward approach to this task is extracting passages of the document for each dimension. This is typically done by human annotators: Based on a set of keywords (i.e. the one used by Seher and Pappi) the annotators search for sentences or passages containing these words and label them (after verifying the topic) with the respective class. We simulate this process by simply searching context around key words for a topic, using the seed words described in the previous section. We decided to extract each 20 words before and after the key word as context[6]. This results in a separate bucket of text snippets, each representing one dimension. The text snippets are filtered for nouns only.

We then compare the similarity between the coalition contract and the party opinion topic wise. To calculate the similarity, we represent the buckets of text snippets for each topic as a word vector, listing all terms with their frequencies.

In addition to this baseline, we compare the results to a majority baseline, which is based on the assumption that the stronger coalition partner gets to hold all ministries.

3.5. Results using LogicLDA for topic creation

In the following, we present the results of our experiments using logicLDA. In particular, we compare the outcome of the application of our method to the actual assignment of ministries to the coalition parties. We present the results based on using Cosine similarity and predicting the party whose topic is more similar to the topic created from the coalition contract to be in charge of the respective ministry. As it turns out, our method makes far less wrong predictions than the baseline method, some of which can even be explained by the specifics of the topics and the coalition. We present the results for each election individually as the parties involved and the ministries finally created differ from each election making it impossible to aggregate results in a meaningful way.

[6]Using sentences as basic units is a valid alternative. However, we dismissed this possibility, as the conversion of the original PDF documents did not always lead to intact sentence boundaries.

Analyzing 6 elections using 9 topical areas results in 54 single ministries to evaluate. Tables 2 to 7 show the results for each year and topic. They list the similarity of the parties and the coalition contract, marking correct predictions for ministries by "+", wrong decisions by "–" and ties by "**?**". The column "truth" shows the party that was actually in hold of the respective ministry.

To give an example, in 1990 (see table 2), the ministry for Social Affairs was actually held by the CDU, as noted in column "truth". Our logicLDA based method (stated on the left side of the table) computed a similarity of 0.10 between the CDU and the coalition contract for this topic, and a similarity of 0.19 for the FDP and the coalition contract. As the latter similarity is higher, our system predicts that the FDP is in hold of the ministry, which is wrong. This is marked by "–" in the next column.

As mentioned before, in some years the topical areas defined by seeds do not correspond directly to one particular ministry. Security and Foreign Affairs, for example, corresponds to the two ministries Foreign Office and Federal Ministry of Defence. Throughout all coalitions, those two ministries are held each by a different party. Therefore, it is not possible to predict the ministries with our method, as it cannot distinguish between the two different posts. The same holds for Economics and Transport corresponding to Federal Ministry for Economic Cooperation and Development / Federal Ministry of Transport, Building and Urban Development (except for 1998 and 2002) as well as for Environment and Regional Planning in 1990.

Summed up, this results in 11 particular items for which we are not able to draw a conclusion about the correctness of the method. In the resulting 43 cases, our method predicts the ministries correctly 32 times (74.4%). The baseline is correct in 20 cases only (46.5%). While our method predicts the wrong party 10 times, and is undecided in one case, the baseline is undecided in 11 cases and predicts the wrong party 12 times. We notice a strong variance in the behavior of the baseline: in 2009, it is a pure majority baseline, predicting CDU for all ministries. In 2002, it predicts either the Greens or is undecided, but never SPD. In 1998, it is undecided in nearly all cases, while in 2005 it is always undecided or wrong, except for one ministry.

The highest error rate is found for the ministry of Justice. This might be caused by the fact that there is no general preference of a certain party to hold this ministry, in contrast to some other ministries that are traditionally strongly bound to one particular party, like for example Agriculture for the CDU or Environment for the Greens.

For the interpretation of the results we would like remind the reader that the purpose of our method is not to seriously predict a ministry, but we just use this as an evaluation scenario. Otherwise, traditional preferences for ministries as well as the proportion of votes for each party would have to be considered as well. In 2002, for example, our method predicts the Greens 5 out of 9 times, though it is obviously unrealistic that the junior partner in a coalition gets more than half of the ministries.

Table 1 shows the output of the LogicLDA analysis for the topic Social Affairs and Labour Market for CDU, FDP and the Coalition contract in 1994. Seed words are printed in italics. The example shows that the term 'Gesellschaft' (society) is of importance to the topic for CDU and the coalition contract, though it had not been included in the seed words. Yet it was detected by the topic model. This ability to detect terms that show a strong relation to the seed words for an individual party makes the method more suitable for the task of identifying topics than the seed-based-only baseline.

In the following, we will describe the results for each election in more detail.

CDU	Coalition	FDP
familie = 0.0803	*arbeit = 0.1209*	*frauen = 0.0652*
frauen = 0.0694	*familie = 0.1060*	*arbeit = 0.0577*
gesellschaft = 0.0663	**gesellschaft = 0.1039**	menschen = 0.0477
kinder = 0.0597	aufgaben = 0.0910	*ausbildung = 0.0410*
arbeit = 0.0512	bürger = 0.0776	*integration = 0.0350*
familien = 0.0383	*frauen = 0.0612*	kinder = 0.0325
generationen = 0.0330	*integration = 0.0612*	*familie = 0.0289*
männer = 0.0315	erhaltung = 0.0590	bedeutung = 0.0205
integration = 0.0257	beitrag = 0.0583	kindern = 0.0203
kindern = 0.0239	form = 0.0497	länder = 0.0186
unterstützung = 0.0227	*ausbildung = 0.0463*	unterstützung = 0.0172
partnerschaft = 0.0203	energieversorgung = 0.0290	einrichtungen = 0.0166
ehe = 0.0182	erwerbsarbeit = 0.0264	zahl = 0.0132
beruf = 0.0173	*drogen = 0.0164*	angebot = 0.0131
angebot = 0.0163	beachtung = 0.0131	**gesellschaft = 0.0119**

Table 1: LogicLDA output for Social Affairs and Labour Market for the election of 1994 (*Seed words* are in italics.)

	LogicLDA			Truth		Baseline	
Policy Area	**CDU**	**FDP**		**Ministry**		**CDU**	**FDP**
Social Affairs	0.10	**0.19**	–	CDU	-	0.90	**0.93**
Culture	0.14	0.14	?	FDP	+	0.88	**0.91**
Agriculture	0.03	**0.11**	–	CDU	-	0.71	**0.78**
Finance	**0.18**	0.06	+	CDU	+	**0.49**	0.47
Justice	0.00	**0.14**	+	FDP	+	0.76	**0.82**
Internal Affairs	**0.18**	0.13	+	CDU	-	0.85	**0.88**
Environment	**0.35**	0.20	+/–	CDU / FDP [7]	–/+	0.88	**0.90**
Economics	0.51	**0.68**	+/–	FDP / CDU [8]	+/–	0.92	**0.94**
Security	0.01	**0.02**	+/–	FDP / CDU [9]	–/+	**0.60**	0.59

Table 2: Result of the Analysis of the German national elections 1990 using LogicLDA

German National Elections 1990 and 1994. Results for the elections of 1990 and 1994 are listed in tables 2 and 3. With three falsely predicted respectively undecided ministries per election, for these two years we received the worst results throughout all elections, scoring exactly as low as the baseline. The bad results for those two years can partly be explained by technical reasons resulting from the original PDF documents. In the early 90s, PDF documents did not directly contain the content as text data. To extract the content, they have to be converted to text via OCR based PDF converters. This is especially problematic for the party manifestos and general programs of the FDP, as their documents have a two column layout. Converting those documents, the order of the text blocks is not always kept correctly.

It is notable that in 1994 the similarity scores for Finance are especially low for both parties. This might be explained with the fact that there are only two seed words for this topic, namely 'Steuern' (taxes) and 'Finanzen' (finance). In the coalition contract of 1994, the first term occurs only once, and the latter one does not appear at all except from being part of compounds, where it cannot be identified.

German National Election 1998. For the elections of 1998 (table 4), the presented method only makes one false prediction, which is for the ministry of Justice. This might be explained by the fact that neither the Greens nor the SPD has a strong traditional focus on this domain. Our system clearly outperforms the baseline, which results in tie situations for 4 ministries.

	LogicLDA			Truth		Baseline	
Policy Area	**CDU**	**FDP**		**Ministry**		**CDU**	**FDP**
Social Affairs	**0.59**	0.53	+	CDU	-	0.93	**0.94**
Culture	**0.61**	0.59	+	CDU	?	0.92	0.92
Agriculture	**0.16**	0.13	+	CDU	+	**0.84**	0.80
Finance	0.01	**0.02**	–	CDU	+	**0.64**	0.63
Justice	**0.17**	0.13	–	FDP	+	0.86	**0.87**
Internal Affairs	0.31	**0.38**	–	CDU	–	0.92	**0.94**
Environment	**0.57**	0.37	+	CDU	+	**0.90**	0.88
Economics	0.62	**0.70**	+/–	FDP / CDU[10]	–/+	**0.97**	0.96
Security	**0.35**	0.02	–/+	FDP / CDU [11]	–/+	**0.87**	0.86

Table 3: Result of the Analysis of the German national elections 1994 using LogicLDA

	LogicLDA			Truth		Baseline	
Policy Area	**SPD**	**GRE**		**Ministry**		**SPD**	**GRE**
Social Affairs	**0.74**	0.72	+	SPD	+	**0.96**	0.95
Culture	**0.55**	0.44	+	SPD	?	0.93	0.93
Agriculture	0.13	**0.22**	+	GRE [12]	?	0.81	0.81
Finance	**0.03**	0.01	+	SPD	+	**0.74**	0.69
Justice	0.23	**0.39**	–	SPD	?	0.90	0.90
Internal Affairs	**0.32**	0.27	+	SPD	?	0.94	0.94
Environment	0.08	**0.12**	+	GRE	+	0.88	**0.89**
Economics	**0.27**	0.19	+	SPD	+	**0.93**	0.92
Security	0.00	**0.09**	+/–	GRE / SPD [13]	+/–	0.75	**0.82**

Table 4: Result of the Analysis of the German national elections 1998 LogicLDA

German National Election 2002. In table 5 we can see that our method correctly predicted most of the ministries. The method made a mistake on the area of Economics and Transport, this mistake can be explained, however, by the high relevance of environmental issues which is traditionally a green topic for the Transport area. Another mistake was made on Social Affairs and Labour Market where the method predicted the Greens to be in charge, whereas the ministry was taken by the SPD. Overall, we can see that the method was able to correctly predict six out of eight unambiguous areas.

In contrast, the baseline was not able to correctly predict the ministry in 5 cases.

	LogicLDA			Truth	Baseline		
Policy Area	**SPD**	**GRE**		**Ministry**		**SPD**	**GRE**
Social Affairs	0.67	**0.77**	–	SPD	?	0.96	0.96
Culture	**0.69**	0.56	+	SPD	?	0.96	0.96
Agriculture	0.10	**0.59**	+	GRE	+	0.91	**0.92**
Finance	**0.17**	0.04	+	SPD	–	0.78	**0.82**
Justice	**0.38**	0.21	+	SPD	–	0.92	**0.93**
Internal Affairs	**0.45**	0.43	+	SPD	?	0.96	0.96
Environment	0.53	**0.59**	+	GRE	+	0.94	**0.96**
Economics	0.45	**0.63**	–	SPD	?	0.97	0.97
Security	0.08	**0.14**	+/–	GRE/SPD[14]	+/–	0.85	**0.90**

Table 5: Result of the Analysis of the German national elections 2002 LogicLDA

German National Election 2005. For the 2005 election, we obtain a similar picture as shown in table 6. Making one mistake only on the Ministry of Justice, the system clearly outperforms the baseline, which makes 4 wrong predictions and has 2 ties. It is interesting to see that the values for the ambiguous cases (Economics and Transport which is represented in the Ministries of Economics and Technology occupied by the CDU and the Ministry of Transport which was given to the SPD) are very close to each other indicating an almost identical influence of the parties in the respective topics.

German National Election 2009. The best result was obtained on the 2009 election as we show in table 7. Here all unambiguous cases were correctly predicted by our method. The baseline contains one wrong prediction, however, it is a majority baseline in this case predicting CDU in all cases.

Finally, we briefly compare our method to a majority baseline. A majority baseline classifier assigns all ministries of a year to the party that holds the majority of ministries. In 2009, for example, it would predict that all ministries are held by CDU. Throughout all 6 elections we regarded in this experiment, the majority baseline classifier would make 11 wrong predictions for 54 ministries, our system 10. Please note that first of all, the majority baseline is hard to beat as in most years the ministries are highly unbalanced

	LogicLDA			Truth		Baseline	
Policy Area	**CDU**	**SPD**		**Ministry**		**CDU**	**SPD**
Social Affairs	0.39	**0.44**	+	SPD	–	**0.97**	0.96
Culture	**0.71**	0.58	+	CDU	+	**0.97**	0.96
Agriculture	**0.07**	0.05	+	CDU	–	0.86	**0.87**
Finance	0.06	**0.22**	+	SPD	–	**0.84**	0.75
Justice	**0.46**	0.03	–	SPD	–	**0.93**	0.91
Internal Affairs	**0.56**	0.49	+	CDU	?	0.96	0.96
Environment	0.18	**0.26**	–	SPD	?	0.93	0.93
Economics	**0.67**	0.66	+/–	CDU/SPD[15]	+/–	**0.97**	0.96
Security	**0.03**	0.01	+/–	SPD/CDU[16]	–/+	**0.91**	0.88

Table 6: Result of the Analysis of the German national elections 2005 LogicLDA

	LogicLDA			Truth		Baseline	
Policy Area	**CDU**	**FDP**		**Ministry**		**CDU**	**FDP**
Social Affairs	**0.62**	0.37	+	CDU	+	**0.97**	0.95
Culture	**0.81**	0.72	+	CDU	+	**0.98**	0.96
Agriculture	**0.52**	0.16	+	CDU	+	**0.94**	0.87
Finance	**0.23**	0.04	+	CDU	+	**0.83**	0.82
Justice	0.07	**0.32**	+	FDP	–	**0.93**	0.90
Internal Affairs	**0.41**	0.30	+	CDU	+	**0.96**	0.94
Environment	**0.22**	0.20	+	CDU	+	**0.95**	0.92
Economics	**0.79**	0.72	+/–	FDP/CDU[17]	–/+	**0.98**	0.96
Security	0.08	**0.55**	+/–	FDP/CDU[18]	–/+	**0.91**	0.88

Table 7: Result of the Analysis of the German national elections 2009 LogicLDA

between the parties. Furthermore, it is not our purpose to create a prediction system for coalitions, as this would have to consider many other factors beside the party manifesto, but we just want to verify whether our system is able to detect political positions stated in text. In the election of 2005 that resulted in a grand coalition between CDU and SPD with nearly equally distributed ministries, our system only makes 2 false predictions.

Using all content words. In 2.1, we explained that we keep nouns only for our experiments. Before deciding on this, we ran several experiments on the

influence of the kept word types. As nouns clearly outperformed other variants and as this is not a surprising outcome, we will keep the reporting about these experiments short: we just give some numbers for performing experiments keeping all content words. Those include nouns, verbs, adjectives and adverbs while dismissing pronouns, conjunctions and the like.[19]

Running our system with logicLDA keeping all content words on all elections from 1990 - 2009, only 25 ministries are correctly predicted and 17 falsely, the rest is ties. In contrast, the same system keeping nouns only results in 31 correctly predicted ministries and 10 mistakes, whilst the rest being ties.

3.6. Results using Labeled LDA for topic creation

To investigate the influence of the tool used for the topic modeling, we repeat the experiments for the years 2002-2009 with Labeled LDA. We observe a performance similar to LogicLDA. Throughout those years, there are 13 ministries for which the baseline makes false predictions or cannot predict the correct party, compared to 9 false predictions made by the system using Labeled LDA.

In 2002, shown by table 8, our system makes three wrong predictions. Like for LogicLDA, Social Affairs is one of the erroneously predicted ministries. The two other false predictions are Internal Affairs and Environment.

For the grand coalition in 2005 (results shown in table 9), the performance is worse than that of LogicLDA. However, in most cases of wrong prediction the similarity scores of both parties do not show a big difference: For culture, the similarity of the CDU with the coalition contract is 0.25, while that of SPD with the coalition contract being 0.27. Accordingly, for Justice we observe the similarities 0.27 (CDU) compared to 0.25 (SPD), and for Internal Affairs 0.21 (CDU) compared to 0.22 (SPD). It would be interesting to have an expert's opinion on whether the two parties indeed do have very similar positions towards those topics.

The system using Labeled LDA made two mistakes for the coalition in 2009, shown in table 10: Culture and Internal Affairs.

[19] We also experimented with stemming. As it did not change the results significantly, we omit to report the results and focus on more expressive experiments.

	Labeled LDA			Truth		Baseline	
Policy Area	**SPD**	**GRE**		**Ministry**		**SPD**	**GRE**
Social Affairs	0.08	**0.15**	-	SPD	?	0.96	0.96
Culture	**0.54**	0.43	+	SPD	?	0.96	0.96
Agriculture	0.33	**0.56**	+	GRE	+	0.91	**0.92**
Finance	**0.16**	0.09	+	SPD	–	0.78	**0.82**
Justice	**0.35**	0.33	+	SPD	–	0.92	**0.93**
Internal Affairs	0.17	**0.18**	-	SPD	?	0.96	0.96
Environment	**0.44**	0.33	-	GRE	+	0.94	**0.96**
Economics	**0.53**	0.45	+	SPD	?	0.97	0.97
Security	0.52	**0.54**	+/–	GRE/SPD[20]	+/–	0.85	**0.90**

Table 8: Result of the Analysis of the German national elections 2002 using Labeled LDA

	Labeled LDA			Truth		Baseline	
Policy Area	**CDU**	**SPD**		**Ministry**		**CDU**	**SPD**
Social Affairs	0.37	**0.45**	+	SPD	–	**0.97**	0.96
Culture	0.25	**0.27**	–	CDU	+	**0.97**	0.96
Agriculture	**0.28**	0.16	+	CDU	–	0.86	**0.87**
Finance	**0.19**	0.10	–	SPD	–	**0.84**	0.75
Justice	**0.27**	0.25	–	SPD	–	**0.93**	0.91
Internal Affairs	0.21	**0.22**	–	CDU	?	0.96	0.96
Environment	0.04	**0.22**	+	SPD	?	0.93	0.93
Economics	**0.44**	0.33	+/–	CDU/SPD[21]	+/–	**0.97**	0.96
Security	**0.60**	0.47	–/+	SPD/CDU[22]	–/+	**0.91**	0.88

Table 9: Result of the Analysis of the German national elections 2005 using Labeled LDA

3.7. Impact of the seed terms

The choice of suitable topics with appropriate seed terms seems crucial for our task. To investigate the impact of the used seed terms, we ran experiments with a different seed. In addition, we will discuss statistics of the occurrence of the initial seed terms.

As an alternative to the political areas defined by Seher and Pappi [14], we generated a seed set for each of the following ministries:

- Federal Ministry of Defence ('Bundesministerium der Verteidigung')

	Labeled LDA			Truth		Baseline	
Policy Area	**CDU**	**FDP**		**Ministry**		**CDU**	**FDP**
Social Affairs	**0.38**	0.31	+	CDU	+	**0.97**	0.95
Culture	0.57	**0.63**	–	CDU	+	**0.98**	0.96
Agriculture	**0.48**	0.30	+	CDU	+	**0.94**	0.87
Finance	**0.31**	0.27	+	CDU	+	**0.83**	0.82
Justice	0.13	**0.26**	+	FDP	–	**0.93**	0.90
Internal Affairs	0.13	**0.15**	–	CDU	+	**0.96**	0.94
Environment	**0.30**	0.22	+	CDU	+	**0.95**	0.92
Economics	**0.40**	0.37	–/+	FDP/CDU[23]	–/+	**0.98**	0.96
Security	0.43	**0.46**	+/–	FDP/CDU[24]	–/+	**0.91**	0.88

Table 10: Result of the Analysis of the German national elections 2009 using Labeled LDA

- Foreign Office ('Auswärtiges Amt')
- Federal Ministry of Education and Research ('Bundesministerium für Bildung und Forschung')
- Federal Ministry of Food, Agriculture and Consumer Protection ('Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz')
- Federal Ministry of Health ('Bundesministerium für Gesundheit')
- Federal Ministry of the Interior ('Bundesministerium des Innern')
- Federal Ministry of Labour and Social Affairs ('Bundesministerium für Arbeit und Soziales')
- Federal Ministry for the Environment, Nature Conservation and Nuclear Safety ('Bundesumweltministerium')
- Federal Ministry of Transport, Building and Urban Development ('Bundesministerium für Verkehr, Bau und Stadtentwicklung')

For each ministry, we looked up its description in Wikipedia and extracted all nouns appearing in the article. The links to the Articles can be found in the appendix. We repeated the above mentioned experiment while just replacing the expert created seed set by the fully automatically generated one.

For each of the 6 years, we analyze 9 ministries, which leads to 54 single predictions. Out of those, the system using Wikipedia-generated seeds makes 28 correct predictions and 25 false ones, while it cannot decide for one ministry. This performance is clearly lower as that of using the manually created seed set which produces up to 74.4% correct predictions.

This suggests that the quality of seeds defined by an expert makes a large difference. We assume that the results of the experiment could be increased even more with a seed set tailored to this task.

In order to get an impression of the overlap between the seed words and the analyzed documents, we calculated some basic statistics listed in table 11. The second column states the amount of seed words for the topic on the left, e.g. there are 24 seed words indicating the topic Social Affairs. The column "average occurrences" gives the number how often each of those seed words occurred on average per document (consisting of either the party manifesto and program of a party or of the coalition contract) and year. So on average, each of the 24 seeds of Social Affairs occurred on average 8 times per analyzed document. Instead of giving the average standard deviation, we decided to calculate the standard deviation per seed and give only the highest value we observed. This means, for the topic Social Affairs, one seed had a standard deviation of 45.84, which is very high. This means, some seed words occur with a high frequency in one document whereas they are hardly observed in another one.

Comparing the amount of seeds per topic, we notice large differences: while there are 27 seed words for Economics, there are only 2 for Finance. However, this does not seem to influence the quality of the results: in our experiments with LogicLDA, we predict the false party for the corresponding ministry only in one out of 6 elections. Furthermore, there is a large span of average occurrences per seed, ranging from only 2.22 to 8. Considering also the sometimes very high standard deviations for the occurrences of seeds per document, it becomes salient that there is a very high variance in the occurrence of seeds. Thus, the seeds are unequally important for each document, and it is hard to predict how the lack of one single seed influences the performance of the whole approach, as it strongly depends on the seed term itself and the analyzed document.

Topic	# seeds	∅ occurrences	max stdev
Social Affairs	24	8	45.84
Culture	18	7.41	17.79
Agriculture	11	2.79	10.2
Finance	2	5.64	7.96
Internal Affairs	21	5.75	26.88
Justice	21	2.22	10.91
Internal Affairs	21	5.75	26.88
Environment	22	3.01	13.17
Economics	27	6.42	15.93
Security	8	2.44	4.25

Table 11: Statistics of the occurrence of seed terms for topics in the data.

4. Application Example

In this section, we want to give an expression how our method could be used by political scientists to analyze party positions. We apply the above described method to all party manifestos (enriched by the general programme) of the 5 parties that were elected into the government in 2009, namely CDU, Greens (GRE), FDP, PDS, SPD. We calculate the similarity among the parties pairwise for each topic. Due to the paper limitations, we cannot display the numbers for all topics, so we decided on giving the topics Economics and Transport as well as Internal Affairs as an example.

The result of the analysis for one topic are the similarities for the ten possible pairs of the 5 parties. As it is inconvenient to the majority of the readers to interpret those numbers, we chose a to some extent visual representation as a pyramid, sorting the pairs by their similarity in a descending order (see figures 4 and 5). The two parties in the first row have the highest similarity, the one in the last row the lowest.

Please note that as we are no political scientists, we do not want to interpret the results rather than showing the output of our method. Therefore, we try to give some comments on the results based on an average citizen's knowledge about politics in Germany.

Traditionally, the parties are placed on a left-right scale in the following order: PDS, Greens, SPD, CDU, FDP.

Figure 4 shows the results for Economics. At a very first glance, the

```
      CDU SPD          (0.65)
     CDU   FDP         (0.56)
    FDP     SPD        (0.53)
   FDP       GRE       (0.45)
  CDU         GRE      (0.43)
 CDU           PDS     (0.42)
GRE             PDS    (0.39)
GRE              SPD   (0.39)
FDP               PDS  (0.29)
```

Figure 4: Pairwise similarity for Economics (2009)

most similar pair being CDU and SPD is somewhat surprising, as they are always considered as the two big competing opponents. However, they are the parties in the middle of the traditional scale and they are both considered as moderate parties, and their views on economics might in deed not vary that much. Next, let us have a look at the similarity of the other parties to the most right-wing party FDP. The party closest to it is CDU followed by SPD and Greens, and with the largest distance finally the PDS. This is consistent with the common left-right scale. Another salient observation is the position of the PDS: all pairs including it range in the bottom part of the pyramid. This can be due to the fact that this party in deed has very specific opinions. Furthermore, due to the fact that they have never been participating in a governing coalition, they do not have to consider possible coalition possibilities when stating their goals before an election, but can keep their to some extent extreme position.

Figure 5 shows the results for Internal Affairs. Like for Economics, CDU and SPD have very close positions, although here the most similar parties are PDS and SPD. The most salient observation is that the parties that are considered left-wing (PDS, SPD, Greens) all have very similar positions with their pairs ranging on the upper part of the pyramid.

Interestingly, CDU and FDP, both being considered as right wing parties, have the lowest similarity considering Internal Affairs. This is not further surprising, as it is well known that in deed they do have different views on inner state security: while the FDP aims towards a liberal state, the CDU typically takes a law and order position on this question.

```
      PDS SPD          (0.49)
     CDU   SPD         (0.47)
    CDU     GRE        (0.43)
   GRE       PDS       (0.36)
  FDP         SPD      (0.33)
 GRE           SPD     (0.32)
CDU             PDS    (0.30)
 FDP              GRE  (0.27)
  CDU               FDP (0.25)
FDP                  PDS (0.24)
```

Figure 5: Pairwise similarity for Internal Affairs (2009)

5. Conclusions

The work presented in this paper shows that it is possible to use topic modeling as a basis for multi-dimensional analysis of political documents pushing the limits of automatic content analysis in the social sciences. Our experiments show that we can relatively reliably determine topics related to predefined policy areas and compare individual topics across documents. The method consistently outperforms a baseline directly extracting context around seed terms. As we have mentioned before, these results do not have a direct value for research in political science. Yet they provide a proof of concept that we can build upon for addressing questions in the area of party competition which currently cannot be addressed without the need for manual coding. There are a number of issues that have to be investigated in more detail in future work before we can apply this method to open research questions in political science.

The most central problem is the choice of the right policy areas and seed information for a given question. As mentioned before, we used a coding scheme introduced by Seher and Pappi for policy analysis on the regional level. This already led to some problems when applying it to the national level and we had to exclude some areas not relevant in the context of national elections. On the European level, again different policy areas with different

scope are relevant[25]. While for example Competition is a central area on the European level, it almost does not play any role on the regional level. In a similar way, policy areas as well as their focus change over time. While in the early phase of German politics post-war issues like Compensation and Denazification was a major issue, later periods were dominated by the cold war, these topics are not relevant today any more. Similarly new areas like environmental protection have emerged and gained importance. This means that the determination of relevant topics and seed information is a scientific problem that requires expertise in political science. From a technical point of view, the extraction of relevant seed information from existing knowledge resources is a problem that needs to be addressed. In addition, as the expert created seed set showed good results, it should be investigated whether an existing seed word set could be automatically extended in a useful way.

We have restricted our attention to the generation of topics and the distance between topics. Most related work in the political sciences, however, focuses on the projection of party positions on different scales (i.e. left-right or liberal-conservative). In future work we will investigate the projection of the generated models on a multidimensional scale. This will support researchers to carry out well established scale-based analysis while taking the different topic areas into account. Such an approach would solve the problems of one-dimensional analysis outlined in the motivation.

The main contribution of our work is to analyze documents containing more than one topic in a full automatic manner. We would like to measure directly how big the improvement acually is compared to traditional methods which manually divide the texts into thematically coherent sentences or paragraphs, neglecting the fact that even very small units of text might contain several topics. The Manifesto Project[26] is an important research project providing sentence-based hand-coded party programmes from manifold countries and years. We aim at finding a context in which we can make use of those data in a setting that enables the comparison with our automatically created topic positions.

A possible application of the proposed method could be to locate legislative bills in coalition agreements and party manifestos. The purpose of this application is to investigate on the one hand whether a government has

[25]compare http://europa.eu/pol/index_en.htm

[26]https://manifesto-project.wzb.eu/

previously stated their intentions for laws in the coalition agreement, and on the other hand if one of the parties has previously mentioned the issue in their manifesto. To approach this task with the presented method, the terms used in the bills can be employed as seeds, while treating each proposal as a separate topic. After creating the topic models on the investigated party documents (party manifestos, coalition agreements), the distance between the resulting topic and the original bill text can be calculated. If it is small enough, it indicates that the bill was initiated within this document.

Acknowledgements

We want to thank Sven-Oliver Proksch of the Mannheim Centre for European Social Research (MZES) for providing the idea to evaluate our approach by comparing party positions to the coalition agreements and predicting ministries.

Appendix A. Seed Words

- Social Affairs and Labour Market
  Arbeit Soziales Ausländerintegration Behinderte Drogen Ehrenamt Familie Frauen Generationen Gesundheit Gleichstellung Integration Jugend KITA Kindergarten Betreuung Verwahranstalt Mindestlohn Qualifikation Pflegeberuf Ausbildung Prostitution Senioren Wohlfahrt

- Culture and Education
  Kultus Bafög Studiengebühren Bildung Denkmalschutz Doping Ethik Forschung Frühkindliche Bildung Gentechnik Juristenausbildung Kirche Kultur Kunst Medienpolitik Schule Sport Vorschuljahr Weiterbildung Wissenschaft

- Agriculture
  Landwirtschaft Ernährung Fischerei Forsten Gentechnik Jagd Kleingärten ländliche Räume Tourismus Verbraucherschutz Weinbau

- Finance
  Finanzen Steuern

- Justice
  Justiz 218 Abtreibung Asyl Bankgeheimnis Bürgerrechte Datenschutz Frauenhaus Frauenhandel Menschenhandel Gefängnis Strafvollzug Gewalt Kriminalität

- Internal Affairs
  Inneres Auswanderungswesen Bezirksverwaltung Verwaltung Bürgerbegehren Bürokratieabbau Demokratie Einwanderung Katastrophenschutz
- Environment and Regional Planning
  Hochwasserschutz Lärmschutz Wirtschaft Verkehr Atomausstieg Auenhandel Bahn Energie Existenzgründung Hafen Infrastruktur Innovation Kreditwesen Banken Medienstandort Infrastruktur Staatskanzlei Ministerpräsident
- Security and Foreign Affairs
  Auenpolitik Sicherheitspolitik Entwicklungshilfe Militär Verteidigung Wehrdienst Wehrpflicht Zivildienst

Appendix B. Articles used to generate the Wikipedia seed set

- http://de.wikipedia.org/wiki/Bundesministerium_der_Verteidigung
- http://de.wikipedia.org/wiki/Ausw%C3%A4rtiges_Amt
- http://de.wikipedia.org/wiki/Bundesministerium_f%C3%BCr_Bildung_und_Forschung
- http://de.wikipedia.org/wiki/Bundesministerium_f%C3%BCr_Ern%C3%A4hrung,_Landwirtschaft_und_Verbraucherschutz
- http://de.wikipedia.org/wiki/Bundesministerium_f%C3%BCr_Gesundheit_(Deutschland)
- http://de.wikipedia.org/wiki/Bundesministerium_des_Innern
- http://de.wikipedia.org/wiki/Bundesministerium_fr_Arbeit_und_Soziales
- http://de.wikipedia.org/wiki/Bundesumweltministerium
- http://de.wikipedia.org/wiki/Bundesministerium_f%C3%BCr_Verkehr,_Bau_und_Stadtentwicklung

We used the versions of the pages from October 5th, 2012.

References

[1] M. Laver, J. Garry, Estimating policy positions from political texts, American Journal of Political Science 44 (3) (2000) 619–634.

[2] M. Laver, E. Sergenti, Party Competition: An Agent-Based Model, Princeton University Press, 2011.

[3] K. Benoit, S. Mikhaylov, M. Laver, Treating words as data with error: Uncertainty in text statements of policy positions, American Journal of Political Science 53 (2) (2009) 495–513.

[4] J. Slapin, S.-O. Proksch, A scaling model for estimating time-series policy positions from texts, American Journal of Political Science 52 (3) (2008) 705–722.

[5] F. Pappi, N. Seher, A.-S. Kurella, Das politikangebot deutscher parteien in den bundestagswahlen seit 1976 im dimensionsweisen vergleich: Gesamtskala und politikfeldspezifische skalen, Working Paper 142, Mannheimer Zentrum für Europäische Sozialforschung (MZES) (2011).

[6] A. Volkens, O. Lacewell, P. Lehmann, S. Regel, H. Schultze, A. Werner, The Manifesto Data Collection, Manifesto Project (MRG/CMP/MARPOR), Wissenschaftszentrum Berlin für Sozialforschung (WZB), 2011.

[7] M. Hearst, Texttiling: Segmenting text into multi-paragraph subtopic passages, Computational Linguistics 23 (1) (1997) 33–64.

[8] H. Schmid, Improvements in part-of-speech tagging with an application to german, in: In Proceedings of the ACL SIGDAT-Workshop, 1995, pp. 47–50.

[9] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research (JMLR) 3 (2003) 993–1022.

[10] G. Casella, E. George, Explaining the gibbs sampler, The American Statistician 46 (3) (1992) 167–174.

[11] D. Ramage, D. Hall, R. Nallapati, C. D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 248–256.
URL `http://dl.acm.org/citation.cfm?id=1699510.1699543`

[12] D. Andrzejewski, X. Zhu, M. Craven, B. Recht, A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic., in: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), 2011.

[13] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th conference on Uncertainty in artificial intelligence, UAI '04, AUAI Press, Arlington, Virginia, United States, 2004, pp. 487–494.
URL http://dl.acm.org/citation.cfm?id=1036843.1036902

[14] N. Seher, F. Pappi, Politikfeldspezifische Positionen der Landesverbände der deutschen Parteien, Working Paper 139, Mannheimer Zentrum für Europäische Sozialforschung (MZES) (2011).

[15] H. Stuckenschmidt, C. Zirn, Multi-dimensional analysis of political documents, in: Natural language processing and information systems : 17th International Conference on Applications of Natural Language to Information Systems, NLDB, Groningen, 2012.