Fine-grained Position Analysis for Political Texts

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Cäcilia Zirn
aus Rottweil

Mannheim, 2016

Dekan:	Professor Dr. Heinz Jürgen Müller, Universität Mannheim
Referent:	Professor Dr. Heiner Stuckenschmidt, Universität Mannheim
Korreferent:	Professor Dr. Michael Strube, Heidelberger Institut für Theoretische Studien

Tag der mündlichen Prüfung: 6. September 2016

Abstract

Meinungsanalyse auf politischen Textdaten hat im Bereich der Computerlinguistik in den letzten Jahren stets an Bedeutung gewonnen. Dabei werden politische Texte zumeist in voneinander diskrete Klassen unterteilt, wie zum Beispiel *pro* vs. *contra* oder *links* vs. *rechts*. In den Politikwissenschaften dagegen werden bei der Analyse von politischen Texten Positionen auf Skalen mit fließenden Werten abgebildet. Diese feingranulare Darstellung ist für die dort gegebenen Fragestellungen erforderlich. Das Feld der "quantitativen Analyse" - der automatisierten Analyse von Texten - die der traditionellen qualitativen Analyse gegenüber steht, hat erst kürzlich mehr Beachtung gefunden. Bisher werden Texte dabei zumeist lediglich durch Worthäufigkeiten dargestellt und ohne jegliche Struktur modelliert.

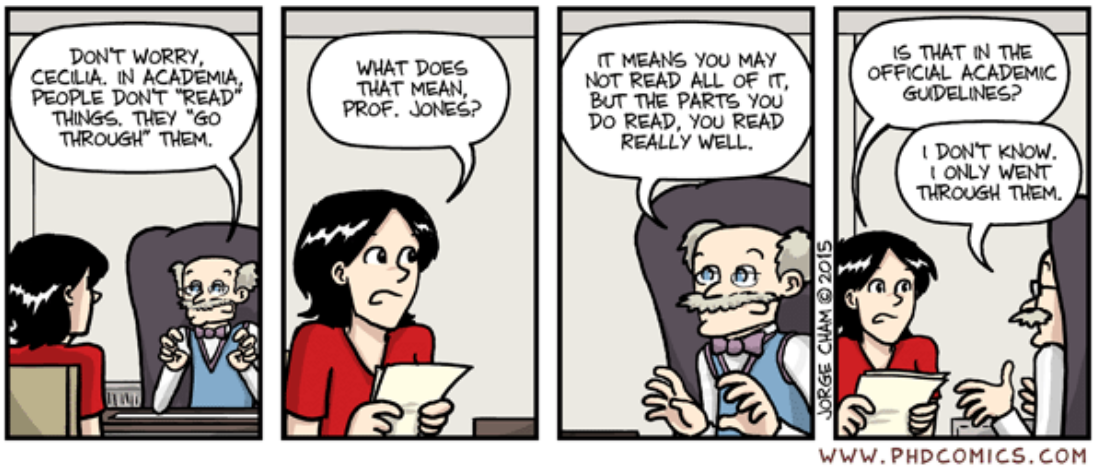
Wir entwickeln in dieser Dissertation Ansätze basierend auf Methoden der Computerlinguistik und der Informatik, die gegeignet sind, politikwissenschaftliche Forschungsfragen zu untersuchen. Im Gegensatz zu bisherigen Arbeiten in der Computerlinguistik klassifizieren wir nicht diskrete Klassen von Meinungen, sondern projizieren feingranulare Positionen auf fließende Skalen. Darüber hinaus schreiben wir nicht Dokumenten ganzheitlich eine Position zu, sondern bestimmen die Meinungen zu den jeweiligen Themen, die in den Texten enthalten sind. Diese mehrdimensionale Meinungsanalyse ist nach unserem Kenntnisstand neu im Bereich der quantitativen Analyse.

Was unsere Ansätze von anderen Methoden unterscheidet, sind insbesondere folgende zwei Eigenschaften: Zum Einen nutzen wir Wissen aus externen Quellen, das wir in die Verfahren einfließen lassen - beispielsweise integrieren wir die Beschreibungen von Ministerien des Bundestags als Definition von politischen Themenbereichen, mit welchen wir automatisch Themen in Parteiprogrammen erkennen. Zum Anderen reichern wir unsere Verfahren mit linguistischem Wissen über Textkomposition und Dialogstruktur an. Somit gelingt uns eine tiefere Modellierung der Textstruktur.

Anhand der folgenden drei Fragestellungen aus dem Bereich der Politikwissenschaften untersuchen wir die Umsetzung der oben beschriebenen Methoden:

1. Multi-Dimensionale Positionsanalyse von Parteiprogrammen
2. Analyse von Themen und Positionen in der US-Präsidentschaftswahl
3. Bestimmen von Dove-Hawk-Positionen in Diskussionen der amerikanischen Zentralbank

Wir zeigen, dass die vorgestellten Lösungen erfolreich feingranulare Positionen in den jeweiligen Daten erkennen und analysieren Möglichkeiten sowie Grenzen dieser zukunftsweisenden Verfahren.



"Piled Higher and Deeper" (PhD), 9/28/2015

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

When talking about political opinions, the first thing that comes to one's mind are opposing positions like *liberal* vs. *conservative*, *pro* vs. *contra* or *left* vs. *right*. We subscribe those positions particular values and ideologies. *Left* and *right*, for example, take a particular stand towards certain topics like economy or family policies.

For most political issues, however, it is not possible to clearly divide positions into two opposing views. Regarding the German party landscape, we find the *AfD* (short for *Alternative für Deutschland*) on the right side of the spectrum, and *Die Linke* on the left side. Due to this, they should have completely opposing views on political issues. It might be surprising that despite this, they do have similar opinions towards certain topics. To give an example, they are both skeptic towards the European Union and the Euro. *Die Linke* has concerns about the current structure and political direction of the European Union, and they aim at strengthening its social and democratic dimensions. During the European debt crisis that began in 2009, this political party questioned the common Eurozone. When the target rate was at its record low in 2013, Oskar Lafontaine, one of *Die Linke*'s former leading politicians, even suggested the return to separate currencies[1].

The same suggestion is supported by the *AfD*: the party claims in their manifesto that European debts and bank liabilities are detrimental to the competitiveness of all countries in the Eurozone, and that Germany should therefore leave it[2].

Albeit for different reasons, both parties share their resentments towards the current state of the European Union and the Eurozone. This example shows that general distinction of the political spectrum into *left* and *right* is not adequately capturing political landscapes. For this reason

[1] https://www.die-linke.de/politik/themen/detail/artikel/raus-aus-dem-euro/
[2] https://www.alternativefuer.de/programm-hintergrund/fragen-und-antworten/zu-euro-und-waehrungspolitik/

we suggest that it is necessary to consider political positions as being **multi-dimensional**, consisting of separate opinions towards particular topics.

The field of quantitative analysis in political science assesses text of political documents in a statistical way to answer research questions. These research questions often correspond to opinion mining tasks: a typical application is for instance to analyze party ideologies based on the parties' election manifestos. Common approaches assume one single political ideology per document. They place documents on ideological scales in order to compare and analyze the contained positions. These scales usually reach from *left* to *right* or from *conservative* to *liberal*. As such approaches assume one single ideology - thus a single position - per document, they are one-dimensional. This way of modeling the task, however, does not fit our assumption that ideologies have to be seen with respect to particular topics. For a proper analysis of party positions expressed in manifestos, the positions should be compared on the level of the topics they cover. In other words, there is the need for a multi-dimensional analysis where the dimensions correspond to the topics tackled in the manifestos.

In the area of computational linguistics, there exist some approaches for opinion analysis that focus on extracting positions towards particular topics rather than accessing the overall position of a document. These methods mostly classify positions into distinct stances: taking the legalization of abortion as an example, such stances could be *pro* and *contra*.

Revisiting our example from the beginning about the stance towards the European Union and the Eurozone, we can see that there might be different reasons for taking a particular position which induce different strengths in opinion: both *Die Linke* and *AfD* are EU critic, but while the first requests a change of the European Union and the Eurozone, the latter clearly rejects those institutions in general. Furthermore, there might be a different level of inter-party agreement towards certain issues. This shows that in the context of complex political disputes a classification of views into two distinct, opposing classes does not model potential positions sufficiently. In fact positions should be rather modeled in a more *fine-grained* way according to their strength. This is what we find in the quantitative analysis approaches: scholars of political science prefer to place positions on a continuous scale.

Our goal is to provide methods that allow for addressing actual **research questions from the field of political science**, namely analyzing positions stated in coalition contracts, positions of presidential candidates expressed in public speeches, and positions taken in discussions of a central bank's steering committee. These methods should be appropriate to work on documents that are written in a highly domain-specific language by domain experts, so called political agents. Examples are party manifestos, speeches or transcriptions of discussions in

expert committees. This type of language is more difficult to analyze than content produced by political non-experts - like on-line debates - as they require domain-knowledge to be properly understood. In contrast to previous approaches, we provide multi-faceted approaches capable of determining opinions towards particular topics and placing those on a fine-grained continuous scale. We increase the effectiveness of common methods in computer science, computational linguistics and political science two-fold: First, we integrate **external knowledge sources**, such as descriptions of ministry responsibilities in order to determine topics in party manifestos. Second, we make use of **linguistic knowledge** about structure, discourse and coherence of a text. As an example, consider the order of particular topics in a party manifesto: after addressing family policies, it is more likely to tackle labor or education policies rather than military issues. This information can be added to a model as a constraint.

In the remainder of this chapter, we will first give an introduction into opinion analysis on political documents, followed by elaborating the objectives of this thesis in Section 1.3. In Section 1.4, we will give an overview over the organization of this thesis and introduce the particular tasks that we addressed in our work.

1.2 Introduction to Opinion Mining

Since the beginning of the millennium, sentiment analysis and opinion mining are major topics in the field of computational linguistics. This research was motivated by the attempt to cope with the huge amount of product or movie reviews on the web which can be found on platforms like Amazon[3] or Rotten Tomatoes[4]. Pang et al. [Pang et al., 2002] are well-known for one of the first approaches of sentiment analysis. They applied machine learning to predict movie ratings from review texts. Their study revealed interesting insights when they compared the most important words humans would expect and the ones statistically extracted for the distinction between positive and negative reviews: good cues for negative sentiment are for example the punctuation marks "?" and "!", due to sentences like "What was the director thinking?".

Later approaches focus on opinion analysis of texts that are more difficult to analyze than opinions recommending or dissuading a product or movie, such as personal opinions in social networks and on-line platforms.

[3] `http://www.amazon.com`
[4] `https://www.rottentomatoes.com/`

The stances taken in political opinion exchanges are not necessarily *pro* or *contra* the disputed topic; participants can rather be grouped into supporting *democrats* or *republicans*, taking the U.S. American political landscape as an example. Having a closer look at the argumentation why people support or oppose something, one can even observe different stances within the same group, as people have different reasons why they support or oppose something. An example is the recent discussion in the United States whether abortion should be forbidden ("pro life") or allowed ("pro choice") and if allowed, up to which status of the pregnancy.

The next level achievements in opinion analysis were approaches for analyzing on-line debates. These approaches grouped participants of on-line debates into subgroups based on shared stances they took towards the discussed subject, and they employ either machine learning or unsupervised clustering methods.

In such on-line debates that are open to public, the stances taken by the users are mostly formulated in a compact, straight forward way revealing a clear stance. We pick some examples from the platform CreateDebate[5] related to the current presidential election campaign:

> *"Trump is the best candidate to Make America Great Again! Also, he will stop illegals from ruining this county"*[6]

> *"Clinton. Because Trump has NOTHING BUT 'generalized stances and rhetoric' "*[7]

The authors rather use common vocabulary than very domain-specific language. Furthermore, their language is more emotional. They tend to criticize opposing opinions and promote their personal view. This is different for opinion statements of political agents such as politicians or parties, which are found in transcriptions of speeches as well as discussions of political organs, to name only a few. In this kind of text we find another type of language that on the one hand derives from the domain knowledge an expertise of the authors or speakers. On the other hand, the intention of these agents slightly differs from non-domain experts: rather than criticizing opposing personal views, they advertise a particular ideology and try to convince the readers or audience.

To give an impression of the differing language used in the latter type of disputes, we show a brief example extracted from U.S. Congressional floor debates:

[5] http://www.createdebate.com/
[6] //www.createdebate.com/argument/newarg/85204/150563/735437
[7] //www.createdebate.com/argument/newarg/85282/150677/735945

> "*[...] we made a significant reduction to the domestic MOX plant because of the large unexpended prior year balances in that project caused by the continued liability dispute with the Russians [...]*"[8]

However, the number of approaches dealing with this kind of specific language produced by domain experts is limited.

Another task performed on this kind of data is to predict votes for or against legislations, party affiliation of speakers or speaker ideology. Examples are approaches by [Høyland et al., 2014], [Thomas et al., 2006] or [Diermeier et al., 2007], to name only a few. However, Hirst et al. [Hirst et al., 2014] suggest that supervised approaches such as the previous two might rather learn the difference between speakers of the governing party and speakers of the opposition than their ideology.

The previous approaches classify speakers or documents. In the last couple of years, opinion mining tasks evolved to more fine-grained challenges, such as detecting ideological bias on sentence level. Most recent development in the opinion analysis domain is the task of argumentation mining: instead of just classifying stances, arguments for the respective positions are determined.

So far, we described the development of opinion mining in the field of natural language processing. The analysis of political texts in social science has a different focus: in order to draw conclusions from political texts, inherent positions are usually placed on a scale between two opposing concepts. Common scales are *right-left* or *liberal-conservative*. Scholars of political science refer to such approaches that apply statistical methods to text as quantitative analysis. The field of quantitative political science has made a large progress in the last couple of years (cf. [Grimmer and Stewart, 2013]). Typical types of texts that are analyzed are party statements, speeches and discussions. Typical approaches for position analysis in political science scale texts based on word frequencies and co-occurrences as described by Grimmer and Stewart [Grimmer, 2010] or Quinn et al. [Quinn et al., 2010], to name only a few. There are many attempts to place texts on scales according to their positions, namely Laver and Garry, [Laver and Garry, 2000], Laver and Benoit [Laver et al., 2003], Keman [Keman, 2007] or Sim et al. [Sim et al., 2013], among others. A typical research question in the domain of political science is for example to measure the positions of parties using party election programs presented by Slapin and Proksch [Slapin and Proksch, 2008] which then allows for analyses about the change of party positions over time or in relation to certain events.

[8] `https://www.congress.gov/congressional-record/2005/5/24/house-section/article/H3780-2`

Summing up, approaches to detect positions in political texts are found in social sciences as well as in the field of computational linguistics. Both fields developed practically independent of each other. The most significant difference is that approaches of social science project positions on continuous scales, while the natural language approaches classify discrete stances. Moreover, the majority of the former assume one position per document, while in the latter field there are approaches on a more fine-grained level such as utterances or sentences. However, the main part of the natural language processing research focuses on documents containing colloquial rather than expert language.

1.3 Objectives

Our goal is to develop methods that enable scholars of political science to analyze actual research questions within their domain. We focus on documents created by political agents, i.e. domain experts, that are written in formal language and require some domain-knowledge to understand, such as party manifestos and transcriptions of political speeches as well as discussions.

Current methods in quantitative analysis regard a document as containing one single stance and place the whole document accordingly on a one-dimensional, continuous scale. We showed in our introductory example in Section 1.1 about the Euro-criticism shared by a left and a right party, that we consider opinions as being dependent on the discussed topic. The documents we intend to analyze cover various topics. Hence we develop techniques that access the opinions towards these topics as separate dimensions. This enables the comparisons of opinions across documents on topic level.

As mentioned before, there are already approaches in the domain of computational linguistics that determine opinions on topic rather than document level. The majority of these approaches focuses on on-line debates and other text authored by non-experts. In these texts, we find a colloquial and to a certain extent emotional language, and the objective of the authors is to criticize opposing ideas rather than propagate an own ideology. Approaches for that type of text are not directly applicable to the language in our documents. To give an example, they highly rely on agreement and disagreement cue phrases like *that's wrong* or *whoever believes this is an idiot*, which are hardly found in formal political texts. We will compare such approaches to our work in more detail in the following section and in the related work descriptions of the particular tasks we perform.

Another issue that we pointed out is that most approaches in computational linguistics classify stances towards a topic, i.e. they distinguish between a fixed set of distinct views on a topic, such as *pro choice* and *pro life* in the discussion about the legality of abortion. Analyses in political science usually rely on fine-grained positions that enable interpretation of distances between items. For this reason we only develop methods that do not classify opinions into fixed classes, but reflect their distances to each other and project fine-grained positions on a scale.

The methods in quantitative analysis commonly model a text as bag-of-words and operate on word frequencies and word co-occurrences. We enhance this model of a document by including linguistic information. To be more precise, we include knowledge about dialog and discourse structures as well as text coherence. Meeting transcriptions, for example, include various types of conversational contributions: On one side, one participant may address the others to elaborate on his or her opinion, on the other side, a discussant could ask a follow-up question to another one's contribution. We select contribution types according to their relevance for opinion analysis. Furthermore, we impose constraints based on principles of text coherence. For example, in a party manifesto we assume to find several subsequent sentences tackling the same item. Furthermore, there are sequences of topics that are more probable than others. Another improvement we contribute to determine topics in documents is the insertion of external knowledge. Thus we explore different techniques to access external descriptions of topics by linking the respective vocabulary to the vocabulary found in the documents, thereby giving meaning to particular words.

For the tasks we envisage, supervised methods would be appropriate. However, such methods require to be trained on text that is annotated. Fine-grained annotations in the field of political opinions on topic level do rarely exist and their creation requires political experts. As a consequence, we we focus on unsupervised or only weakly supervised approaches to determine topics and opinions.

1.4 Outline

In the previous sections, we introduced the task of opinion mining in natural language processing as well as in the field of political science and elaborated on the difference based on the independent development of both fields.

We will continue with an overview over related work from both fields in the following Chapter 2. We also include the few recent approaches that constitute a step towards connecting both areas, like it is also the goal of our work presented here. Chapters 3, 4 and 5 contain the heart of this work and describe the particular research questions we focus on and the approaches we developed to address them. Each of these chapters contains a section with related work that is particularly relevant for the respective research question and adds to the items presented in 2. Furthermore, each of the chapters contains an evaluation of our developed methods in order to analyze their possibilities and limitations. We finally draw conclusions of our work in Chapter 6.

Our contributions can be grouped regarding three aspects, which we will describe in the following paragraphs.

Computer Science Methods for Problems of Political Science. We investigate the applicability of state-of-the art models in computer science to the problem of scaling positions, such as various extensions of Topic Models, Machine Learning, Support Vector Regression and Markov Logic Networks.

Informed Approaches. We develop methods to interlink various information sources and include external knowledge resources for more informed approaches.

Knowledge about Text Structure. We employ methods of natural language processing and make use of knowledge about structure of texts, such as dialog structure or sequences of topical items.

These objectives are addressed in the following particular tasks:

1. Multi-dimensional position analysis of Party Manifestos
2. Fine-grained party manifesto classification and analysis of speeches in the U.S. Presidential Election Campaign
3. Scaling dove-hawk positions in the discussions of the Federal Open Market Committee of the U.S.

We have published parts of the research regarding these tasks in the following conference papers and journal articles: [Stuckenschmidt and Zirn, 2012], [Zirn and Stuckenschmidt, 2014] [Zirn, 2014], [Zirn et al., 2014], [Nanni et al., 2016], [Zirn et al., 2016].

In the remainder of this section, we will give more detailed descriptions of these tasks and relate them to our objectives.

1.4.1 Multi-Dimensional Position Analysis of Party Manifestos

Above described approaches like [Slapin and Proksch, 2008] place parties on a one-dimensional scale based on their manifesto programs. Our hypothesis is that the position of a party is rather a composition of several dimensions which refer to various political topics than only one dimension. Different parties might agree on one topic, but take different stands on another. We develop a method to capture all positions contained by a manifesto through all its topical dimensions. We further show how the topical positions can be compared by measuring their distance. This approach is completely unsupervised and does not require any training data. To detect the latent topics as dimensions, we apply topic models. In order to compare topics across documents, we need a way to take influence on the topic modeling process. To address this challenge, we compare different variants of topic models that allow for the use of seed words. These seed words serving as key words for political topics are automatically gathered from the definitions of ministries: Seed terms are extracted from ministry descriptions as well as ranked and selected with techniques of natural language processing. After applying the informed topic model, we analyze the distance between the positions underlying the topics. We compare different measures to best capture the positions. Finally, the success of the method is evaluated by predicting actual ministries from German coalition contracts.

1.4.2 Analyzing Topics and Positions in the U.S. Presidential Election Campaign

Prior to the U.S. Presidential Elections, potential candidates advertise themselves in public giving speeches. In those speeches, the candidates tackle manifold topics. However, the expressed positions towards the topics might be adapted to the particular audience and occasion of the speech and vary depending on the closeness to the elections. Answering this research question demands reliable methods to identify the topics in question and measure the underlying positions. Although this task shows similarities to the previously described one - measuring positions towards topics - it has different preconditions. The most significant difference is that in party manifestos, we can assume the positions as fixed. Here, however, scholars of political science are interested in potential changes within the speeches of the same person. Rather than measuring the distance between two (topical) positions, we intend to place each

position transported in a speech as a separate point on scale. The tool WordFish developed by Slapin and Proksch [Slapin and Proksch, 2008] is perfectly suitable for such a task, given that one input document contains exactly one position. In order to scale the positions towards all topics covered in a speech separately, we need to split the speech into documents containing a single topic. This again asks for topic classification. The most straightforward way is to classify units of a speech, i.e. sentences or paragraphs, with its topic. This is an easy task with supervised models using training data. However, we are not aware of an appropriate training set, and hand-coding such a resource set is time-consuming and costly. As an alternative, we propose a bootstrapping approach to create a training data set. Starting from a small set of party manifestos hand-coded on sentence level with their political topic by the Comparative Party Manifesto Project (cf. [Volkens et al., 2015]), we build a system to increase the training data with more, unlabeled manifestos. For this system, we first train state-of-the art classifiers on those manifestos. In a second step, we combine their predictions for sentences of unlabeled manifestos with knowledge about the structure of those documents and about typical sequences topics. A Markov Logic framework assess these features in mutual dependence to find the optimal global topic classifications for all sentences of an unlabeled manifesto. The total set of hand-coded and automatically labeled manifestos is then used in a state-of-the art machine learning model, which is then applied to the speeches to detect its topics.
After detecting the topics in the speeches and splitting them accordingly, we can use WordFish to scale the resulting topical sub-documents.

1.4.3 Scaling Dove-Hawk Positions in the Discussions of the Federal Open Market Committee of the U.S.

In the third task, we analyze central bank communication. In this domain, financial agents are associated with *dovish* or *hawkish* positions. *Hawks* are defined to be afraid of inflation and thus in favor of stability. Doves, on the other side, aim to stimulate economy by quantitative easing. Placing agents of the central bank on a *dove-hawk* scale is a difficult task even for experts. The textual communication released by central banks mostly contains complex, highly domain-specific language. Yet this communication has a huge influence on markets. Therefore, scholars in the financial domain rely on methods to estimate positions of central bank members that are reproducible and robust as basis for further analyses. We study this problem at the example of meeting transcriptions of the Federal Open Market Committee in the U.S. To our knowledge, there are no appropriate external knowledge sources to induce *dove-hawk* positions. We thus base our research on a supervised approach using a hand-coded dataset of

meeting transcriptions, on which we apply sophisticated regression techniques. The challenge lies in the appropriate modeling of the data: we improve upon a bag-of-words model by a deeper understanding of the dialog structure and modeling features in an according way.

2
Related Work

In Section 1.2, we introduced the development of opinion mining in the field of computational linguistics as well as quantitative position analysis in the area of political science. We now continue by giving an overview over research in these areas related to our work. We will focus on opinion mining in on-line debates, on formal political documents and on political speeches. Approaches that are closely related to one of the three particular tasks we address in chapters 3 to 5 will be discussed in the respective related work sections within these chapters.

2.1 Opinion Mining on On-line Debates

Approaches for opinion analysis in the political domain are manifold. A large share of research addresses stance classification in on-line debates.

Malouf and Mullen [Malouf and Mullen, 2008] present a supervised approach to classify the ideology of users based on their posts on a U.S. political discussion site. On this platform, every user has a profile and chooses a general label for his or her political affiliation. The authors first categorize these labels manually into *left*, *right* and *other* ideology. Posts on this platform are organized in threads and might contain citations of other posts. They contain political content, but are written in an informal, colloquial language. Malouf and Mullen merge all posts of a user into one document and apply a Naive Bayes classifier to learn the ideology of a user, which are *left*, *right* and *other*. The authors improve their approach by using network structures based on user interactions in the community. An analysis of the corpus shows that 77.5% of the citations refer to posts by users of the opposite ideology. Malouf and Mullen hypothesize that if users show similar citing patterns, they share the same political orientation. By clustering users according to their co-citations and classifying the resulting clusters with the Naive Bayes classifier, the performance is increased.

Further supervised approaches for stance classification in on-line-debates are presented by Anand et al.[Anand et al., 2011] and Hasan and Ng [Hasan and Ng, 2013], which build upon another.

The objective of Anand et al. is, as they say, "*to understand the discourse and dialogic structure*" of this type of debates. The discussion platform their data is obtained from offers the possibility of direct rebuttal posts. Besides some playful debates (e.g. *cats* vs. *dogs*), their data set mainly consists of ideological and serious topics such as *euthanasia* or *abortion*. Posts addressing the latter topics show more investment of the authors, which is indicated by several posts of an author per topic, and they contain more linguistic markers of dialogic interaction.

The authors tackle two tasks: first, they try to identify whether a post is a rebuttal of another one. Second, they predict the stance of a post. The main contribution of their approach is to make use of rebuttal-links between posts. According to the authors, crucial features for the final result are second person pronouns, negations and punctuation like question marks. The reason for this is the huge amount of rebuttals which contain interactions and references between the authors. This is supported by the comparison with Somasundaran and Wiebe [Somasundaran and Wiebe, 2009], who achieve much better results with a unigram-based approach on data that contain opinionated texts about the same topics, but no rebuttals. Furthermore, Anand et al. report that the classification is strongly dependent on the presence of sentiment terms: with absence of cues the performance decreases significantly.

These two insights lead to the assumption that those techniques might work well for user-generated content, but will not be applicable to the tasks we address. The above described methods are all designed to classify on-line debates with user generated content. The language in this type of documents is mostly colloquial and to a certain extent emotional. Malouf and Mullen further point out that the focus of contributions in public debates lies on criticizing opposing ideologies. In contrast, conversation of political agents - such as party members or financial committees - serves the purpose of convincing others of the agent's own views. We expect the data sets of our tasks, which are mainly party manifestos and meeting transcriptions of central bank communication, to be written in a highly formal language to the greatest extent devoid of sentiment terms.

Furthermore, the presented work basically distinguishes between two opposing stances only. Providing methods for scholars of political and economy science, we need to determine precise positions using the whole spectrum of a scale. All of the above mentioned approaches have in common that they regard the structure underlying their data, in particular the relations that hold within the contributions of the dialogs. In our bootstrapping approach to label party manifestos

on sentence level with their political topics that we mentioned above (see subsection 1.4.2), we make use of the structure and relations between sentences and topics in a similar way. Furthermore, our experiments on the transcriptions of the central bank discussions outlined in subsection 1.4.3 confirm the observation that when discussants directly address each other, the language differs from statements taking a standpoint without a direct addressee.

Hasan and Ng [Hasan and Ng, 2013] build upon the above described work of Anand et al.[Anand et al., 2011] to classify stances in on-line debates by refining and expanding the relation constraints of model. First, they encode *user interaction* constraints which assumes adjacent posts have opposite stances. They model this as a sequence labeling task with Conditional Random Fields.

In a second step, they introduce *inter-post* constraints, which apply to the particular discussion structure of this debating platform: When the same author writes comments in different threads which discuss the same topic, these posts are assigned the same ideology. Finally, the *ideology* constraint assumes that users of a particular ideology tend to comment on specific topic, which means there is a correlation between stances and certain topics. Users who comment on similar topics thus share the same ideology.

Hasan and Ng formulate the probabilities for stances they receive from the Conditional Random Field in combination with the two further constraints as an integer linear programming problem. We follow a very close idea in our model for party manifesto classification on sentence level (cf. subsection 1.4.2). We encode relations between sentences and categories as constraints, and formulate the classification for all sentences of a document at an optimization problem using Markov Logic.

Similar approaches that include relations among users and posts for the classifications of on-line debates are further presented, among others, by Lu et al. [Lu et al., 2012], which focus on agreement-disagreement constraints between discussants, or Ranade et al. [Ranade et al., 2013], who employ a Gradient Ascent method to analyze the "health" of the debate structure to improve stance prediction.

2.2 Opinion Mining on Formal Political Documents

As we pointed out, the above described techniques are tailored to detect stances in informal discussions and are mainly dependent on sentiment cues, which are absent in the most part of formal political documents. In Section 1.2, we already directed our attention to-

wards the field of quantitative analysis in political science. Approaches like Grimmer and Stewart [Grimmer and Stewart, 2013], Hillard et al. [Hillard et al., 2007] or Laver et al. [Laver et al., 2003] focus on formal political documents. Instead of relying on sentiment markers or dialog structure, stance is defined by the topics the document elaborates on. All above described research has in common that the whole document or agent is assigned a single position. In our work, we go beyond this and retrieve multi-dimensional positions which comprise positions towards the topics an agent speaks about.

In the following, we give an overview over approaches that apply methods of computer science and computational linguistics to formal political texts. This is consistent with the objectives of our work.

Thomas et al.[Thomas et al., 2006] as well as Burfoot [Burfoot, 2008] predict votes for or against a legislation from transcriptions of U.S. Congressional floor debates. In the first paper, the authors label speech segments as supporting or opposing a legislation. They introduce positive strength links. The first constraint defines that speech segments by the same speaker tend to have the same label. The second constraint states that a speaker who references another another one in agreement shares their label. They combine lexical features and the constraints with Support Vector Machines and differentiate the speakers that are *for* or *against* the legislation using Minimum Cuts.

Burfood builds upon this work refining the agreement relations. On the one hand, there is party agreement: speakers tend to vote in agreement with their party affiliation. Therefore, speakers of the same party have the same label. Speaker agreement on the other hand can be captured by measuring similarity between their texts: if speakers use similar words, they have the same position. Burfood concludes that they were not able to show the improvement of the results by introducing the refined constraints, as their integration in the Minimum Cut Framework was not able to exploit their potential.

An unsupervised approach to identify and cluster arguing expressions in on-line debates is addressed in Trabelsi and Zaıane [Trabelsi and Zaıane, 2014]. They assume an opinionated document is actually a text that covers several topics related to the discussed subject. On each topic, the author has a certain viewpoint, which is revealed by the words he uses to talk about the topic. The whole of these viewpoints define his opinion towards the subject. They model latent topics of the discussions and the implicit viewpoints they express in a Joint Topic Viewpoint model, which is an extension of Latent Dirichlet Analysis [Blei et al., 2003]. The author's way of modeling views and topics resembles our scenario of multi-dimension topic analysis on party manifestos (cf. subsection 1.4.1).

Another approach modeling latent topics and two opposing views is shown by Gottipati et al. [Gottipati et al., 2013]. They work with data from from Debatepedia[1], which collects supporting and opposing arguments from articles on political topics.

We partly share Trabelsi and Zaıane's as well as Gottipati et al.s' definition of opinionated documents for our multi-dimensional topic analysis of party manifestos. The difference between Gottipati et al.s' and our model is that we have the advantage of knowing the topics in advance: we assume that the latent topics are the broad political domains as defined by the ministries, that is why we can take influence on the topic creation process via seed words. Trabelsi and Zaıane, on the other hand, need to make assumptions about the amount of topics that are covered for each subject discussed in the on-line debates. Both Trabelsi and Zaıane and Gottipati et al. further have to specify the amount of possible viewpoints, which they define as distinct classes, namely *supporting* and an *opposing*. On our case, we do not need to model viewpoints: we are interested in the opinion of a party which is written down in a manifesto, and every manifesto thus contains exactly one view towards each topic. So in our case, it is enough to run a separate topic model on each manifesto. Please note that a topic model expects multiple documents as input, so we partition the manifestos into several subdocuments. As we know the tackled topics in advance due to the use of keywords, we are able to map the referring topics in the different manifestos to each other.

2.3 Opinion Mining on Political Speeches

An interesting approach presented by Sim et al. [Sim et al., 2013] is highly related to two of the tasks we tackle in this thesis. First, their task is similar to our in that they measure positions in U.S. Presidential election campaign speeches. We will compare their results to our findings. Second, the ideas behind their approach are related to some of ours.

Sim et al. start by defining a fixed set of 12 ideologies, which are interconnected by distance relations according to the assumption that some ideologies are closer related than others. The authors make the assumption that speakers might mingle with ideologies, yet changes between closer ideologies are more probable. A speech is considered as a sequence of clues for an ideology and filling terms, which are seen as "lags". Longer lags indicate a higher probability of an ideology change. Then, a Hidden Markov Model is used to model the transitions between the ideologies. To adjust the model parameters, Sim et al. use regression. The cues for the ideologies are learned from a labeled corpus of political writings. The authors analyze

[1] http://www.debatepedia.org

their results with a qualitative evaluation in which they discuss the conformity with political hypotheses which they refer to as “sanity checks”.

Sim et al. and our work have in common that we both follow hybrid approaches that combine external knowledge to define positions: in their case, they extract cues for ideologies from political writings, while we extract cues for political topics from ministry definitions. In our multi-dimensional position analysis, we measure positions towards topics relative to each other. Our approach could yet as well be used - as we show in our evaluation - to place a further document, like for example a coalition contract, on a scale spanned between two opposing views, which are defined by two party manifestos. In that case, we have a very similar setting to Sim et al.: we both have ideologies, defined by political writings in their case and election manifestos in our case. Further, we both are able to measure the position of new text - such as speeches or coalition contracts - in our case relative to party’s positions, in their case as proportions of ideologies. However, they do not distinguish between particular topics, while we measure positions on topic level, thus on multiple dimensions. Furthermore, the same applies to our approach to analyze the positions of the candidates of the Presidential election campaign: here, again, we analyze positions towards the topics discussed in the candidate’s speech.

3

Multi-dimensional Position Analysis of Party Manifestos

When coalitions are formed after governmental elections, scholars of political science are interested in analyzing the influence of the participating parties. With existing methods of quantitative analysis, it is possible to analyze the general dominance of the coalition partners. In this chapter, we present a method to measure the influence of parties in a more fine-grained way, namely with respect to particular topics. The results of coalition negotiations are written down in the coalition contract. Hence we consider this document as an appropriate basis for our analysis. We develop an approach that allows to compare the positions towards a topic recorded in the coalition contract to the positions described in the respective parties' election manifestos.

We extract the stances taken towards a topic from the coalition contract as well as from the election manifestos. To determine the topics, we access external resources rather than learning them in a supervised way. After the extraction of the views we develop methods to compare the positions of the coalition contract with the positions of the parties by measuring their similarities.

3.1 Motivation and Related Work

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 [Laver and Garry, 2000], where policy positions that have been identified automatically based on text can for example be used as input for simulations of party competition behavior [Laver and Sergenti, 2011]. The approach to text analysis adopted by researchers in this area is still strongly influenced by statistical methods

used to interpret survey data [Benoit et al., 2009]. 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. We point out 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. [Pappi et al., 2011]). In contrast, there is a good reason why party manifestos have been the primary subject of attempts to identify party positions [Volkens et al., 2011], 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 take these different topics into account.

In this task, 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 3.3, 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.

This chapter is organized as follows. In Section 3.2, 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 3.3 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 3.4. We conclude

the chapter with a discussion of the results and the implications for computer-aided content analysis in the social sciences.

3.2 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 political topic areas and that the documents to be analyzed actually contain information related to these areas. The second fundamental assumption is that the political 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.

3.2.1 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

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

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 [Hearst, 1997], 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 ([Hearst, 1997], 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.

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 [Schmid, 1999] 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.

3.2.2 Topic Creation

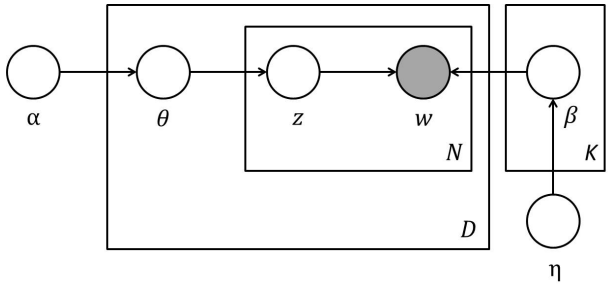


Figure 3.1: Graphical model for standard LDA

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. [Blei et al., 2003]. Figure 3.1 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 $\beta_i \sim$ Dir(η) for $i \in \{1, ... , K\}$, while Dir(η) being a Dirichlet prior with parameters η.
2. Then, for each document in D draw the topic proportions $\theta \sim$ Dir(α).
    a) For each word in the document, draw the per-word topic assignment $Z_{d,n} \sim$ Mult(θ_d) with Mult(θ_d) being a Multinomial mixture distribution depending on the topic proportions θ_d.
    b) For each word, draw the word $W_{d,n} \sim$ Mult($\beta_{z_{d,n}}$) with 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 [Casella and George, 1992].

Topic Creation with LabeledLDA

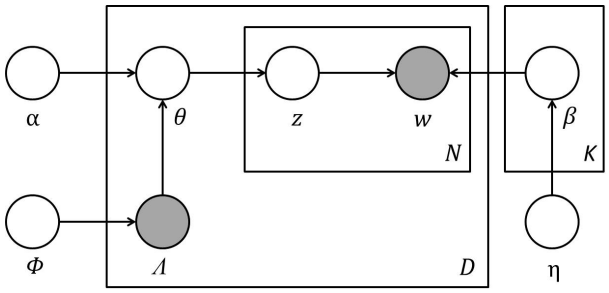


Figure 3.2: Graphical model for Labeled LDA

Labeled LDA by Ramage et al. [Ramage et al., 2009] 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 3.2 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 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 3.2.1, 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.

Topic Creation with LogicLDA

LogicLDA [Andrzejewski et al., 2011] by Andrzejewski et al. is an extension of LDA that offers the possibility to include first order knowledge.

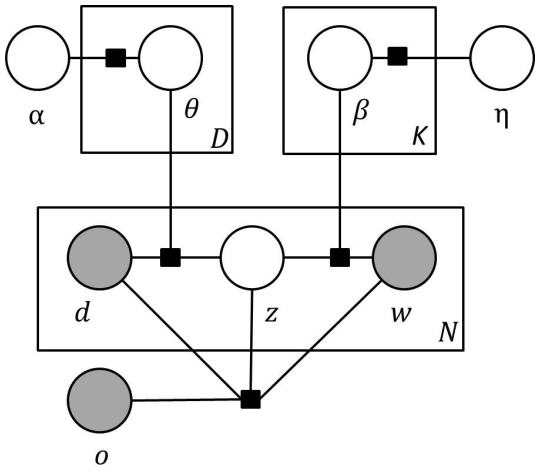


Figure 3.3: Factor graph for logicLDA

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 3.3 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:

$$W(i, Euro) \Rightarrow Z(i, 2) \tag{3.1}$$

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.

3.2.3 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 [Rosen-Zvi et al., 2004], 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.

3.3 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.

3.3.1 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 [Seher and Pappi, 2011] which will be described in more detail in Section 3.3.2. 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.

3.3.2 Data Sources

In our experiments, we analyze the data of the German national elections between 1990 and 2009. In all six elections (1990,1994,1998,2002,2005,2009), 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.

[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.

In [Seher and Pappi, 2011] 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 We map the topics of their scheme to the German ministries[5] 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')
- **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')

[5] German ministries as they were defined in 2011.

- **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.

We extract the following keywords from the ministry descriptions:

- 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 Außenhandel Bahn Energie Existenzgründung Hafen Infrastruktur Innovation Kreditwesen Banken Medienstandort Infrastruktur Staatskanzlei Ministerpräsident
- Security and Foreign Affairs
  Außenpolitik Sicherheitspolitik Entwicklungshilfe Militär Verteidigung Wehrdienst Wehrpflicht Zivildienst

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.

3.3.3 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 [Schmid, 1999], 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 3.3.4 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.

3.3.4 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.

3.3.5 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

[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.

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.

Analyzing 6 elections using 9 topical areas results in 54 single ministries to evaluate. Tables 3.2 to 3.7 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 3.2), 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 3.1 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.

German National Elections 1990 and 1994. Results for the elections of 1990 and 1994 are listed in tables 3.2 and 3.3. 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 oc-

[7] Regional Planning: FDP
[8] CDU: Traffic
[9] CDU: Defence

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 3.1: 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 3.2: Result of the Analysis of the German national elections 1990 using LogicLDA

curs only once, and the latter one does not appear at all except from being part of compounds, where it cannot be identified.

[10] CDU: Traffic
[11] CDU: Defence

	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 3.3: 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 3.4: Result of the Analysis of the German national elections 1998 using LogicLDA

German National Election 1998. For the elections of 1998 (Table 3.4), 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.

German National Election 2002. In Table 3.5 we can see that our method correctly predicted most of the ministries. The method made a mistake on the area of Economics and

[12] First 12 days: SPD
[13] SPD: Defence

	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 3.5: Result of the Analysis of the German national elections 2002 using LogicLDA

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.

German National Election 2005. For the 2005 election, we obtain a similar picture as shown in Table 3.6. 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 3.7. 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.

[14] Greens: Foreign Affairs, SPD: Defence
[15] CDU: Economics, SPD: Transport
[16] SPD: Foreign Affairs, CDU: Defence

	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 3.6: Result of the Analysis of the German national elections 2005 using 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 3.7: Result of the Analysis of the German national elections using 2009 LogicLDA

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 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

[17]FDP: Economics, CDU: Transport
[18]FDP: Foreign Affairs, CDU: Defence

coalition between CDU and SPD with nearly equally distributed ministries, our system only makes 2 false predictions.

Using all Content Words. In 3.2.1, 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.

3.3.6 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 3.8, 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 3.9), 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.

[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.

[20] Greens: Foreign Affairs, SPD: Defence

	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 3.8: 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 3.9: Result of the Analysis of the German national elections 2005 using Labeled LDA

The system using Labeled LDA made two mistakes for the coalition in 2009, shown in Table 3.10: Culture and Internal Affairs.

[21] CDU: Economics, SPD: Transport
[22] SPD: Foreign Affairs, CDU: Defence
[23] FDP: Economics, CDU: Transport
[24] FDP: Foreign Affairs, CDU: Defence

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

3.3.7 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 [Seher and Pappi, 2011], we generated a seed set for each of the following ministries:

- Federal Ministry of Defence ('Bundesministerium der Verteidigung')
- 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[25]. 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

[25] Namely, we used the following links:
http://de.wikipedia.org/wiki/Bundesministerium_der_Verteidigung
http://de.wikipedia.org/wiki/Auswärtiges_Amt
http://de.wikipedia.org/wiki/Bundesministerium_für_Bildung_und_Forschung
http://de.wikipedia.org/wiki/Bundesministerium_für_Ernährung,
_Landwirtschaft_und_Verbraucherschutz
http://de.wikipedia.org/wiki/Bundesministerium_für_Gesundheit_
(Deutschland)
http://de.wikipedia.org/wiki/Bundesministerium_des_Innern
http://de.wikipedia.org/wiki/Bundesministerium_für_Arbeit_und_Soziales
http://de.wikipedia.org/wiki/Bundesumweltministerium
http://de.wikipedia.org/wiki/Bundesministerium_für_Verkehr,_Bau_und_
Stadtentwicklung
We used the versions of the pages from October 5th, 2012.

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 3.11. 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.

3.4 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 program) of the 5 parties that were elected into the government in 2009, namely CDU, Greens (GRE), FDP, PDS, SPD. We calculate the similarity among

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

```
      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 3.4: Pairwise similarity for Economics (2009)

the parties pairwise for each topic. We show the topics Economics and Transport as well as Internal Affairs in figures 3.4 and 3.5.

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 3.4 and 3.5). 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.

```
      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 3.5: Pairwise similarity for Internal Affairs (2009)

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

Figure 3.4 shows the results for Economics. At a very first glance, the 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 3.5 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.

3.5 Conclusions

We presented an approach that demonstrates 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 topics can be relatively reliably related to predefined policy areas and be compared 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[26]. 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

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

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 actually 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 is an important research project providing sentence-based hand-coded party programs 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 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.

4

Analyzing Topics and Positions in the US Presidential Election Campaign

During the presidential election campaigns in the United States of America, the candidates give speeches in public to communicate their positions. There are different phases of the campaign: first, the candidates have to gain support of their party, and need to position themselves among other candidates of their party. After they have been chosen as candidate, they have to position themselves against the candidates of the competing party. Hence the candidates might adjust their positions respectively throughout the different phases.

In this chapter, we present a computational method that identifies topics in presidential election speeches and determines the candidates' positions towards each of these topics. We combine a topical classifier and WordFish [Slapin and Proksch, 2008], which is commonly used for quantitatively estimating candidate positions in political science analyses [Grimmer and Stewart, 2013]. We experiment with two different ways to create the topic classifier. On the one hand, we manually label a subset of the speech data. On the other hand, we explore the performance of training the same classifier on a semi-automatically annotated corpus, which we bootstrap from a small set of manually labeled party election manifestos.

In order to show why there is the need for a more fine-grained position analysis on topic level, we apply our method to speeches delivered in presidential election campaigns. In a qualitative analysis, we discuss how candidates' positions do not only vary with respect to topics, but how they also change in different phases of an election campaign. In other words, we show how some topic-based positions of some candidates change from pre-primaries, over primaries, to general election.

4.1 Motivation

The competition for votes in US elections provides an opportunity for candidates to communicate their positions. Evidence suggests that campaign statements are designed to inform voters of the types of policy a candidate will pursue in legislative [Ringquist and Dasse, 2004] and executive offices [Marschall and McKee, 2002].
Converging on a position, however, is a complicated process. Candidates must not only satisfy the interests of voters in the general election, but also win in primary elections where party identification is shared among candidates and support is ultimately won from informal organizations within the party [Masket, 2009].

Adequately capturing this process, namely the development of candidates' positions and reputations in campaigns, is a challenging empirical problem that relies on processing large amounts of political texts. Significant advancements in quantitative methods from the field of natural language processing (NLP) have enabled coarse-grained analyses of texts produced in presidential campaigns [Medzihorsky et al., 2014, Sim et al., 2013, Gross et al., 2013].

Democrats as well as Republicans have a tendency to either be vague on their position towards particular topics, such as immigration – especially in 2008 during the primaries. This is why we claim that an analysis of the candidates' positions has to be performed on topic level rather than general. In order to analyze the candidates' positions and their movements in a profound way, scholars of political science hence have a need for position analysis towards particular topics.

A common method in quantitative analysis to measure positions in text in a fine-grained way is WordFish [Slapin and Proksch, 2008]. As it measures positions not on topic, but on document level, it cannot be directly applied to our task. The speeches might address several topics and thus as many positions. In order to apply WordFish, we first have to detect the topics of the speech and split the text accordingly.

Therefore, our approach consists of two parts. In the first part, we recognize topics prevalent in the speech. In the second part, we split the speech transcriptions accordingly and perform position analysis on the resulting subdocuments.

Supervised text classification is widely used in computer science to classify topics in text. However, supervised methods require a training corpus labeled with the respective topics. In contrast to that, we aim at keeping manual input for our approaches as little as possible. In order to avoid hand coding of the topics in the speeches, we investigate the usage of a semi-

automatically labeled corpus from a different domain to train the topic classifier. For this, we take a small set of topically labeled party election programs and bootstrap a classifier in order to increase the training data. The process of the bootstrapping approach will be described in detail in Section 4.3. We will compare the performance of the resulting classifier to a classifier trained on a manually labeled subset of the presidential speeches.

4.2 Related Work

During the last decade, there has been a consistent growth in application of natural language processing methods in political science research [Grimmer and Stewart, 2013]. Here we cover the most relevant lines of work.

The detection of topics in political documents has been performed adopting unsupervised techniques such as latent semantic analyses (LSA) [Hofmann, 1999] and latent Dirichlet allocations (LDA) [Blei et al., 2003] as well as supervised adaptations like supervised LDA (sLDA) [Mcauliffe and Blei, 2008] and labeled LDA (lLDA) [Ramage et al., 2009]. For example, [Quinn et al., 2010] present a method that estimates a hierarchical structure of topics in political discussions, while [Balasubramanyan et al., 2012] describe an adaptation of sLDA for studying the topic-based polarization of debates in the U.S., and [Gottipati et al., 2013] explore the potential of Debatepedia for determining political topics and positions. [Nanni and Fabo, 2016] combine entity linking [Rao et al., 2013] and labeled LDA in order to overcome the most common limitation of unsupervised topic modeling techniques, namely the interpretability of the results.

Fully supervised approaches for topic detection have also been performed (see for example [Hillard et al., 2007]). However, as these solutions rely on expert knowledge for establishing a set of relevant topics in advance and on annotating a large set of training data, they generally are more time-consuming to build. In contrast, we show that for our approach a small set of annotated data is enough, and we explore the use of externally annotated training sources.

While there has been a long term interest in modeling ideological beliefs using automated systems (see for example [Abelson and Carroll, 1965]), only in recent years have we seen a growth of advanced computational techniques for performing the task. In 2003, Laver, Benoit and Garry presented Wordscores [Laver et al., 2003], a supervised approach that relies on a set of pre-defined reference texts to determine the position of political documents in an ideological space. Inspired by it, in 2008 Slapin and Proksch developed WordFish

[Slapin and Proksch, 2008], a completely unsupervised solution for scaling documents on a single dimension.

The techniques presented above analyze coarse-grained political positions on document level and do not fully exploit the potential of topic-based political scaling.

In the last decade, computer-based analysis of political campaigns has attracted the attention of journalists [Silver, 2012] and academics [Foot et al., 2003]. [Scharl and Weichselbraun, 2008] studied trends in political media coverage before and after the 2004 U.S. presidential election applying NLP methods. Recently, [Prabhakaran et al., 2014] studied the topic dynamics of interactions during the 2012 Republican presidential primary debates. Transcriptions of speeches have been employed by [Gross et al., 2013] adopting the method presented in [Sim et al., 2013] to study the US 2008 and 2012 campaigns and in particular to test the Etch-a-Sketch hypothesis. We will address the same hypothesis in our qualitative evaluation in subsection 4.5.

4.3 Bootstrapping a Topic-Labelled Training Set

In order to classify topics in the presidential candidates' speeches, we are looking for a labeled corpus to train a topic classifier on. The corpus needs to fulfill the following requirements: first, it needs to cover all political topics that are potentially addressed in political speeches. Second, the language in its text should not be politically biased towards one or another political camp, so either it has to contain various political positions or it has to be neutral.

There are only few newspapers that allow the usage of their data and contain topically labeled articles. These articles, however, might be biased towards a political camp. This can be balanced by using articles of different news papers, but on the one hand we did not succeed in finding enough freely available corpora, nor do the tagsets of the topical labels of various corpora match.

Nevertheless, we found a resource that seems appropriate. The Comparative Manifesto Project [Volkens et al., 2015] collects party election programs from several countries of the world, among them the United States, and labels the content with its topics on sentence level. It fulfills our requirement of covering all potential topics that might be addressed in the election campaign speeches, as it covers all topics mentioned in party manifestos. Furthermore, it contains the manifestos of all parties the candidates might belong to, and thus compounds all

possible views. The downside is that only a small subset of the manifestos collected by the project do provide the topic labels, which is not enough to sufficiently train a classifier on.

We propose a bootstrapping method to label the remaining part of the manifestos in order to generate a semi-automatically labeled training corpus for a supervised topic classifier.

4.3.1 Comparative Manifesto Project

The comparative manifesto project [Volkens et al., 2015], currently maintained by MAPOR[1], collects party election manifestos in order to compare political parties across countries. The database covers party programs of more than one thousand parties from over 50 countries in 5 continents and comprises data from the last 70 years. It is updated constantly.

The main purpose of the project is to allow for cross country comparison of political parties and their positions. Therefore, the statements of the manifestos are hand-coded with the political category they refer to. A detailed manual[2] gives instructions to the coders and explains the use of the 56 political categories, which are organized in 7 domains which refer to policy areas:

1. External Relations
2. Freedom and Democracy
3. Political System
4. Economy
5. Welfare and Quality of Life
6. Fabric of Society
7. Social Groups

To assure comparability and reproducibility across the coders, they first have to undergo an intense training with a supervisor followed by a test. They advise the coders to annotate keeping in mind the following advice [Werner et al., 2014]:

> *"The central question of manifesto coding is: What message is the party/presidential candidate trying to convey to voters? Which are the issues the party/presidential candidate regards as important?"*

[1] Manifesto Research on Political Representation
[2] https://manifestoproject.wzb.eu/down/papers/handbook_2014_version_5.pdf

With its coverage of manually labeled political statements across various years, countries and political parties with different positions, the Comparative Manifesto Database constitutes a valuable resource for structured text describing political topics and stances.

4.3.2 Bootstrapping the Manifesto Classifier

Our bootstrapping approach to label the whole set of election manifestos consists of three components. The first component is a local sentence-level classifier that predicts a topic based on the information extracted from the sentence. The second part consists of sentence-pair classifiers that learn whether two consecutive sentences have the same or different labels. For the classifiers in the first and second component, we use a Support Vector Machine.

The third component combines the results of the first two components and adds additional knowledge about topic labeling in manifestos to find the optimal global classification of all sentences of a manifesto in mutual dependence. As a framework for the optimization component, we use Markov Logic Networks.

In the following, we will briefly describe Support Vector Machines and Markov Logic Networks, followed by a detailed description of the three components.

Support Vector Machines Among the variety of machine learning algorithms, Support Vector Machines (SVM) have been shown to work especially well for text classification (cf. [Joachims, 2002], among others). However, the use of deep learning algorithms recently gained currency in the area of natural language processing. The reason why we still choose Support Vector Machines is that they are more suitable for relatively small corpora, as is the case in our scenario (cf. [Caruana and Niculescu-Mizil, 2006], [Raschka, 2016]).

Support Vector Machines [Vapnik, 1995] search for hyperplanes that best separate the data into classes. If the data is not linearly separable, they transform the data into a higher dimension with a kernel function. As the algorithm tries to find the hyperplane that best separates the data, which is the one with the largest margin to the data, it is also referred to as maximum margin classifier. The support vectors (SVs) are the data points that are used to define the plane; they lie either on or within the margin.

Assume we have a set of m data points assigned to two classes:

$$\{(\mathbf{x}_i, y_i) | i = 1, \ldots, m; \mathbf{x} \in \mathbb{R}; y \in \{-1, 1\}\}. \tag{4.1}$$

We are looking for the hyperplane separating the data classes, which is defined by a normale vector $\mathbf{w}$ and a bias b, as given in equation 4.2.

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \tag{4.2}$$

If the data is not linearly separable, it needs to be transformed from its dimension d_1 to a higher dimension d_2:

$$\phi \colon \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}, \mathbf{x} \mapsto \phi(\mathbf{x}) \tag{4.3}$$

This can be computed with a Kernel function $K(\mathbf{x}_i, \mathbf{x})$.

The resulting classifier for an unseen data point $\mathbf{x}$ is thus defined in equation 4.4, with α_i being the parameter that describes the mapping of a data point into the higher dimension space.

$$f(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b) = \operatorname{sgn}\left(\sum_{i \in SVs} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right) \tag{4.4}$$

There are various choices for Kernel functions. In our experiments, we will make use of two commonly used functions: the linear kernel function given in equation 4.5 and the radial basis function (RBF), originally proposed by [Broomhead and Lowe, 1988], given in equation 4.6.

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{4.5}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0 \tag{4.6}$$

Markov Logic Networks Markov logic [Richardson and Domingos, 2006] can be interpreted as a template language combining first-order logic with maximum entropy models. The user can specify types of data and encode prior knowledge about the information used in the classification scenario, and it searches the most probable possible world given the evidence.

A Markov network $\mathcal{M}$ is an undirected graph whose nodes represent a set of random variables $\mathbf{X} = \{X_1, ..., X_n\}$ and whose edges model direct probabilistic interactions between adjacent

nodes. More formally, a distribution P is a log-linear model over a Markov network $\mathcal{M}$ if it is associated with:

- a set of features $\{f_1(D_1), ..., f_k(D_k)\}$, where each D_i is a clique in $\mathcal{M}$ and each f_i is a function from D_i to $\mathbb{R}$,
- a set of real-valued weights $w_1, ..., w_k$, such that

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{k} w_i f_i(D_i)\right),$$

where Z is a normalization constant.

A Markov logic network is a set of pairs (F_i, w_i) where each F_i is a first-order formula and each w_i a real-valued weight associated with F_i. With a finite set of constants C it defines a log-linear model over possible worlds $\{\mathbf{x}\}$ where each variable X_j corresponds to a ground atom and each feature f_i is the number of true groundings (instantiations) of F_i with respect to C in possible world $\mathbf{x}$. Possible worlds are truth assignments to all ground atoms with respect to the set of constants C. We explicitly distinguish between weighted formulas and *deterministic* formulas, that is, formulas that always have to hold.

Given a set of first-order formulas and a set of ground atoms, we wish to find the formulas' maximum a posteriori (MAP) weights, that is, the weights that maximize the log-likelihood of the hidden variables given the evidence.

Component 1: Local Sentence-Level Classification. The local sentence-level classifier makes predictions taking into account only the information from the sentence itself. We classify manifesto sentences into the seven topical classes that correspond to the domains in the Comparative Manifesto Project categorization scheme. We use the following set of lexical and numerical features:

1. The bag-of-words term-vector of the sentence;
2. The topic of the preceding sentence;
3. The semantic similarity between the current and preceding sentence. The semantic similarity between the sentences is computed by greedily aligning most similar words from the two sentences. Let P be the set of greedily aligned pairs (w_1, w_2) of words (where

w_1 is from the first sentences, and w_2 is from the second sentence). The raw semantic similarity between the sentences is then given as:

$$sim(s_1, s_2) = \sum_{\substack{(w_1,w_2)\in P \\ w_i \in s_i}} \cos(v_{w_1}, v_{w_2})$$

where v_w is the semantic embedding vector of the word w. We used the pre-trained set of 200-dimensional GloVe embeddings[3] [Pennington et al., 2014] to compute the raw semantic similarity score. Because the similarity given by the above-mentioned formula depends on the length of the sentences, we normalized the score by the length of each of the two sentences and provided the harmonic mean of the two normalized scores as the final value of the semantic similarity feature;

4. For each topic class we computed a numeric feature indicating the level of relative relevance of the sentence words for that class. We computed the relative frequencies of lemmas in sentences belonging to each of the topic classes on the train set. For example, if the word *"social"* appeared n times in all sentences of the train set labeled with the topical class *"social security and welfare"* and these sentences together contain N words, then $\frac{n}{N}$ is the relative relevance of the word *"social"* for the *"social security and welfare"* topic. Let $rr(w, c)$ be the relative relevance of the word w for the topical class c. The relevance score of the sentences s for the class c is then computed as follows:

$$rs(s, c) = \frac{\sum_{w\in s} rr(w, c)}{|s|}$$

where $|s|$ is the total number of words in the sentence s. Only content words (nouns, verbs, adjectives, and adverbs) were considered for the computation of class-relevance scores of sentences. For each sentence, one relevance score (i.e., one feature) is computed for each of the topical classes.

Component 2: Sentence-pair Classification. The classifiers for the detection of topic shifts between adjacent sentences are binary classifiers that predict whether the two given adjacent sentences are on the same topic or not. We used the following set of features for the detection of local topic shifts:

1. Bag-of-words term-vector of the first sentence (f^1);

[3]Built from Wikipedia 2014 and Gigaword 5, obtainable from `http://nlp.stanford.edu/data/glove.6B.zip`

2. Bag-of-words term-vector of the second sentence (f^2);
3. Length (in no. words) of the first sentence (f^3);
4. Length (in no. words) of the second sentence (f^4);
5. Semantic similarity between the two sentences (f^5, cf. Section 4.3.2);
6. Ngram overlap between the two sentences (f^6) – this is merely the number of content words shared by the two sentences. The number of shared words is normalized by the length of the sentences and the two normalized scores are averaged.

Considering the large size of the feature space, caused by the lexical BoW features f^1 and f^2, we first attempted to feed all features to a single linear SVM classifier. However, we observed that adding the numerical features (f^3–f^6) yielded no improvements in classification performance over using only the bag-of-words vectors of the sentences. We then fed only the numerical features to the SVM classifier with a non-linear RBF kernel and obtained similar cross-validation performance on the train set as when using the linear SVM classifier with only the bag-of-words features. Considering that the two classifiers, (1) the linear SVM using the bag-of-words features and (2) the RBF SVM with four numeric features, address the same task with completely disjoint sets of features, we decided to incorporate local (i.e., sentence-level) predictions of both classifiers (together with the predictions of the topic classifier) in an inference framework for finding globally optimal topic predictions for a sequence of sentences.

Component 3: Global Optimization. In addition to the information we gain from the content of the sentences, we make use of knowledge we have about the sequence and distribution of topics in manifestos. One salient observation is the fact that topics are usually tackled in several consecutive sentences. So we assume that successive sentences tend to share the same label. If we take this observation a step further, we can measure the frequency of transitions of one category to another in two successive sentences from the training corpus. This does not only help us to decide whether two consecutive sentences share the same label, but gives us an estimate for probable sequences of categories. So for this component, we combine such knowledge with the information gained from the first two components in order to classify all sentences of a manifesto in mutual dependence. Above all, the goal is the best classification for the manifesto as a whole, satisfying all constraints. We use Markov Logic Networks to search for the optimal solution to this classification problem.

We model each sentence of the manifesto as a constant $s \in S$. In the same manner, categories 1-7 are represented as constants. First, we specify that each sentence s is mapped to exactly one category c as a deterministic formula:

$$\forall s, c : |c| map(s, c) = 1$$

As we intend to predict the correct mappings, $map(s, c)$ is our hidden predicate. We introduce the predicate $next(s1, s2)$ stating that sentence $s1$ is followed by $s2$ to model the sequences of sentences in a manifesto. This allows us to encode our observation that subsequent sentences share the same category:

$$\forall s, c : next(s_1, s_2) \wedge map(s_1, c) \Rightarrow map(s_2, c)$$

In contrast to the first formula, this one can be violated with a certain penalty, thus the formula is given a weight. Estimations about the transition between two particular categories are modeled alike by replacing c by particular variables c_1, c_2.

The predictions from the local sentence classifiers are modeled with the predicate $localConf(s, c, conf)$, where $conf$ represents the confidence for sentence s to be mapped to a particular category c. We use this confidence as the weight for the corresponding formula:

$$\forall s, c : localConf(s, c, conf) \wedge map(s, c)$$

Each of the sentence-pair classifiers is modeled (separately) via a predicate called $flip$.

$$\begin{aligned} \forall s, c : & shift(s_1, s_2, conf) \wedge map(s_1, c) \\ & \Rightarrow \neg map(s_2, c) \end{aligned}$$

It expresses the confidence of a sentence pair belonging to two different categories: the label of the first sentence is "flipped" if the formula is true, i.e. if the confidence by the classifier (included as the weight for the formula) is high enough.

4.3.3 Evaluation

In our experiments we used six U.S. manifestos (Republican and Democrat manifestos from 2004, 2008, and 2012 elections). In all experiments, we perform folded cross-validation and report the micro-averaged results over folds.

Topic	P	R	F_1
External Rel.	83.7	86.6	85.1
Freedom & Dem.	68.0	59.9	63.7
Pol. system	69.7	65.7	67.6
Economy	73.9	77.4	75.6
Welfare & QoL	72.8	72.8	72.8
Fabric of Soc.	74.8	76.0	75.4
Soc. Groups	71.2	67.9	69.5
Micro-avg.	74.9	74.9	74.9

Table 4.1: Local topic classification, 10-fold CV (%).

Model	P	R	F_1
Linear, bow feat.	56.6	54.6	55.6
RBF, num. feat.	98.5	27.4	42.9

Table 4.2: Topic-shift classification, 10-fold CV (%).

Results of Component 1: Topic Classification. Table 4.1 shows the results of the local topic classifier obtained via the 10-fold cross validation. The classification performance is best for *External relations* (more easily recognizable due to re-occurring country names) and worst for *Freedom and democracy* (as lexical clues typical for this class tend to frequently appear in sentences of other topics as well).

Results of Component 2: Sentence-pair Classification. The performance of the two topic-shift classifiers is given in Table 4.2. These results indicate that detecting topic shifts is a more difficult task than predicting the topics of individual sentences. This is expected, as correctly identifying the topic shift logically amounts to correctly predicting topics for two consecutive sentences.

Results of Component 3: Global Optimization. The predictions of local classifiers are combined with the topic distribution information in a Markov Logic Network (MLN). We use RockIt [Noessner et al., 2013] as the MLN engine.

Setting	MaP	MaR	MaF_1	MiF_1
L	73.5	72.3	72.8	74.9
L,T	80.7	73.1	75.2	78.3
L,S	78.3	74.5	75.9	78.3
L,S,P_{bow}	74.2	73.0	73.6	75.6
L,S,P_{num}	78.6	76.7	**77.5**	**79.3**
L,S,P_{bow},P_{num}	74.4	73.2	73.7	75.8

Table 4.3: Global classification (validation-set): MaP/MaR/MaF_1 = Macro precision/recall/F_1-measure; MiF_1 = micro F_1-measure.

To evaluate the impact of each component, we start the experiments with a reduced set of formulas and incrementally add more constraints. As a baseline, we simply use the predictions by the local classifier (setting L). In the second setting, we encode rules for transitions (setting T) between particular topics. This is directly compared to a simpler setting S where we just assign consecutive sentences the same label instead of adding an own transition rule for every possible sequence of topics. The results of these combinations applied to the validation set are shown in the first part of the Table 4.3. Adding the information about consecutive sentences and transitions improves over the local classifier performance for by points, reaching 78.3%.

As precision and recall are more balanced for setting S and it needs significantly less rules, we prefer it over setting T for the following experiments. We now employ the predictions of the topic-shift classifiers: P_{bow} are the predictions of the linear SVM model with BOW features and P_{num} denotes the predictions of the non-linear SVM using numerical features. We first test each one separately, then both together (setting L + N). The lower part of table 4.3 shows the results. The best performance of 79.3% F_1 score is obtained for the model using predictions P_{num}. The combination of both sentence pair classifiers drops performance, which is not surprising due to the performance of classifier P_{bow}.

4.4 Identification of Topics in Speeches

As described in the introduction of this chapter, our first step to analyze presidential election campaign speeches is to identify the topics that are discussed in the speeches with a supervised machine learning algorithm. Like for the classification problem in Section 4.3, we decide to

use Support Vector Machines, as they perform well on such text classification tasks. The features we use are the same as for the local sentence level classifier in the previous section:

1. The bag-of-words term-vector of the sentence.
2. The topic of the preceding sentence.
3. The semantic similarity between the current and preceding sentence.
4. The relevance score of the sentence for that class.

We choose the classification scheme introduced by the Comparative Manifesto Project [Volkens et al., 2011] (for a description, see subsection 4.3.1) not only because we intend to use the party manifestos as a training corpus, but also because they categorize the whole spectrum of political topics that might appear in the candidates' speeches in a reasonable granularity. We assume that those domains, which are used to capture all topics tackled in party election programs, correspond to major coarse-grained topics of interest in electoral speeches.

We train this local classifier on two different datasets and compare their performance on a gold standard of speeches manually annotated for their topics on paragraph level.

Training Set: Manifestos. We train the classifier on party manifesto programs labeled on sentence level as described in the previous section. The advantage of such a domain transfer approach is the fact that we do not need any manual topic annotations on speeches. The downside is, however, that the language of manifestos might differ from the language used in speeches. In the next section, we quantify the drop in performance due to the domain change.

Training Set: Annotated Speeches. We manually annotated a small part of the presidential election campaign speeches on paragraph level with their categories. We train the above described system on this data and, in the next section, report the results. We explore whether investing human resources for annotating speeches pays off with more accurate classification results.

We decide to classify the text of the speeches at paragraph level. Whole paragraphs most often belong to the same topic, because politicians tend to express their arguments coherently. So please note that when using the manifesto training set, we train on sentences, but apply the classifier on paragraphs.

Model	F_1
Manifesto Corpus	36.2
Standard SVM	71.2
Speech Corpus	78.6

Table 4.4: Topic classification performance, micro F1-score, 10-fold CV (in %).

4.4.1 Evaluation of Topic Classification

We assess the correctness of the topic classification on a manually-labeled evaluation dataset of speeches. We compare three different settings to classify the topics in the speeches. This comparison will reveal if it is necessary, for political scientists interested in conducting a similar analysis, to train a classifier in an in-domain dataset in order to detect topics in political speeches, or whether an existing data set from a similar domain might lead to an acceptable performance, too.

- **Baseline**. As a baseline, we apply a Support Vector Machine (SVM) using a simple bag-of-words features on the gold standard performing 10-fold cross validation.
- **SpeechCorpus**. We train and apply the classifier described at the beginning of Section 4.4 using 10-fold cross validation directly to the manually annotated gold standard.
- **ManifestoCorpus**. We train the same classifier on the set of the semi-automatically labeled manifestos, and apply it to the manually labeled gold standard.

Gold Standard Annotation

We asked two scholars of political science to annotate a subset of 10 speeches from the US presidential election campaigns of 2008, 2012 and 2016. The set comprises samples of seven candidates. Our annotators labeled each of the 779 selected paragraphs with one of the 7 topical classes listed in subsection 4.4. The inter-annotator agreement across the seven topical classes is $\kappa = 0.55$, which is only moderate and thus confirms the difficulty of the task.

Results of Topic Classification

The results of the three evaluation settings are shown in Table 4.4.

As it is evident from Table 4.4, the baseline performs quite well with an F_1-score of around 71% , re-confirming the already well-known efficiency of the simple bag-of-words-based supervised topic classification models. The drop in performance caused by the domain adaptation (i.e., the low performance of the model trained on manifestos) indicates that, even if the topics discussed in electoral manifestos and in political campaigns are the same, the language in which they are convened seems to be significantly different. Finally, the best performance is achieved by the local topic-classifier trained on a small set of manually labeled speeches. The fact that this model drastically outperforms the model trained on manifestos shows that having little of in-domain annotations (i.e., annotated speeches) matters more than having a lot of annotations on out-of-domain texts (i.e., manifestos). For this reason, we will use the classifier trained on the speech corpus as foundation for our topic-specific position analysis performed in the following section.

4.5 Qualitative Analysis of Topic-Specific Positions

To determine the positions of politicians based on their speeches, we apply the above described classifier trained on the set of manually annotated speeches and employ WordFish [Slapin and Proksch, 2008].

Our goal is to determine fine-grained positions towards the topics contained in the speeches instead of the overall position of the whole speech. We therefore first apply our topic classifier on the speeches on paragraph level in order to identify the topics within a speech. We divide a speech into subdocuments by merging the paragraphs according to their most prevalent topics, so that we receive one subdocument per topic. Finally, we apply WordFish on the topical documents.

To our knowledge, there is no straightforward way to evaluate the correctness of the derived topic-wise positions in a manner common for computational linguistic tasks, such as comparing them to a gold standard. We are not aware of any appropriate data set, nor do we believe that humans are capable of capturing the degree of position shifts particularly without being subjective or using backgound knowledge or knowledge from further resources. Which is, in actual fact, the reason why political scientists long for quantitative methods in the first place.

A similar type of evaluation is also performed by [Sim et al., 2013]. They analyze the content of election campaign speeches in relation to pre-defined ideologies (not on topic, but on speech

level though). They evaluate their results by investigating the inherence to strong and moderate hypotheses of the output of their method.

WordFish WordFish, developed by Slapin and Proksch is is widely used for such tasks in political science research [Grimmer and Stewart, 2013]. This method is designed to take documents as input and estimates their positions on a one-dimensional scale based on word frequencies. It assumes that words are distributed according to a Poisson distribution:

$$y_{i,j} \sim Poisson(\lambda_{ij})$$

thus the model is described by the following equation:

$$y_{i,j} = exp(\alpha_i + \psi_j + \beta_j * \omega_i)$$

where y_{ij} is the frequency of word j in a document by speaker i. α is a set of speaker fixed effects, ψ is a set of word fixed effects, and β is an estimate of a word specific weight capturing the importance of word j in discriminating between positions, and ω is the estimate of speaker i's position (cf. [Slapin and Proksch, 2008]). These parameters are then estimated with an expectation maximization algorithm.

The Three Phases of an Election Campaign. Election campaigns are a long and complex process that represents the essence of contemporary democracies. In the United States, the practice of selecting candidates for the presidential elections spans more than a year, being a major focus of American and international media. More specifically, political scientists distinguish between three major phases in the presidential race: a) the pre-primaries, when politicians announce their candidacy for president and begin to establish their positions; b) the primaries: when candidates sharpen their profile in order to win the support of the party; and c) the presidential elections: when party nominees have to satisfy the interests of a spectrum of voters as large as possible.

After collecting speeches made by the most prominent Republican and Democrat candidates during the three most recent general elections (i.e. 2008, 2012, 2016), we divide them into three temporal groups, namely: before primaries (i.e. before the 1st of January of the election year), primaries (between January and June of the election year) and elections (after June of the election year). Using the above described classification model, we topically label all

of the collected political speeches at paragraph level. Next, we group together all topical subdocuments throughout the same time period (e.g. all text from all Barack Obama's primary campaign speeches labeled with topic *External Relations*). Finally, we run WordFish on the collection of temporally and topically divided speeches.

We use the R-implementation of WordFish included in the Austin package [Lowe, 2015]. The preprocessing - stemming and stopword filtering - is performed with the R-package jfreq [Lowe, 2011].

In order to understand the usefulness of our fine-grained analysis (i.e., the combination of the two dimensions – time and topic), we compare the output with another more coarse-grained study which we present first. In both studies, we run the analyses per campaign, i.e. for every election year we perform a separate analysis.

Temporal Dimension. In the first study we consider only the temporal dimension, i.e. the three phases of pre-primaries, primaries and election. This means for every candidate, we create three documents: one for each phase. We then run WordFish on each of the phases separately.

Temporal and Topical Dimension. Finally, we consider time as well as the topical dimension. Thus we merge all paragraphs regarding a particular topic of a candidate per phase, resulting in three documents per candidate. We perform the analyses for each of the topics separately.

4.5.1 Results

In the following, we describe and discuss the results of applying our approach on data of the presidential election campaign in the year 2008. We included the speeches of Hillary Clinton, John McCain, and Barack Obama. Hillary Clinton and Barack Obama were candidates of the Democrats, John McCain ran for the Republicans. Obama and McCain won the candidacy of their parties, consequently Clinton does not appear in the election phase in our study.

General Positions. Figure 4.1 shows the results for the first study, in which we partition the data into three temporal phases: pre-primaries, primaries and elections. We find the results conform with our expectations: during all three phases, we find the Republican McCain on one end of the scale, the Democrat Obama on the other end. The scale might be interpreted as

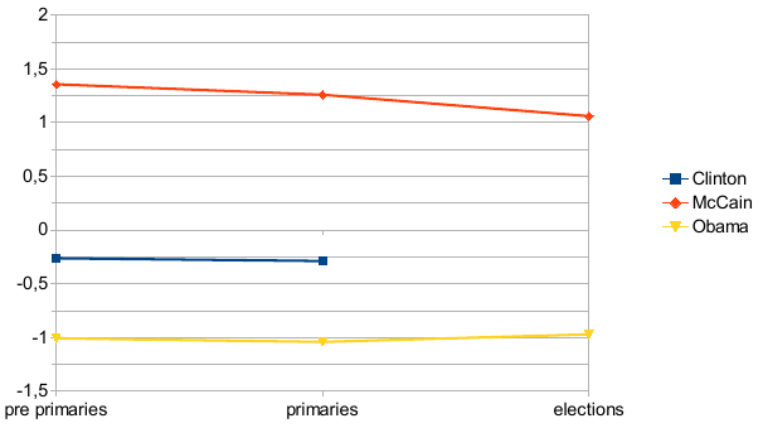


Figure 4.1: Temporally partitioned set of speeches by Hillary Clinton, John McCain and Barack Obama in 2008 scaled with WordFish.

a left/right scale, where positive values indicate a right-orientated political ideology, negative values a left-orientated ideology. This left-right dimension is also found by other studies applying WordFish on their data, for instance in the original paper by Slapin and Proksch [Slapin and Proksch, 2008], but also by further scholars like Klüver [Klüver, 2009] or Seher and Pappi [Pappi and Seher, 2009]. The fact that we find Hillary Clinton in between McCain and Obama might be surprising, as she is considered by many as being left of Obama. If we have a look at Clinton's and Obama's voting record in Senate shown in Figure 4.2, we can see find Clinton's position right of Obama's. Her general position might also be explained by her position on particular topics, such as *Welfare and Quality of Life*, which we will describe below in when discussing the analysis on topic level.

Temporal Dimension. In the previous paragraph we examined the validity of the positions scaled by WordFish. The second observation that we make in Figure 4.1 is the movement of Obama's and McCain's positions towards the center inbetween the primaries and the election phase. This behavior is typical: in the first phases of the campaign, the candidate have to earn the support of their party and compete with other party members, thus they have to clearly differentiate their positions from those of their competitors. After havin been confirmed as official candidate by their party, they have win the people's favor over the candidate of the opposing party. To appeal to more electors, they place themselves in a more moderate direction.

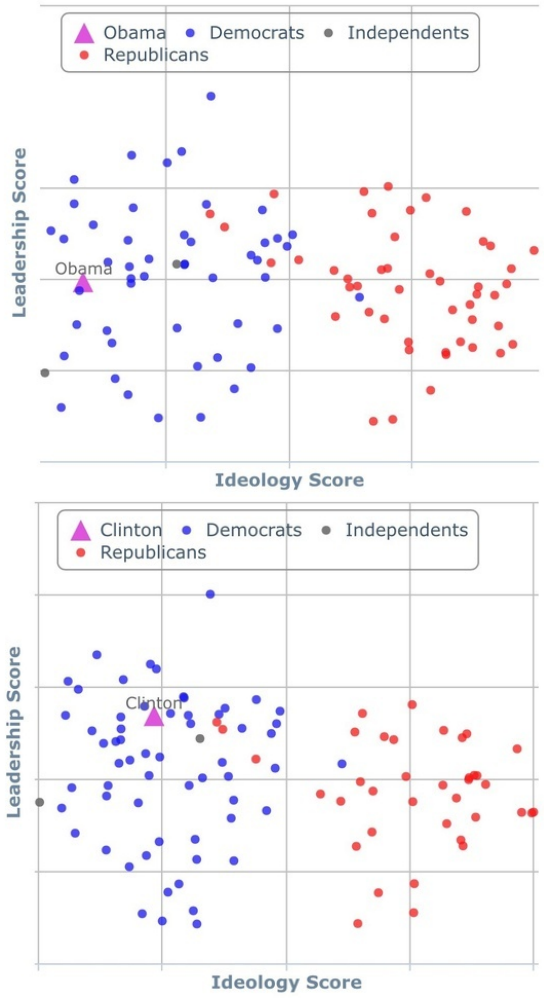


Figure 4.2: Positions of members of the Senate based on their voting record and sponsored/co-sponsored bills.[4], produced by GovTrack `https://www.govtrack.us/`

This movement is referred to as the *Edge-a-Sketch* assumption. This phrase was shaped by Eric Fehrnstrom, senior campaign advisor of Mitt Romney, another candidate in previous presidential elections. Fehrnstrom openly stated his opinion about positioning of the candidates for the last phase of the election campaign starting in fall:

> *"I think you hit a reset button for the fall campaign. Everything changes. It's almost like an Etch A Sketch. You can kind of shake it up and restart all over again."*
>
> Eric Fehrnstrom, CNN interview on March 21, 2012

"Etch A Sketch" refers to the magnetic drawing board, actually a kid's toy, where a complete picture is entirely erased by simply shaking the board.

The Etch-a-Sketch assumption was also confirmed for the same data set by Sim et al. [Sim et al., 2013], whose work we previously mentioned in section 2.3. The authors determine the proportion of a set of 12 ideologies in presidential election campaign speeches. They test the Etch-a-Sketch hypothesis for Obama and McCain. We show their results in Figure 4.3. The authors sort the set of ideologies according to their left-right connotation, and the diagrams list the proportion of these ideologies within the candidates' speeches in the respective election phase. Please note that the diagram for Barack Obama also shows the further development up to 2012. Only the the left part of the diagram until the year 2008 is relevant for comparison with McCaine's ideology proportions and with our diagram in Figure 4.1.

The findings in these diagrams support the Etch-a-Sketch assumption and are consistent with our findings: Obama as well as McCain first position themselves at the outer parts of the spectrum, before their speeches show larger proportions of centered ideologies.

We interpret the matching result of their results and ours as well as the compliance of our output to general observations as confirmation that our method produces valid results.

Temporal and Topical Dimension. In the following, we have a look at the results of our method for topic-specific position analysis. The results are shown in figures 4.4, 4.5 and 4.6. We present and discuss the results for the topics *Economy*, *External Relations* and *Welfare and Quality of Life*.

Topic: External Relation. Figure 4.4 shows the results for the topic *External Relation*. Like in the analysis of the entire speeches, we can observe the Etch-a-Sketch movement towards the middle after the primaries, so we can see that for this topic, the candidates are placed as expected. Here, we find Hillary Clinton close to Obama, which is conform with the general expectation, as they belong to the same party.

The most dominant issue in the 2008 presidential election campaign was the discussion about the involvement in the Iraq War, that had been supported by the current president George W. Bush. John McCain supported the war while Barack Obama opposed it; yet we assume that also McCain's position has a rather left tendency compared to the acting president Bush's stance. By that time, the public opinion had turned to a mostly negative attitude towards

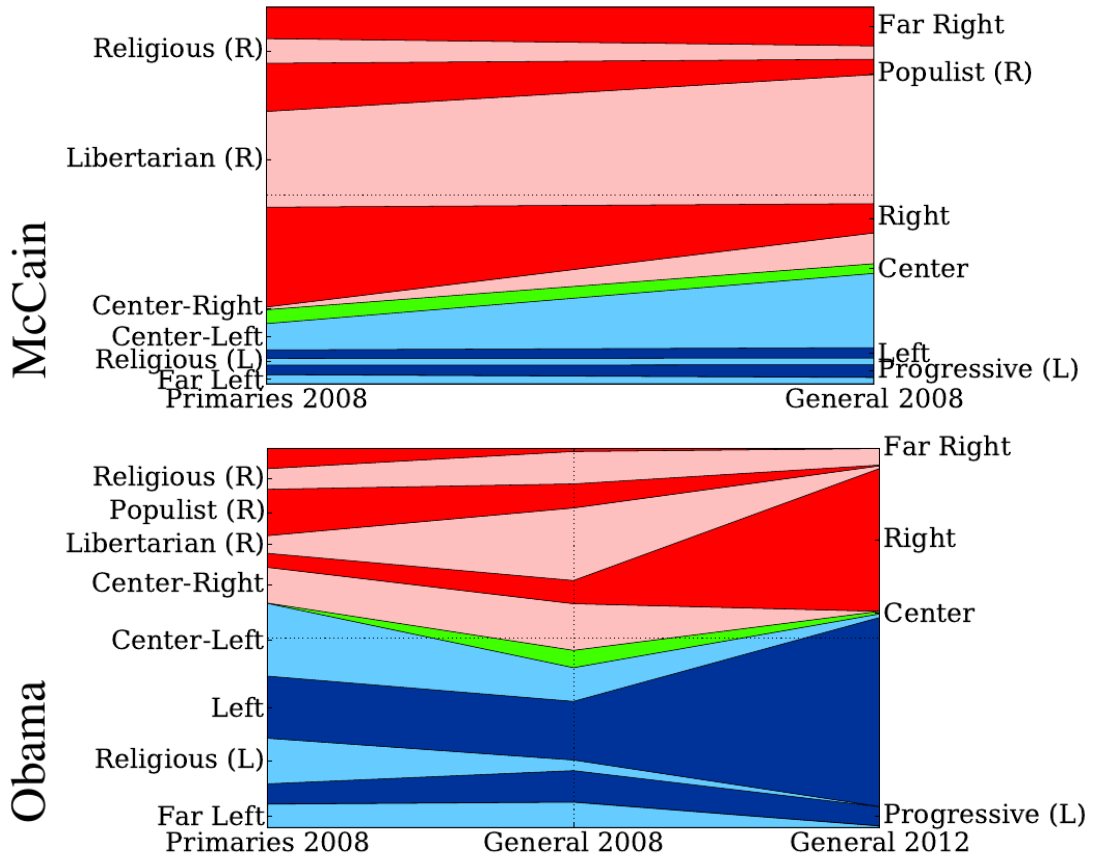


Figure 4.3: Results of the ideology analysis by Sim et al.: Proportion of time spent in each ideology by McCain and Obama during the 2008 ans 2012 election campaigns. Figure taken from [Sim et al., 2013], p.100.

the war. Bush's policy was not supported anymore, and public wanted a change[5]. Absolute positions within the left-right ideology, however, cannot be captured by the analysis, as long as no extremes of both ends of the spectrum are included in the analysis. The scale thus is not bound to any external definition of left and right ideologies. Nevertheless, the advantage of our method is that it reveals the dimension on which the input texts differ, and the relative movements of positions.

Welfare and Quality of Life. Results for the analysis of the topic *Welfare and Quality of Life* are shown in Figure 4.5. In this diagram, there are two interesting observations. First of

[5]However, Bush was not allowed to run as a candidate anyways due to his third period of being president

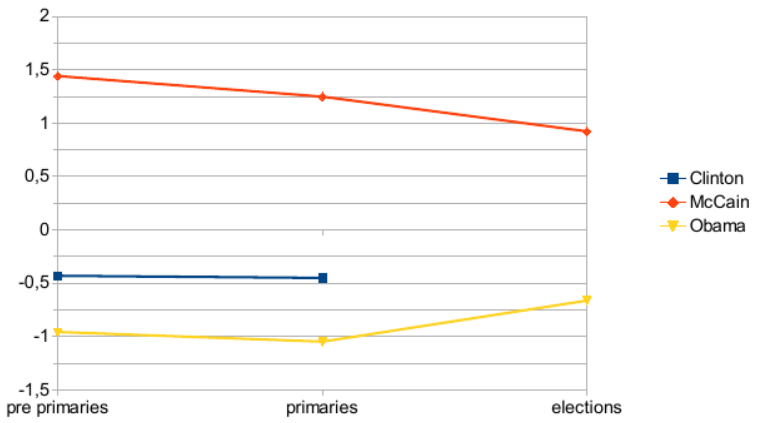


Figure 4.4: Topic *External Relations*: Temporally and topically partitioned set of speeches by Hillary Clinton, John McCain and Barack Obama in 2008 scaled with WordFish.

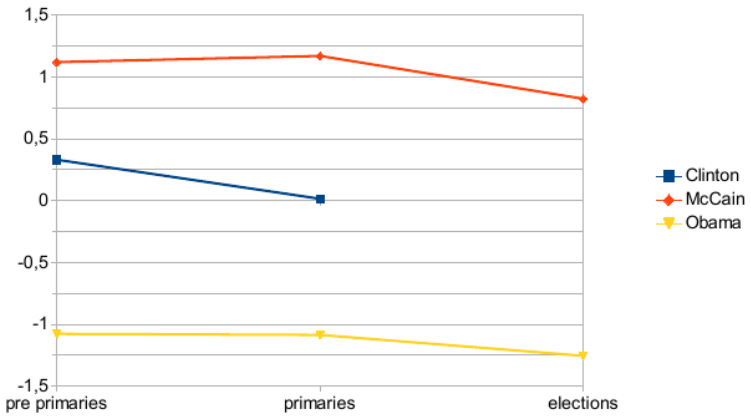


Figure 4.5: Topic *Welfare and Quality of Life*: Temporally and topically partitioned set of speeches by Hillary Clinton, John McCain and Barack Obama in 2008 scaled with WordFish.

all, we see that all the candidates move left (assuming that the revealed dimensions still refer to the common notation of a left-right ideology). Issues within this topic were, among others, health care, the school system and poverty. We have to regard the results with respect to the financial crisis, which headed towards its peak, which it hit in fall of 2008. In this context, social topics gained more importance.

The second salient observation in the diagram, which we noticed also before in Figure 4.1, is the range of Clinton's position: Why is Hillary Clinton closer to McCain as to Obama? We

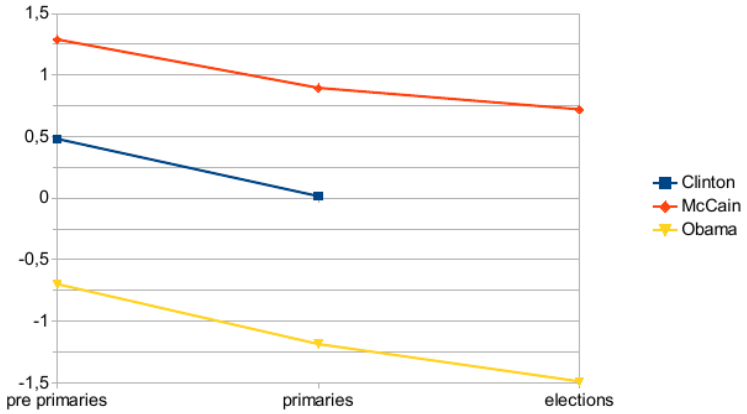


Figure 4.6: Topic *Economy*: Temporally and topically partitioned set of speeches by Hillary Clinton, John McCain and Barack Obama in 2008 scaled with WordFish.

assume the answer can be found in the issues that the candidates focussed on: While Obamas focus was mainly on health care, the Clinton and McCain gave more room to abortion and poverty compared to Obama. In addition, one one particular issue Clinton differed strongly from Obama, putting her closer to McCain: while Obama supports abortion, Clinton is skeptical. Though she does not completely refuses it, she says that "abortion should remain legal, but it needs to be safe and rare".[6].

Furthermore, she supports religion in the public square, though not as strongly as McCain, who considers America as a nation founded by Christians, yet distant from Obama, who supports a strong separation of church and state.

Topic: Economy. In Figure 4.6, we show the results for the topic *Economy*. As mentioned when discussing the topic *Welfare and Quality of Life*, 2008 was the year of the Great Recession, in which the economic crises peaked in September. McCain did not hit the nerve of the people when he did not remember how many houses he owns when being asked in an interview, and reassuring that believes that the fundamentals of the economy were strong. At the same time, Obama won the favor of the people by reacting adequately to the crisis. Given this background, the candidates reacted to the crisis and the sense of the people and moved their positions more and more to the left.

[6] http://www.pewforum.org/2008/11/04/religion-and-politics-08-hillary-clinton/

4.6 Conclusions

We presented a method for fine-grained position analysis on topic level. The qualitative discussion of the results in the previous chapter shows the applicability and usefulness of this approach. In fact, our analysis on the entire text of the speeches did not add any new knowledge and reconfirming already well known facts, such as the global position of candidates over the political spectrum and a common trend in political campaigns, namely the convergence to the center of the selected party candidates after the primary race, known as Etch-a-Sketch phenomenon. The same applies to similar studies such as done by Sim et al. [Sim et al., 2013], which we compared to our findings. In contrast, the fine-grained temporally and topically sliced analysis proposed in our study enables to dig deeper into the process of the candidates converging on a specific position. As a matter of fact, it presents a more clear understanding on how candidates have been positioning themselves regarding different relevant political issues, such as *External Relations* or *Welfare and Quality of Life*. Additionally, it highlights interesting variations on the established idea of positioning during political campaigns (e.g. the shift to-the-left of Barack Obama for *Economy* and *Welfare and Quality of Life*), which are completely ignored by a coarse-grained overview of the election campaign.

The method we presented is language independent. The only precondition is the availability of a training set to determine and partition the topics tackled in the speeches. We compared two strategies of training the topic classifier. On the one hand, we manually annotated a part of the data with topics. On the other hand, we tried to create a training corpus by bootstrapping the labeling process of the training data. The advantage of the second strategy is that it does not require any manual input. Furthermore, as it builds upon the data of the Comparative Manifesto Project, which provides data for many languages and countries, it could be applied within our method for multiple languages. In addition, it would have provided an external, independent definition of topics which is an important property for subjective and reproducible analyses. The downside of the semi-automatically labeled corpus is that it does not lead to a sufficient performance of the topic classifier.

As a matter of fact, the bootstrapping process to label the corpus of party election programs automatically worked well and achieved a good results. It shows that adding external knowledge about text composition improves the performance of fine-grained topic classification. Yet we need further investigations to transfer the topic classifier trained on the party manifesto corpus to corpora containing documents of a different text type.

5

Doves and Hawks in the FOMC

5.1 Motivation and Related Work

The Federal Open Market Committee (in the following referred to as FOMC) is the most important monetary policymaking body of the central banking system of the United States. It consists of the president of the Federal Reserve Bank of New York, the seven members of the Federal Reserve Board of Governors, and four of the eleven Reserve Bank presidents, who receive voting right on a rotating basis. The other seven presidents still attend the meetings and are involved in the discussions. The FOMC meets on a regular basis at least eight times a year to discuss the national and international economic situation as well as the state of the economy in the members' local districts. Furthermore, they give predictions on the future development of the situation and decide on monetary policy actions. At the end of each meeting they vote on a target level for the federal funds rate, which is a short-term objective for the Fed's open market operations.

However, not only the decisions of the FOMC influence the economic situation, but also the communication of the FOMC to the public has a huge influence on it. [Blinder et al., 2008] surveyed various approaches of central bank communication analysis and draw the following conclusion:

> *"... communication can be an important and powerful part of the central bank's toolkit since it has the ability to move financial markets, to improve the predictability of monetary policy, and the potential to help the monetary authorities achieve macroeconomic objectives such as low and stable inflation."*

The awareness that transparency in communication stabilizes the economy did not rise before the early 90s. In 1994, the FOMC began to release their decisions on the fund rate, which was extended since 1999 by adding a bias and giving fuller statements. The minutes of the meetings were only published since 2005.

Why is the transparency and communication so important for the economy? William Poole [Poole, 2001] claims that when markets can predict central bank actions, it enables them to make more efficient decisions. Also, Kohn and Sack [Kohn and Sack, 2003] show that the volatility of various asset prices reacts significantly to statements by the FOMC and its members.

Further evidence for the positive effect of central bank communication are given by [Demiralp and Jorda, 2002], who find that transparency raises efficiency of policy implementations, and by [Ehrmann and Fratzscher, 2007], who show that speeches and interviews by FOMC members influence interest rates.

To sum up, all those studies clearly show that there is a high need for quantitative analyses of central bank communication.
The findings in recent studies go even further and show that not only the presence of information affects the economy, but also the quality is an important factor.

[Fracasso et al., 2003] find that higher quality reports lead to smaller policy surprises and reduce the uncertainty of private agents. They further find that also the quality of the writing style matters. Similar findings are reported by [Jansen, 2011] who say that clarity reduces volatility of interest rates.

[Blinder et al., 2008] conduct a study in which they model interest rate expectations including - among other factors - communication as a variable reaching from “crystal clear” (in case of numerical targets) till “cryptic” (some unclear verbal statements). They find that information about short term rates and future short rates given by the members of the central banks has more influence on the coordination of financial market agents as the actual overnight rate. This shows that communication is important, but it also shows that some of the literal communication is difficult to interpret for financial agents. There might thus be the need for methods of quantitative text analysis methods that assist in understanding difficult verbatim content.

So, what is the information communicated by the central bank actually about?

According to [Blinder et al., 2008], central banks communicate information on four areas:

- They state objectives and the current strategy towards the monetary policy.
- They give motivations for certain decisions.
- They try to give an economic outlook based on the current situation.

- They make predictions for future monetary policy decisions.

[Blinder et al., 2008] state that central banks usually do not publish a general, precise policy rule. Financial agents can yet derive the policy rules by observing the central bank's communication and decisions. An important fact is that each central bank has a general mandate. In case of the FOMC, the highest objective is price stability and full employment. For the European Central Bank, for example, it is stability of the Euro.

Central banks differ in the amount of information they publish on the forecast of inflation. The FOMC publishes quarter-annually reports and frequent official forecasts of output measures. FOMC regularly releases statements of forward-looking assessments, also referred to as bias or balance of risks.

In their survey paper, [Blinder et al., 2008] list several studies analyzing central bank communication with quantitative methods, such as [Jansen and de Haan, 2005], [Ehrmann and Fratzscher, 2007], [Rosa and Verga, 2007] [Musard-Gies, 2006]. For such analyses, scholars hand-code separate statements using pre-defined keywords or annotate according to their expert knowledge. Blinder et al. point out that those classifications are biased by subjectivity and misclassifications. These issues can be improved by having various annotators classify statements, yet not be avoided.

The studies clearly show that central bank communication plays an important role for the economy, and that there is a need for reproducible, robust methods to analyze especially the verbal part of the communication.

The type of information published by central bank differs, especially in the case of their votings. While the European Central Bank does not even publish minutes of their meetings, and their decisions are met unanimously, the British Central Bank even publishes individual votes with names. As the members of the latter Central Bank vote directly according to their opinions, dissenting votes appear commonly. In case of the FOMC, things are different: this committee mostly decides unanimously, however there are some - yet rare - no-votes. Apart from its collegiate structure and its nearly always unanimous votes, one can still observe a high diversity in the statements of the single members compared to other central banks. When members of the central bank communicate with disparate voices, it can cause uncertainty (cf. [Blinder et al., 2008]), which deserves particular attention in economy. As a result, measuring

uncertainty is another important task for scholars the financial domain.

Our objective is thus to develop quantitative methods which enable the analysis of verbal communication by central banks through the example of the FOMC. We provide a reproducible, robust and objective measure to estimate members' opinions from text.

The two opposing general positions in the Central Bank context are commonly referred to as monetary **doves** and **hawks**. Monetary hawks are afraid of inflation. Their goal is to retain stability, thus they are in favor of high interest rates and a stable price level. Doves, on the other side, do not anticipate a high inflation. Their objective is low unemployment. To stimulate economy, they support low interest rates and favor quantitative easing.

Observers of the FOMC and the financial situation from time to time publish their estimates of the FED's members' positions, e.g. the 2014 Fed Dove-Hawk Scale released by Thomson Reuters[1]. However, there is nothing like an official publication of positions for the members of the FOMC, and the observer's estimates are very subjective and may diverge. In 2014, Thomson Reuters considered Narayana Kocherlakota, president of the Federal Reserve Bank of Minneapolis in that time, as being hawkish, whereas a scale provided by Neil Soss from Credit Suisse[2] positions him to be very dovish for the same year. Even though publishing dove-hawk scale representations of the FED's members has recently become more popular, there is hardly any information about the agents positions before 2008.

The examples given above show that for quantitative analyses there is a need for comparable and reproducible methods to derive positions. In this work we therefore aim at a robust approach to place members of the FOMC on a dove-hawk scale that meets these requirements.

In the non-public meetings of the FOMC, we expect the members to openly state their opinions. We hence expect the utterances of a member transcribed in the meeting minutes to reflect the members' positions on the monetary policy. This means the position of a speaker on a dove-hawk scale can be inferred from the content of his utterances in the transcriptions. We will test our hypothesis by predicting the dove/hawk positions of FOMC members based on the content of the transcriptions without using background knowledge about the speaker. We compare our results to the positions determined by [Eijffinger et al., 2015], who hand-coded the same transcriptions.

[1] `https://graphics.thomsonreuters.com/F/10/scale.swf`
[2] `http://www.businessinsider.com/the-feds-hawkdove-scale-2013-10?IR=T`

So far, we sketched related work in quantitative analysis on central bank communication data. The reported approaches and studies all stem from the domain of political science and economics. To complete the picture, we briefly refer to related works from the domain of natural language processing. In fact, there is hardly any work concerned with financial data. The most related work to ours is introduced by [Kogan et al., 2009]. The authors predict financial volatility from financial reports. They model their text as a bag of words model using unigrams and bigrams and predict the volatility value on a continuous scale, using Support Vector Regression. They show that including textual features from current financial reports improves predicting volatility scores compared to using historic volatility values only. [Kogan et al., 2009]s' work and ours have in common that we both use Support Vector Regression on text to predict a value on a continuous scale, which is very rarely in natural language processing, as we have discussed in chapters 1.1 and 1.

Another paper on the field of natural language processing that applies regression on politically specialized language is published by [Yano et al., 2012]. They use logistic regression combining non-textual features with the textual data of a bill to predict whether a bill passes the congressional committee.

We briefly mention the work of [Wang and Cardie, 2014] and [Galley et al., 2004], which both work on the classification of agreement and disagreement in dialogs - more precisely, dialogs in Wikipedia discussion pages in the first paper and dialogs of meetings in a computer science institute in the second paper. Both papers formulate the task as a binary classification problem rather than predicting agreement and disagreement as a variable on a scale. The model of [Wang and Cardie, 2014] depends on a socially tuned sentiment lexicon, the approach presented by [Galley et al., 2004] benefits from structural features, such that the addressee of a statement is the previous speaker who shows an opposing position. As the nature of these dialogs is very different from the highly domain-specific discussions of the FOMC members, those approaches are not suitable to our task.

To give an impression for the difficulty of the language, we present some samples taken from the meeting transcriptions.

- *"It's not clear how likely this is, but if it happened, it would be very costly. A spike in long-term yields could be particularly harmful today for elevated housing prices. It would raise long-term mortgage rates directly, obviously. Moreover, it would force us to raise short-term real rates. And in such circumstances I think it would be even harder for us to facilitate this handoff of investment from the housing sector to the business sector without an intervening recession. "* Mr. Lacker, June 2005.

- *" I wanted to ask a different question on the international side. Karen, on the external forecast, how much of a change is this view of where the current account-GDP ratio goes relative to your expectation six months ago or thereabouts? It seems to me that it looks slightly darker."* Mr. Geithner, June 2005.

5.2 Scaling Doves and Hawks

The transcriptions of the FOMC meetings are published as PDF documents on the official web site of the Board of Governors of the Federal Reserve System[3] five years after the meeting date.

Each document contains about 50,000 - 86,000 words. We use CSV versions of the released PDFs converted and shared by John Wilkerson in the context of the ACL Unshared task[4]. Each turn, i.e. one continuous utterance of a member, is listed with its speaker. Our data set contains the meetings of the years 2005 - 2008.

In her book, [Schonhardt-Bailey, 2013] analyzes the structure of the FOMC meetings in that period. The meetings can be divided into 4 parts: In the first phase, the desk manager gives a summary about domestic and international market operations. In the second phase, senior staff economists report about news in national and international economy and give a forecast. The third phase is a go-round in which each member except for the chairman and the previous speakers give a report about the current economic state in their district and their view of the national situation.

Each of the three phases is followed by a short discussion round in which the members can ask questions to the speaker.

The fourth and last phase is begun by the chairman. He summarizes the previous views and suggests the interest rate which is subject to the decision. Then follows a go-around in which every member expresses his or her opinion towards the suggested policy. When a member completely agrees with the suggestion of the chairman, they might just say something like "I agree with you. ". However, especially if they disagree, they clearly state their concerns and give reasons. Phase 4 is concluded with the poll on the policy. Although the members do have different objectives and opinions, the committee usually decides with consenting votes,

[3] `http://www.federalreserve.gov/monetarypolicy/fomc_historical.htm`
[4] `https://sites.google.com/site/unsharedtask2014/`

dissenting votes appear rarely ([Havrilesky and Gildea, 1991], [Adolph, 2013]).

As mentioned in the previous section, [Eijffinger et al., 2015] use the same data set in their work. They estimate the committee members' ideal points on a dove-hawk scale. They hand-code the transcriptions according to a method presented in [Meade, 2005]. The basis for the coding process is the fourth phase of the meetings, where the members are either in favour of an adaption of the interest rate or against it. For each speaker, the coders label single utterances as dovish or hawkish. Mostly the utterances can clearly be assigned one of the labels. Unclear statements were not coded. If a member states that he agrees with the chairman, this statement is assigned the same label as the chairman's suggestion; or the opposite label in case of disagreement. The authors are aware that the annotations might be subjective, yet they checked the inter-annotator-agreement on a subset of documents and found the four independent annotators to agree in all cases. An example for an easy-to-code sentence - dovish, in this case - is the following: *"Therefore, I can certainly support another increase in the funds rate of 25 basis points today" (Yellen, transcripts of June 2006).* Finally, the difference of dovish- and hawkish statements per speaker is calculated and interpolated with the general tendency of a meeting according to the dominating general opinion.

Our hypothesis is that the positions of the speakers can be derived from the content of their utterances in the meetings, thus from the words in the transcripts. We aim to predict the position of each speaker per meeting.

Regression. Regression analysis is the estimation of a function that maps a set of independent variables on a dependent target variable. The regression function can then be used to predict the value of unknown target variables. In quantitative analyses for political science and for economy, Ordinary least squares (OLS) regression is widely used. It is a regression that estimates the best function by minimizing the sum of the squares of the errors that are made when predicting the the dependent variable. However, a shortcoming of OLS is that it cannot cope with a large amount of attributes. Therefore it is not suitable for our problem, in which we model text as a word vector.

Support vector machines (SVM) have been shown to work good with textual features. There are implementations of support vector machines that can be applied to regression problems [Smola, 1996], usually referred to as support vector regression (SVR). Like in other SVMs, the data is mapped into a higher-dimensional feature space in which linear regression is possible.

In the same manner as the non-linear Support Vector Classification approach, a the data is transformed to a higher dimensional feature space where linear regression is performed.

The difference to classification SVMs is an alternative loss function.

We assume the space of input patterns $\mathbb{R}^N$, i.e. the set of independent variables, as our training data: $\{(x_1, y_1), ..., (x_\ell, y_\ell) \subset \mathbb{R}^N \times \mathbb{R}\}$

Based on the training data, we aim to find a function $f(x)$ that predicts y_i with less deviation from the actually obtained values as possible. Assuming the training samples were drawn from a probability distribution, P is the probability function that generated the observations. As we do not know P, we need to estimate a function that minimizes the risk function $R[f]$.

$R[f] = \int_{X \times Y} l(y, f(x)) P(x, y) dx dy$

In contrast to SVM, for SVR the loss function l has to be modified to include a distance measure. The ε-insensitive loss function [Vapnik, 1995] contains the a-priori chosen parameter ε, which specifies the accepted distance $\varepsilon > 0$ between a predicted data point and the actual observation. Errors below this threshold are accepted. The cost c is: $c(\xi) = |\xi|_\varepsilon$

Like for SVC, kernels can be arbitrary chosen. In his practical guide to SVRs, [Hsu et al., 2003] recommends a linear kernel for problems with a large number of attributes.

We use the nu-SVR implementation of [Schölkopf et al., 2000] and a linear kernel contained by Rapid Miner[5], a framework for text mining.

5.2.1 Experiments Determining Relevant Turns

In a first row of experiments, we investigate whether dove and hawk positions can be actually derived from utterances of the members in the meetings. We further analyze which parts of the discussion, i.e. which turns of the speaker, provide relevant information. We compare 3 different sets of turns: As a standard setting, we use all turns of a speaker in a meeting for the regression. In the second experiment, we only use the turns of the last go-around as data, which corresponds to phase 4 of the meetings. This section of the transcriptions that was used in [Eijffinger et al., 2015] for the hand coding. In a third experiment, we distinguish between two types of turns: statement turns, in which the members elaborate on their view of the economy and give an outlook, and shorter turns of more spontaneous nature, which are contributions to discussions that arise from the first type of statements. In the remainder of

[5] https://rapidminer.com/

the section we will give more detailed information about how we model the data and how we extract the subsets of turns and compare the results.

All Turns. We model each speaker per meeting as a separate instance. We consider this setting as the standard system, as we use the complete set of turns as data. The speaker-meeting-instances are modeled as bag-of-words. We filter stopwords and most frequent as well as least frequent words by rank (prune below rank 0.05 and above rank 0.95). The same preprocessing is used in all following settings.

We perform 10-fold cross validation over randomly sampled instances of the data set, with no regard to year or speaker. For the regression, we use nu-SVR with C = 10, nu = 0.1, cache size = 80 and $\epsilon = 0.001$, which is the default setting. We keep these parameters throughout all following experiments.

Phase 4 Turns. In this setting, we only use turns that were uttered in the fourth phase of the meeting, which corresponds to the information that was used in [Eijffinger et al., 2015] to code the speaker's dove-hawk-positions. The coders did not only code typical dove or hawk statements, but they also interpreted agreement and disagreement relations among the members, especially towards the chairman. We can assume that they further had some background knowledge of the domain which helped coding the stances. We investigate whether the vocabulary of the turns in phase 4 is enough to determine the dove-hawk positions of the FOMC members.

To extract only those turns from the transcriptions, we apply a simple heuristic. The fourth phase is finished by the chairman reading out the results of the votes. This statement contains the names of the speakers together with "yes" or "no". The last but one statement of the chairman is the summary of the previous opinions and discussions, and his suggestion for the target rate to vote on. In each transcription, we search for those two particular statements. Then, for each speaker we collect all of his turns in between the chairman's statements. The chairman is not included in this setting, though he could be classified with the resulting model in a similar way by regarding his last but one statement.

Statement Turns. In the previous setting we only regarded the turns of the last phase. However, it is the previous three phases in which the members of the committee give their view on the current economic situation and its expected development. If doves and hawks are defined by their view on the economy, we assume that also these parts of the meeting contain

valuable information for our position estimates. As described in Section 5.1, after each report there is the possibility to ask questions, and discussions come up.

We found the turns of the reports to differ from the turns uttered in these discussions. In the first type, the speakers state their opinion, expressing arguments that they have assumably prepared in advance. The contributions to the discussions are shorter and of spontaneous nature. In the following, we will refer to the different types as statements turns and discussion turns.

Statements turns are prepared and reflect the general position of the speaker. According to research in political science, the political position of a speaker is determined by the topics he speaks about (cf. [Grimmer and Stewart, 2013], [Hillard et al., 2007],[Laver et al., 2003]) . The speaker will use the possibility to expand on the topics he considers important.

The shorter discussion turns are spontaneous reactions to the discourse contributions of the previous speakers. Rather than the topics the speakers considers important they contain an attitude towards previous speeches: the speaker often expresses agreement or disagreement, as in *"I can see why you assume that, but ..."* or *"To be honest, I don't think"*.

We believe that the statement turns contain the vocabulary containing the information of their dove-hawk position, and the discussion turns introduce noise in the estimation process. To proof this hypothesis, we filter the statement turns and remove the discussion turns from the data before performing the regression. We will evaluate whether making use of the structure of the meetings improves the classification results compared to the standard which uses the complete set of turns.

The distinction between statement and discussion turns is done as follows: We first manually annotated one meeting, classifying each discourse contribution as either statement or discussion turn. Figure 5.1 shows the results of plotting the consecutive turns of the meetings with their wird count and their statement/discussion label.

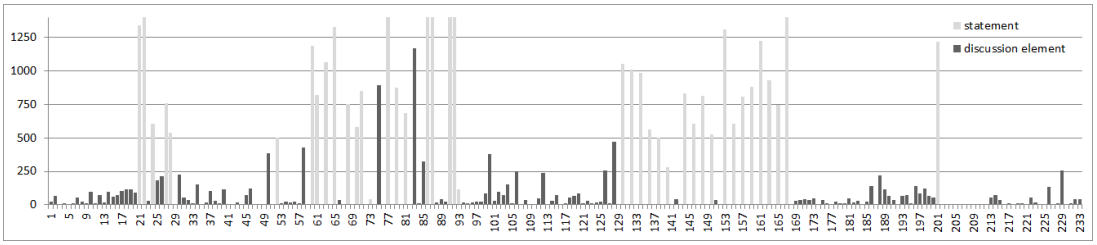


Figure 5.1: Manually annotated discourse contributions and their lengths (word count).

From the diagram, we can see that the threshold between the statement and discussion turns is around 500 words. We use this number as a shallow heuristic to automatically distinguish between the two turn types.

For this setting, we filter and dismiss the discussion terms and keep only the statement turns for the regression. We would like to point out that the turns belonging to phase 4, which are used in the setting described above, are apart from very few exceptions all of discussion type. Therefore, we can consider the data sets for this and the previous setting as nearly non-overlapping.

Results. To measure the results, we report the root mean squared error and the absolute error of the predictions as well as the correlation between the prediction and the gold standard labels.

The **absolute error AE** is the average of the absolute distances between the observation y_i and the prediction $\hat{y}_i$:

$$AE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

The **root mean squared error RMSE** is the square root of the squared distances between observation y_i and prediction $\hat{y}_i$ averaged by the number n of samples:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The **Pearson's correlation coefficient (PCC)** measures the strength of the linear relationship between two sample variables x,y:

$$PCC = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - mean(x)}{s_x}\right)\left(\frac{y_i - mean(y)}{s_y}\right)$$

where $mean(x)$ is the sample mean ($mean(y)$ respectively):

$$mean(x) = \frac{1}{n}\sum_{i=1}^{n}x_i$$

	average baseline	all turns	phase 4 turns	statement turns
RMSE	0.744 (+/- 0.5)	0.494 (+/- 0.047)	0.706 (+/- 0.057)	**0.475** (+/- 0.036)
mikro	–	0.496	0.708	0.476
AE	0.550 (+/- 0.501)	0.346 (+/- 0.044)	0.549 (+/- 0.046)	**0.330** (+/- 0.029)
mikro	–	0.346 (+/- 0.356)	0.549 (+/- 0.448)	0.330 (+/- 0.343)
PCC	0.00	79.6% (+/- 0.055)	44.0% (+/- 0.202)	**83.1%** (+/- 0.043)
mikro	–	78.8%	44.5%	82.1%

Table 5.1: Results comparing subsets of turns. Root mean squared error = RMSE, Absolute error = AE, Pearson's Correlation Coefficient = PCC.

Table 5.1 shows the results of the experiments described above, in which we apply support vector regression to different sets of turns. The scale introduced by [Eijffinger et al., 2015] ranges from -2 (dove) to 2 (hawk). The values in the gold standard lie between 1.75 for Mr. Lacker and -1.35 for Mr. Mishkin. We compare our results to an average baseline that assigns the average dove-hawk score, which is -0.033, to all speakers. As expected, the baseline system provides a very high RSME of 0.744 as well as an AE of 0.55, and it shows no correlation with the gold standard classification. The **all turns** system that makes use of the whole transcriptions has an absolute error of 0.35 and a root mean squared error of 0.49. This says that on average, a prediction has a distance of 0.35 to the gold standard. The correlation is with 79.6% is considerable.

We observe lower results for the **phase 4 turns** setting with a root mean squared error of 0.7, which is quite high, and a low correlation of 44% only. Although the same subset of turns was used for the manual coding of the speaker's positions, the words of those turns themselves do not contain enough information to derive these positions. We assume that the process of hand coding is possible on this subset of the data for two reasons: first, human readers can easily detect agreement or disagreement among the speakers and can thus infer positions via relations between the speakers. For example, if a speaker says that he or she agrees with the chairman, and the coder knows the position of the chairman, they can infer the position. Furthermore, they have background knowledge about the domain.

However, the setting using only **statement turns** significantly outperforms the previous two settings. The absolute error of 0.33 and root mean squared error of 0.47 can be considered low, and it achieves a strong correlation of 83.1 %. We tested the significance of the improvement of using statement turns over using all turns with a one-tailed paired t-test on the prediction

errors per speaker and meeting of both settings, which is $\alpha = 2.64E^{-028}$, which indicates a very high level of significance.

In Figure 5.2, we visualize the results of this system. The red line shows the average dove-hawk values predicted by our system based on statement turns. The dashed gray line represents the positions derived by [Eijffinger et al., 2015]. We can see that both lines overlap to a far extent. The light gray dashed lines on the left and right side mark the minimum and maximum dove hawk values that [Eijffinger et al., 2015] found for a meeting for each speaker. Even for speakers whose positions vary in a broad range of the dove hawk spectrum, we predict a position very close to their average, as for example for Mr. Santomero, Mr. Bernanke or Mr. Warsh. It is remarkable that we can reproduce very close positions to the manually coded ones from a partition of the data (i.e. the statements turns) that does not or only slightly overlap with the part of the data (i.e. the discussion turns) that was used for the manual annotation process. This suggests that the approach is appropriate as a reproducible and robust measure to analyze speaker's positions, avoiding bias from human interpretation to a certain extent. This confirms our hypothesis that the FOMC speakers describe their opinion and thus their dove-hawk position mainly in the statements parts of the meetings. Filtering the dialogs removes noise from the classification. We conclude that employing knowledge about the dialog structure improves position analysis.

Inspecting Support Vectors. We will now give an insight into the learned model. Table 5.2 shows highest and least weighted support vectors trained on the whole data set. A support vector consists of a term and weight. The highest weights bias the score towards a hawkish position, the low weights to a dovish position. This means, the higher a score for a term, the more "hawkish" is the term.

Some of the terms might seem unfitting, like e.g. *narrative* for hawks. It refers to the economic *narrative*, which means the way an economic issue is framed. *Okun* stands for *Okun's law*, which refers to an empirically observed correlation between unemployment and loss of production. *NAIRU* is short for non-accelerating inflation rate of unemployment. *Nonfarm* is a part of the term *Nonfarm payrolls*, which is a monthly released economic indicator which intends to represent the total number of employees form any business. The most distinctive term for hawks is *fifth*. The Richmond Federal Reserve Bank is the head of the *Fifth Federal Reserve District* which comprises Virginia, Maryland, the Carolinas, the District of Columbia and most of West Virginia. Its president since 2004 is Jeffrey M. Lacker, who is known to be a hawk, and is labeled with most hawkish score of our data set. The Fifth District's monthly

report on economic indicators has importance for the decisions of the FOMC. Terms like *Katrina* show that current events have an influence on the discussions of the FOMC and the use of vocabulary changes depending on the particular topics. *Hurricane Katrina* was with an estimated damage of 108 billion US Dollar the costliest disaster in history of the United States ([Knabb et al., 2005]) and caused damage in several states in 2004, so it had a huge impact on the US economy.

Dove-biased support vectors		Hawk-biased support vectors	
term	weight	term	weight
structural	-1.379	fifth	3.188
premium	-1.257	reserves	2.904
countries	-1.215	shipments	1.843
narrative	-1.113	intentions	1.468
okun	-1.052	service	1.408
nairu	-1.017	virginia	1.358
nonfarm	-0.952	prevent	1.329
modal	-0.951	base	1.269
crunch	-0.95	adopt	1.18
stance	-0.93	tips	1.172
loan	-0.879	intervention	1.17
bond	-0.86	katrina	1.135
imports	-0.843	categories	1.113
households	-0.822	ceo	1.108
wording	-0.8	conduct	1.083

Table 5.2: Highest and least weighted support vectors of SVR on statement turns.

5.2.2 Example: Predictions for the Year of the Financial Crisis

To get an insight into the predictions by our model, we plot the predicted dove-hawk scores for FOMC members in the meetings of 2008 in Figure 5.3. We chose this year due to the financial crises that began at the end of 2007 and reached its peak in September 2008. For better readability of the diagram, we omitted speakers that did not attend all of the meetings in that year.

We can see that in the beginning of the year, the majority of the speakers is on the dovish part of the scale. From in March, even Mr. Fisher and Mr. Lacker, who are considered hawkish, move their position closer to the doves. This is exactly what economy expects as reaction to the crises: doves are in favor of low interest rates to stimulate economy. In June and August,

Mr. Lacker, Mr. Fisher and Mr. Plosser take quite hawkish positions again. Maybe this is the reason for the doves Ms. Yellen, Mr. Kohn, Mr. Kroszner and Ms. Pianalto to express even more dovish positions than before. In September and October, the members' positions converge again and and are kept in October, except for Chairman Bernanke, wo moves to the the most dovish positions of the speakers. September is said to be the peak of the crisis [Altman, 2009]; this is also conform with our observation that in November things ease again and the positions of all members (except for Ms. Pianalto) move clearly in hawk direction.

5.2.3 Experiments on Topicality and Amount of the Training Data

Analyzing the support vectors, we can see that the dove-hawk classification strongly depends on the current economic topic. For example, one of the most prominent topics in the last meetings of 2005 and the beginning of 2006. We are interested in the influence of vocabulary change through the meetings on our model, and test for its robustness. For this reason, we perform another experiment in which we train two models: the first on data from the years 2005 - 2006, the second on the years 2006 - 2007. We apply both models to the transcriptions of the meetings of 2008 and compare the performance. As the previous experiment determined that using statement turns works best for the analysis, we use only this subset in the following experiments.

Topicality. To analyze the impact of topicality of the vocabulary for the classification of doves and hawks, we compare three settings in which we train on two consecutive years and apply the model to a (not necessarily directly) following year:

A (0506 to 07). We train on years 2005 and 2006 and apply the model to 2007.

B (0607 to 08). We train on years 2006 and 2007 and apply the model to 2008.

C (0506 to 08). We train on years 2005 and 2006 and apply the model to 2008.

In the last case, there is a "gap" of one year between training and test data. We want to analyze whether this reflects in the performance of the classification.

The results are shown in Table 5.3. For the two settings in which we classify directly subsequent years, we observe a similar rate RMSE of 0.583 compared to 0.588 and AE of 0.401 to 0.417, though with a higher correlation in the first setting. If we compare the classification of 2008 with the directly preceding training data to the training data with a one-year-gap, we

receive the expected results: in the third setting, the error rate rises to a RMSE of 0.629 , the correlation drops to 64%

	A (0506-7)	B (0607-8)	C (0506-8)
RMSE	0.583	0.588	0.629
AE	0.401 +/- 0.423	0.417 +/- 0.415	0.421 +/- 0.467
PCC	77.0%	71.6%	64.5%

Table 5.3: Results for analyzing impact of topicality with setting A, B and C.

We now investigate the influence of the amount of training data. In the second and third setting described above, we trained on two years each. We will compare the performance of a fourth setting, in which we train on all years from 2005-2007 and classify 2008:

B (0607 to 08). We train on years 2006 and 2007 and apply the model to 2008.

C (0506 to 08). We train on years 2005 and 2006 and apply the model to 2008.

D (050607 to 08). We train on years 2005, 2006 and 2007 and apply the model to 2008.

	B (0607-8)	C (0506-8)	D (050607-8)
RMSE	0.588	0.629	0.566
AE	0.417 +/- 0.415	0.421 +/- 0.467	0.399 +/- 0.401
PCC	71.6%	64.5%	74.7%

Table 5.4: Results for analyzing the impact training data amount with setting B, C and D.

The results presented in Table 5.4 are not surprising: adding a year of training data reduces the RMSE to 0.566 and the AE to 0.399, the correlation rises to 74.7%.

The experiments show that no only the amount of training data matters, but also its topicality.

5.2.4 Experiments on Speaker Data Dependence

In the previous experiments, we had samples for every speaker in the training data. We are interested in the influence of the particular speaker vocabulary on the classification results.

Unseen Speakers. To begin with, we run another experiment to analyze how the model works on unseen speakers. We use leave-one-out evaluation by iterating over the speakers. In each iteration, we filter the data of one speaker from the training data and apply the resulting model on it. Like before, we use only the statement turns.

	Random %-Validation	Leave-one-out on Speakers
RMSE	**0.475** (+/- 0.036)	0.529 (+/- 0.444)
micro	0.476	0.792
AE	**0.330** (+/- 0.029)	0.503 (+/- 0.452)
micro	0.330 (+/- 0.343)	0.571 (+/- 0.548)
PCC	**83.1%** (+/- 0.043)	3.6 +/- 0.186

Table 5.5: Leave-one-out evaluation for speakers.

Table 5.5 shows the performance of the leave-one-out evaluation on speakers. We compare it to the random cross-validation performed in the first experiment. Apparently, the performance is strongly dependent on the presence of training data for the particular speakers: the RMSE increases from 0.475 to 0.529, the AE from 0.330 to 0.503. For matter of completeness, we included the correlation in the results, which drops from 83.1% to 3.6%. However, we point out that correlation in this case is not meaningful: in this setting, the test set consists of the various meeting texts for the same particular speaker only, and thus all instances of the test set have the same value (dove-hawk-score) in the gold standard. The respective function determined by the regression that has the perfect correlation would thus have to correspond to a straight line, which is very unlikely. To illustrate the classification results, we plot the average prediction for each speaker in Figure 5.4. The diagram shows that the predictions for all speakers are tied towards the middle of the scale. Mr. Bernanke participated the meetings in 2005 as a normal member, but became chairman in 2006. We distinguish between the roles of a speaker in the meetings, because the role might influence a speaker's attitude. However, we cannot find significant differences in the predictions for the roles, neither for Chairman Bernanke, nor for Vice Chairman Geithner.

Particular Speakers Prediction. In the above experiment we determined that the regression performance depends substantially on training data for each particular speaker. We run another experiment to make sure that the regression process in deed captures the dove-hawk ideologies rather than just learning the respective label for every speaker.

In order to analyze the potential learning of the speakers, we formulate the task as an authorship attribution problem: for each document (which corresponds to the set of turns of a speaker in a meeting) we predict the corresponding author, i.e. the FOMC member. For this experiment we use the Rapid Miner implementation of SVC with a linear kernel, which is the support vector classification algorithm that corresponds to the support vector regression we used above (see Section 5.2).

Table 5.6 shows the results for the task. The classifier predicts the correct speaker for only 13.71% of the instances in our data set. The weighted mean recall - which is the recall of all classes weighted according to their proportions within the test set - is around 7%, the weighted mean precision 19.39%, which indicates a poor performance. The micro precision looks slightly more promising at first glance with 57%. The reason for this are speakers that appear only in few meetings - if these occurrences are predicted correctly, they achieve a high precision, which has a large influence in the non-weighted micro precision.

As a result we conclude that the representation of the data set that we chose is not capable of predicting particular speakers with a reasonable performance. For this reason we consider the regression in subsection 5.2.1 to predict dove-hawk-values rather than speaker language.

	10-fold cross validation
accuracy	13.71% (+/- 6.85%)
weighted mean recall mikro	7.01% (+/- 2.70%) 8.47%
weighted mean precision mikro	19.39% (+/- 8.59%) 57.32%

Table 5.6: Results for speaker prediction.

5.2.5 Experiments Classifying Speeches

Finally, we are curious whether a model trained on the transcription data is able to capture positions in other types of text, such as speeches given by the FOMC members at other occasions. One salient difference between FOMC meetings and speeches is that the latter type are public and are addressed to a particular audience, whereas the meetings are held behind closed doors. We downloaded a set of 200 speeches given by members of the FOMC between 2008

and 2011 for which we had both training data in our data set and a gold standard label. The speeches are downloaded from the public web page of the Federal Reserve[6].

Almost half of the speeches are given by Ben S. Bernanke. The other speeches in our sample are given by Donald L. Kohn, Frederic S. Mishkin, Kevin Warsh, Randall S. Kroszner and Janet L. Yellen. For this reason, like in the previous experiment, we should not overstate the low correlation of 39.2 %. Table 5.7 shows the results for the classification of the speeches. For easier comparison, we repeated the results of the cross-validation on random folds from the first experiment.

A possible research question in the domain of the FOMC is whether FOMC take smoother positions when they speak in public compared to stating their positions behind closed doors. The fact that the committee tends to take unanimous decisions supports this thesis. Answering this question is beyond the scope of this thesis and would require more comprehensive experiments. Still we want to give an impression of such analyses. Figure 5.5 shows the predicted dove-hawk scores per speaker averaged per year. We plot the gold standard score for each speaker for comparison.

The per speaker averaged predictions for speeches all lie in a close range between -0.5 and 0.02. This could be either due to the fact that the language used in speeches differs strongly from the language used in the meetings, and is more similar among the persons giving the speeches. On the other hand, it could be related to the tendency of the FOMC members not to show strong personal views when they address public. Between 2007 and 2008 Bernanke and Mishkin move towards dove position, which might be related to the financial crisis that started in the end of 2007. We point out that from 2009 the predictions become more vague, as those years are not covered in training data. We leave the application of our methods to the actual research questions and the proper interpretation of results to scholars of economy and political science.

	Random %-Validation	Speeches
RMSE	**0.475** (+/- 0.036)	0.484
AE	**0.330** (+/- 0.029)	0.338 (+/- 0.347)
PCC	**83.1%** (+/- 0.043)	39.2%

Table 5.7: Classification of Speeches.

[6] `http://www.federalreserve.gov/newsevents/speech/2008speech.htm`

5.3 Conclusions

We showed an approach that learns a model on transcriptions of FOMC speakers which is capable of predicting the position of a speaker in an unseen meeting on a dove-hawk scale. The model reproduces hand-codings with a correlation of up to 83%. A remarkable fact is that the best performing model uses information from a part of the meeting transcriptions that is non-overlapping with the part of the documents that was used for the manual annotation: while the hand coders rated the statements in the last go-around of the meeting as dovish or hawkish, our analysis relies on statements from the earlier phases of the same meetings. We showed that using knowledge about dialog structures as well as the structure of the meetings improves the predictions, presumably because we omit passages of the conversations that introduce noise in the position measuring task.

The omitted passages mainly consist of the questioning sessions after a speaker's statement, so they are discussions. As a further step of our approach we intend to make use of the contained information in a way different from the technique we applied to the longer statements. In discussions, the participants address each other, and those conversations often contain dialog markers. We plan to apply techniques of agreement-disagreement classification to reveal supporting or opposing positions of the discussants, which can improve the dove-hawk predictions by considering speakers in mutual dependence.

To sum up, we proposed a reproducible, robust and - to a certain extent - objective method for measuring speaker positions on a dove-hawk scale. The objectivity is provided inasmuch as [Eijffinger et al., 2015] claimed that statements can be unambiguously coded as either dovish or hawkish, and on which their annotators consented in all cases. We further showed how the results of our method could be employed in an analysis to keep track of influences on the speakers caused by external events like the financial crisis in 2008.

On the other hand we also examined the limitations and challenges of our approach. The most limiting issue is that the method depends on training data for each of the speakers, yet this does not hinder the practical applicability of this technique to actual data: members of the committee change rarely, and even if the voting positions rotate, the discussants in the meetings remain the same. To apply the approach to a new speaker, it requires a few hand coded meeting utterances that are added to the data set. An alternative to the hand-coding of data for each new speaker could be transfer learning: there are techniques which allow for adapting a model trained on a data set to the properties of a new data set (cf. [Pan and Yang, 2010], [Jiang and Silver, 2014]).

Furthermore, the approach needs to a reasonable extent up-to date training data, as the discussed topics and therefore the vocabulary changes. This issue could be tackled by an on-line learning approach: new meeting transcriptions can be classified by the regression model, finally they can be added to the training data with their predicted dove-hawk value. In this way, the training data is kept up-to date.

We also investigated how the model trained on meeting transcriptions can be applied to text produced by the same speakers but in other contexts, more precisely to speeches by the FOMC members given to public audiences. The dove-hawk scores predicted for those speeches all ranged in a short distance from the center position. This could either mean that the model does not work for the differing language style of the speeches, or it might be caused by the fact that the speeches in deed all show a very moderate position. In Section 5.1 we discussed that within the central banking system of the U.S. the members tend to vote unanimously to demonstrate unity. This supports the theory that in deed the speakers take a more moderate position when addressing the public than when discussing behind closed doors. It requires further analyses to verify whether this hypothesis matches with the human judgment for the same speeches.

One of the main reasons why economists have the need for analyses of central bank communication is that it has an impact on economic agents and thus stability (see the motivation for this work in Section 5.1). A future direction of the work we presented would be to adapt it to the analysis of central bank communication in press. In [Jansen and de Haan, 2004] the authors analyze central bank communication in news wires. They analyze four subcategories (interest rates, money growth, economic growth as well as inflation), which they hand-code based on selected keywords. An approach that first detects these topics automatically and then captures the dove-hawk dimension for the particular topics would be of immense benefit for scholars in that area.

According to [Blinder et al., 2008], central banks communicate information on four areas:

- They state objectives and the current strategy towards the monetary policy.
- They give motivations for certain decisions.
- They try to give an economic outlook based on the current situation.
- They make predictions for future monetary policy decisions.

In their analyses of influence of central bank communication for Czech Republic, Poland and Hungary, [Rozkrut et al., 2007] found that only speeches about monetary policies have

an influence on, but not those on the economic forecast. Here, again, a distinction between those communication areas - within meeting transcriptions, speeches and other kind of central bank related documents - could improve the predictive power of quantitative analyses.

Extending our approach to make it feasible to measure dove-hawk positions in other types of text besides meeting transcriptions could open up another dimension for the analysis of central bank communication, namely cross-country comparison. According to [Blinder et al., 2008], there is no straight-forward way to compare various central banks so far as they differ in their voting behavior and publishing strategies. The method we proposed is generally language independent, it can be applied to data from other central banks as well, given that they provide meeting transcriptions. The possibility of measuring positions from various sources including verbal communication enables further cross-country analyses.

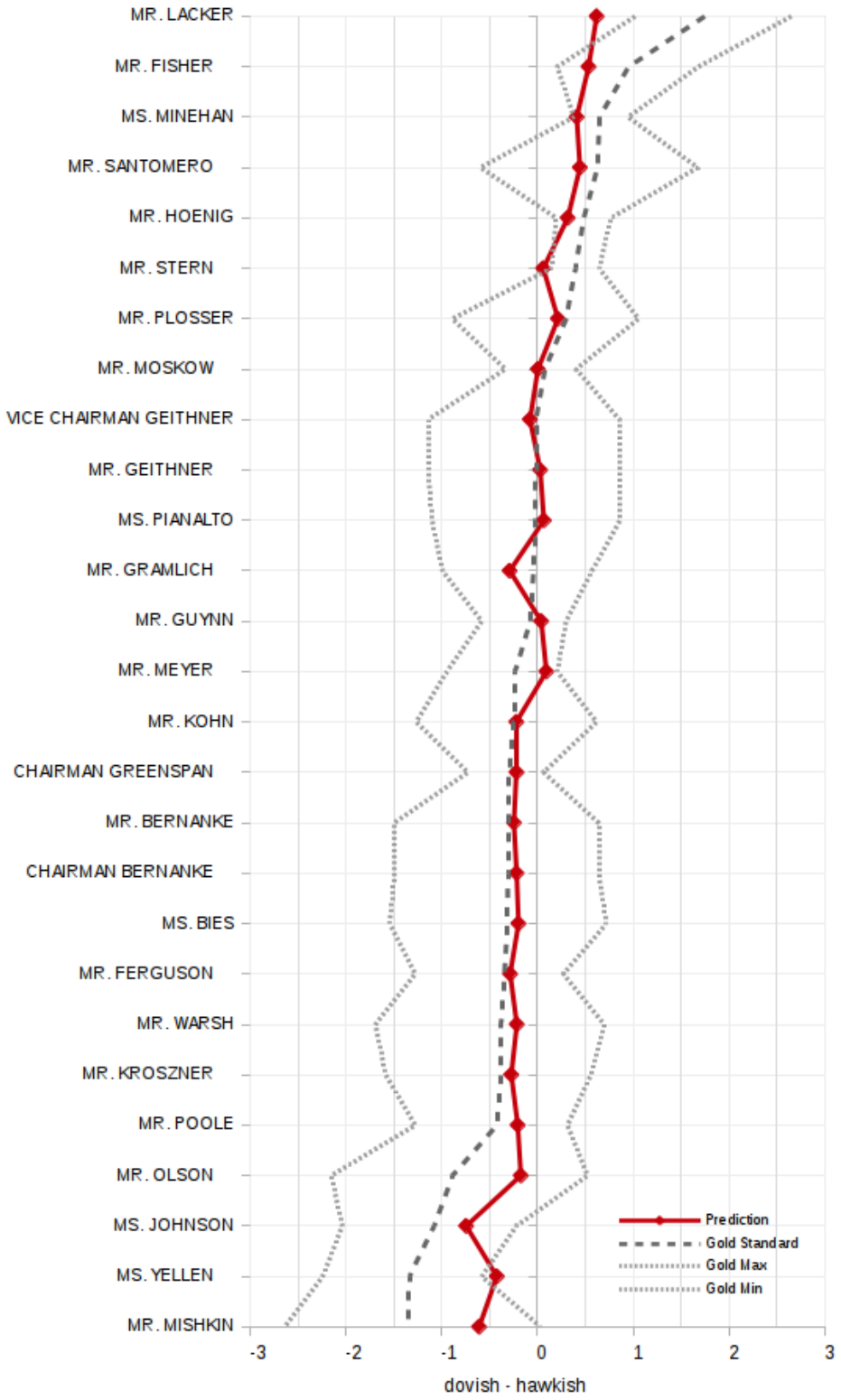


Figure 5.2: Predicted positions for FOMC members. For better readability of the diagram, we connect the single data points - i.e. dove-hawk values of the respective resources - with lines, though they are discreet values.

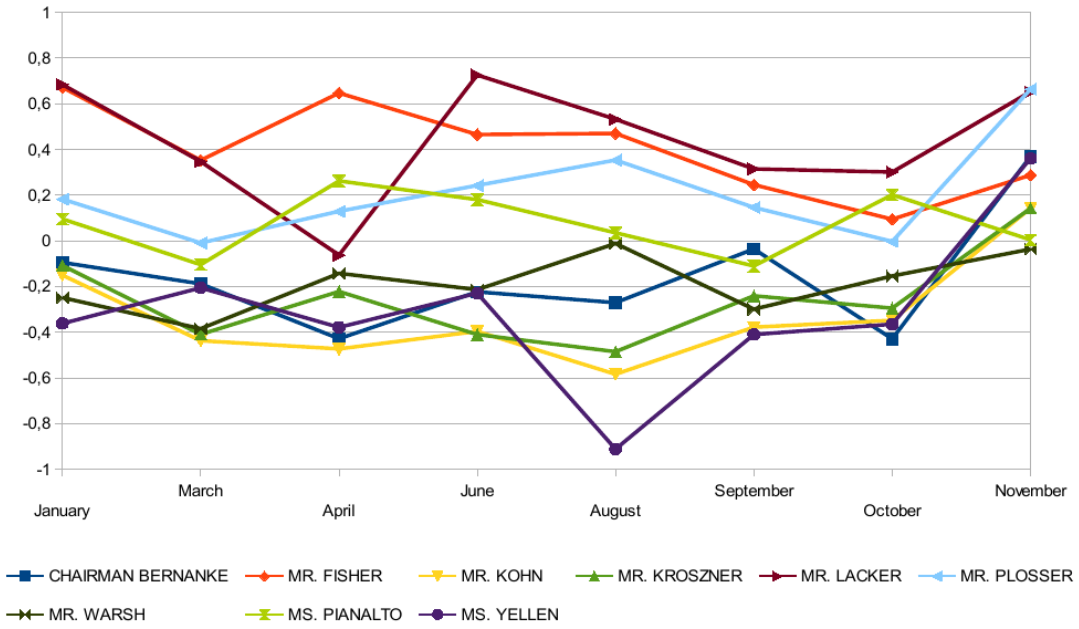


Figure 5.3: Predicted positions for FOMC members in 2008. The y-axis shows dove (negative) and hawk (positive) values.

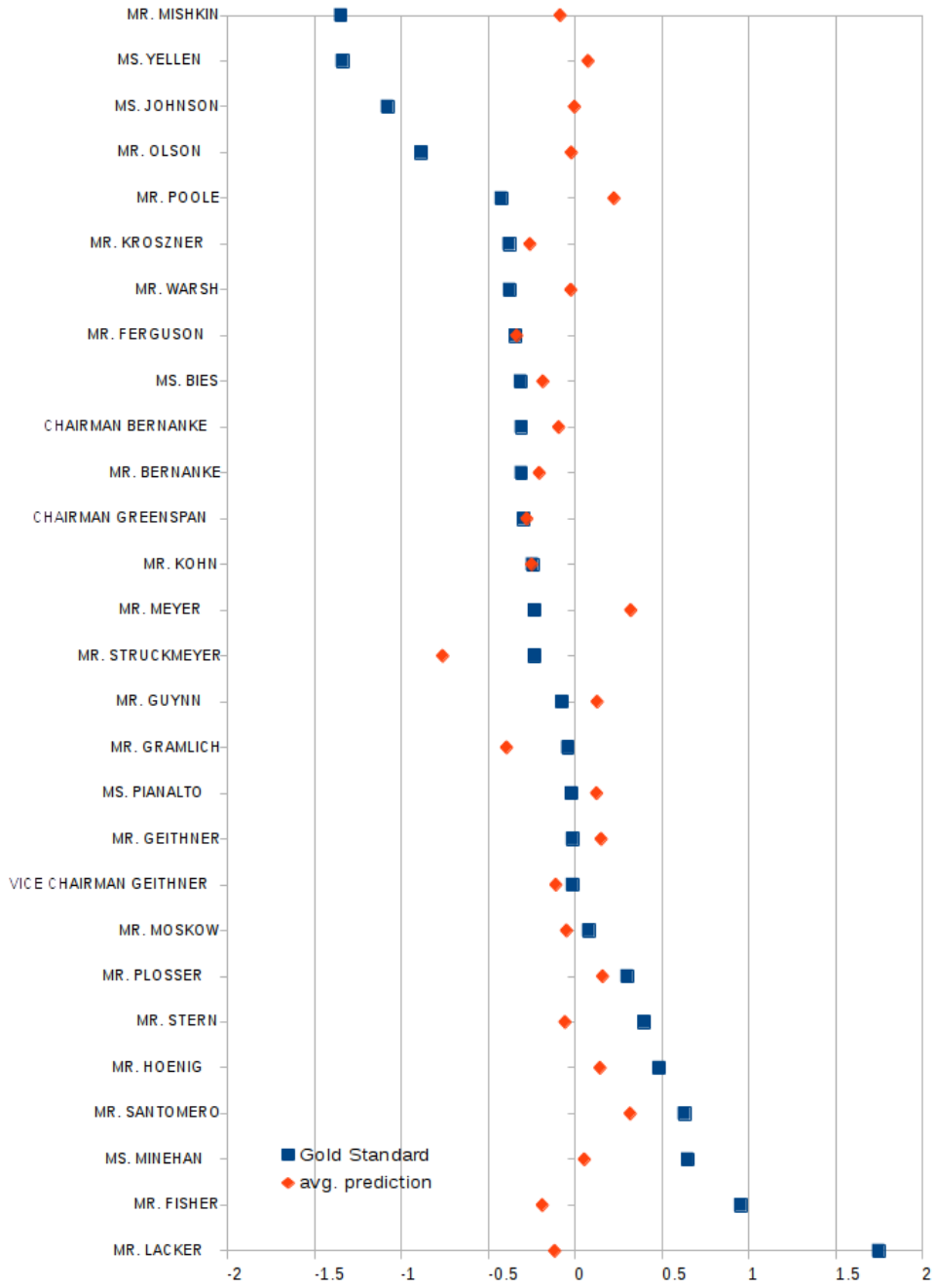


Figure 5.4: Averaged predicted positions for unseen FOMC members.
Dove (negative) / hawk (positive) values are displayed on the x-axis.

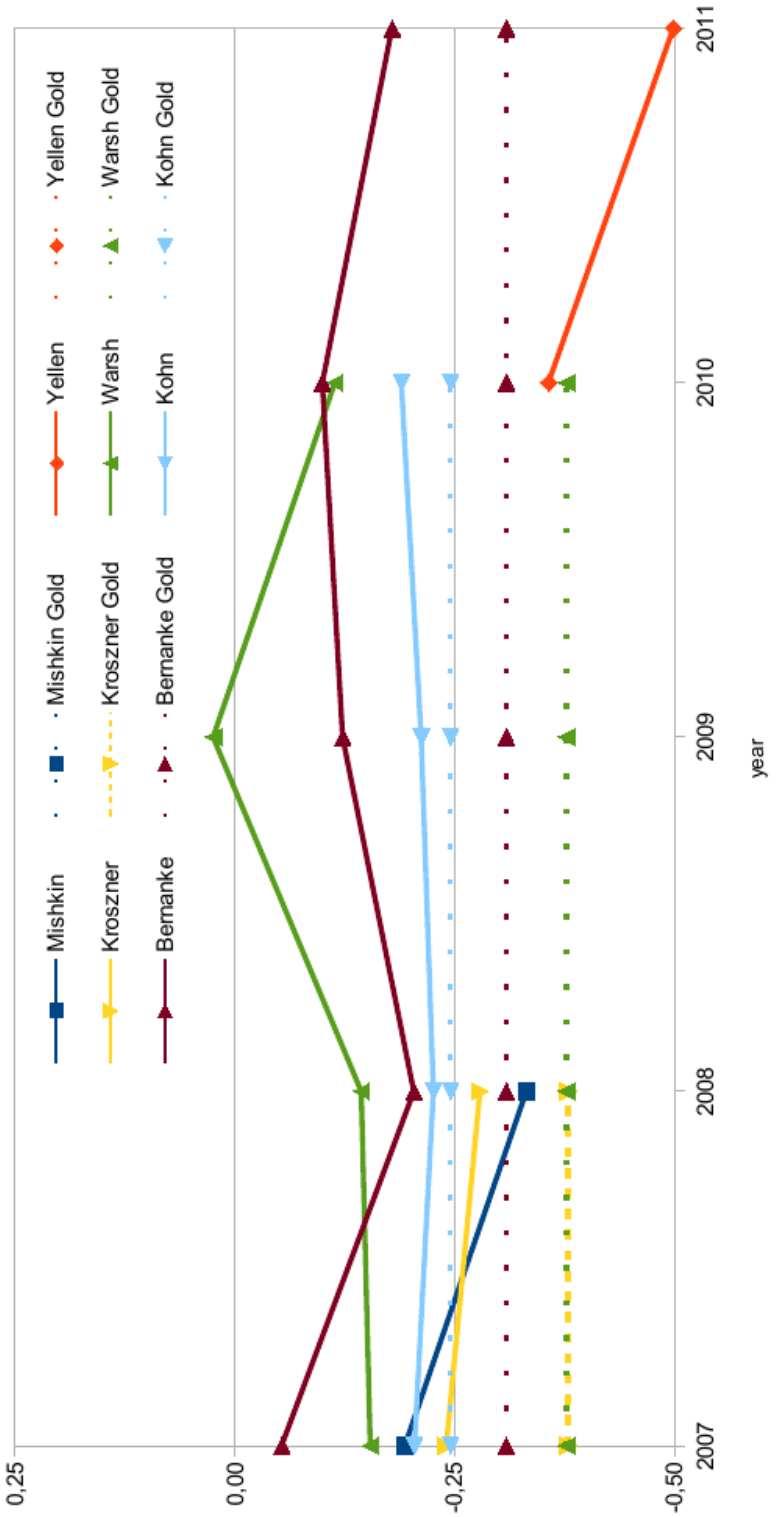


Figure 5.5: Predicted positions on speeches averaged over years. The dashed lines are the gold standard dove-hawk values.

6 Conclusions

In the previous three chapters, we presented our variegated approaches for the analysis of positions in political documents. At this point, we first briefly recap our results, before we discuss in detail the achievements with respect to our objectives.

6.1 Recapitulation of Results

We compared topic-specific positions of parties in a coalition in Chapter 3, which we verified by predicting ministries for the respective topics. Our approach made only 10 wrong predictions out of 54, which can be partly explained with exceptions in the process of appointing the ministers.

In Chapter 4, we first tackled the embedded task to avoid manual annotations. Our bootstrapping system to automatically label a corpus of party election manifestos with topics on sentence-level achieved F-measure scores of 77.5% (macro) and 79.5% (micro).

We further scaled positions of presidential candidates using the speeches they gave during the election campaigns. To check the soundness of our method, we first compared the general predicted positions of the candidates to the output of another system and confirm the accordance of the results. We then discussed the outcome for the topical positions of the candidates, aligning them with background information about the elections, which suggests plausible position predictions.

Finally, in Chapter 5, we predicted dove-hawk positions of members of the Federal Open Market Committee with a correlation of 83.1% and an absolute error of 0.33% given a gold-standard that was hand-coded by scholars of financial economics.

6.2 Discussion of Objectives

The vast majority of quantitative methods in political science that analyze positions in political texts represent a document as a bag-of-representation without modeling any further structure. Positions are then determined on document level, and they are represented as value on a continuous scale. In natural language processing, opinion mining is usually defined as the task of assigning texts to discriminant stance classes rather than a continuous value. Furthermore, those approaches are mostly adapted to common language documents authored by non-experts, which makes them hardly transferable to text produced by political agents.

In this work, we aim at developing methods that enable the analysis of actual research questions in political science by using methods of computational linguistics and computer science as well as linguistic knowledge. In the following, we will discuss the achievement of our objectives.

Computer Science Methods for Problems of Political Science. We investigate the applicability of state-of-the art models in computer science to research questions of political science. There are three salient requirements for these particular tasks. First, the methods need to cope with **highly domain-dependent, complex texts**. Second, the opinions that are determined need to be projected on a **continuous scale**, allowing for revealing fine-grained distances between positions. Third, there is a need for approaches that reach beyond text level and determine **multi-dimensional positions** in text, such as opinions on particular topics covered by the text. To our knowledge, in quantitative analysis there is a lack of methods addressing the third requirement.

In order to position the opinions on a continuous scale, we apply **Support Vector Regression** in our analysis of dove-hawk positions in the Federal Open Market Committee (cf. Chapter 5). Support Vector Regression is to a large extent identical with Support Vector Machines, which are commonly applied to classification problems in computational linguistics and have shown to achieve high performances when applied to text. The difference of the regression variant to the classification variant is the prediction of continuous values. Regression is a standard procedure in quantitative analysis. Traditionally they employ ordinary least squares (OLS) regression, however. The latter regression variant has the drawback of not being able to cope with a large number of features, which is the case when using word feature vectors. In our experiments, Support Vector Regression was able to cope with the amount of word features, had low runtimes, and, above all, it achieved high results at predicting dove-hawk positions.

To determine continuous positions of candidates of U.S. presidential election campaigns (cf. Chapter 4), we employed **Wordfish**, a commonly used tool in quantitative analysis for this type of task. The novelty contributed by us is regarding this task as a the multi-dimensional analysis of positions: we first determine topics addressed in the speeches, for which we then scale the positions. To detect the topics, we employ state-of-the art machine learning techniques using **Support Vector Machines**. The downside of this approach is that it requires a training corpus annotated with topics. We will get back on this issue later in this section when discussing the use of external knowledge.

Above all, we have shown that the methods we provide successfully measure multi-dimensional positions. Besides classifying topics with Support Vector Machines, as described above, we successfully employ variants of topic models to determine topics. While topic models have recently gained popularity in political science, their use was limited, as the standard approaches for topic models do not allow any influence on the creation of particular topics. We employed variants of topic models, namely **LogicLDA** and **LabeledLDA**, which provide the option to input keywords. The use of topic models enables us to analyze topic-dimensions instead of a single position per document. We apply this approach for the analysis of party manifestos (cf. Chapter 3). After extracting the topics from the manifestos with topic models, we measure the positions of the extracted topics relative to each other by calculating their similarities. To this end, we use **cosine similarity**, a widespread standard measure in computer science as well as political science.

Another technique from the field of computer science that we included are **Markov Logic Networks**. They allow for a global optimal classification of multiple objects in mutual dependence: in our case, the topic classification of all sentences for the best possible classification of an entire manifesto (cf. Section 4.3). Markov Logic Networks further provide the combination of various constraints with knowledge from various sources. To our knowledge, these kind of approaches have not been used before in the field of political science.

Informed Approaches. We developed methods to **combine various information sources** and **include external knowledge resources** for more informed approaches.

For the classification of topics in the speeches of the presidential election candidates, we used machine learning, which requires a training corpus containing topic annotations. As we head towards unsupervised approaches that do not require manual input, we intended to **use an existing, external resource** to train the classifier. We assumed that the corpus of party election programs provided by the Comparative Manifesto Project was suitable for this task, as

it contains topic annotations on sentence level. As only a small part of the corpus provided these labels, we bootstrapped a classifier to label the whole set of English manifestos. As the Comparative Manifesto Project provides annotated manifestos in manifold languages, it would serve as a valuable training set to compare topical positions across languages and nations. We therefore applied this semi-automatically annotated corpus to train the topic classifier in order to classify the topics in the presidential candidates' speeches. However, in this case, the use of an external resource to train the topic classifier did not provide the expected results and achieved a very low performance. We will further have to investigate how a classifier trained on external resources can be transferred to data of another domain.

In contrast, external knowledge was successfully employed to determine topics for our multi-dimensional position analysis in Chapter 3. **Descriptions of ministries served as definitions** of political topics. We extracted keywords from these descriptions, and inserted them as seeds for topics in topic models, which we then applied on the manifestos and coalition contracts.

Knowledge about Text Structure. The third objective of our work is to make use of knowledge about structure of texts, such as dialog structure or sequences of topical items.

Text knowledge is especially beneficial for two of our tasks: the bootstrapping process to create a training corpus in Section 4.3 as well as for the analysis of dove-hawk positions in the FOMC in Chapter 5.

When bootstrapping the training corpus, we create a classifier to label sentences of the manifestos with a topic. The sentences are short and thus contain hardly enough information for a classification with standard machine learning. However, we have some knowledge about text composition and pragmatics. Topics discussed in a text usually stretch over multiple sentences, only rarely the topic changes from sentence to sentence. Furthermore, particular sequences of topics are more probable than others. We encode this knowledge as constraints in our Markov Logic Network, which improves the classification of the sentences.

Likewise, we improve the prediction of dove-hawk positions by making use of knowledge about discourse structure. We revealed that in the discussions of the FOMC committee, there are two different types of contributions: On the one hand, there are statement turns, on the other hand, there are discussion turns. In the first type, the speakers elaborate on the current situation of the economy and make forecasts. In the latter type, the members comment on the other speakers' turns or ask questions. Consequently, the statement turns are more appropriate to reveal the positions of the actual speaker. We showed that if we omit the discussion turns, we remove noise form the regression. We assume that the discussion turns could as well

provide valuable information for the classification, as they contain information about agreement and disagreement among the speakers. However, as the language is on a highly domain dependent level, these turns rarely contain discourse structure markers that are usually used in the task of disagreement classification. If this information is accessed, it can be combined with the predictions of the regression on the statement turns in order to globally classify all speakers in a meeting in mutual dependence, just like in the corpus creation process described in Section 4.3.

6.3 Concluding Remarks

This work is a contribution to fill the gap in between opinion mining on political texts in the area of natural language processing and the quantitative analysis in the field of political science. We show the success of more fine-grained analyses compared to previous work by combining methods of computer science, knowledge about text and the inclusion of external resources.

While in the early stages of natural language processing the use of manually created knowledge bases was state of the art, with the development of hardware and algorithms finally learning based methods became popular - first supervised and later also unsupervised approaches. Recently, research is heading towards combining learning approaches with external knowledge, as they provide a way to replace manual supervision by using external resources. Also, by the development of crowd sourcing, more and more huge knowledge bases become available. The presented work shows that opinion analysis of political texts benefits from this development.

Acknowledgements

I would like to thank all the people that supported me on my way to finishing this thesis.

First of all, I am grateful to my advisors Heiner Stuckenschmidt and Michael Strube, who were supportive and helpful at any point. Heiner, you let me freedom for my work and provided me autonomy, yet you where guiding me whenever necessary. Michael, you are one of the three Michaels that had such an impact on my life. Whenever I was stuck at some point, after a discussion with you I knew how to proceed - especially because you have the gift of listening, understanding and asking the right questions.

The second Michael I want to thank is Michael Schäfer. Micha, do you remember when I had to code my first homework, bubble sort? I will never forget your patience when you sat next to me at CN-WG, always pointing me towards the next step, without just telling the solution, so that at the end I had implemented it all by myself. Basically, you taught me programming that night. There followed so many other things to thank you for - thanks. (@)
PS.: Now, I'm waiting for your thesis. But I see, you just started another life-time project...

Steffi, I thank you so much for being with me during the last years. And for over and over reminding me where it is where we are! Pssst ... we both know it. And I hope we both won't ever forget it.

Thank you, Jörg, for being such a good friend and having been there for me during the darkest hours. I'm not sure if there was a thesis without you. The same applies to you, Domi. Thank you for being my BFF and for being there for me whenever I need you. And sorry that I *"left you holding the baby"*, which apparently means *jemandem den schwarzen Peter zuschieben*, according to Linguee ;)

And thanks to Linguee!!!

I'd further like to thank my Monnem Family, for being my anchor and making me feeling at home here.

Special thanks to the man I met right in the middle of the crazy last period of my PhD, and who still liked who I am, and supported me in any possible way. Dankeeeee Thomas!!!

The third Michael I want to thank is Büro-Michi. You brought so much fun to the office (and to Mannheim!), and the paper we wrote together was the most pleasant research so far :)

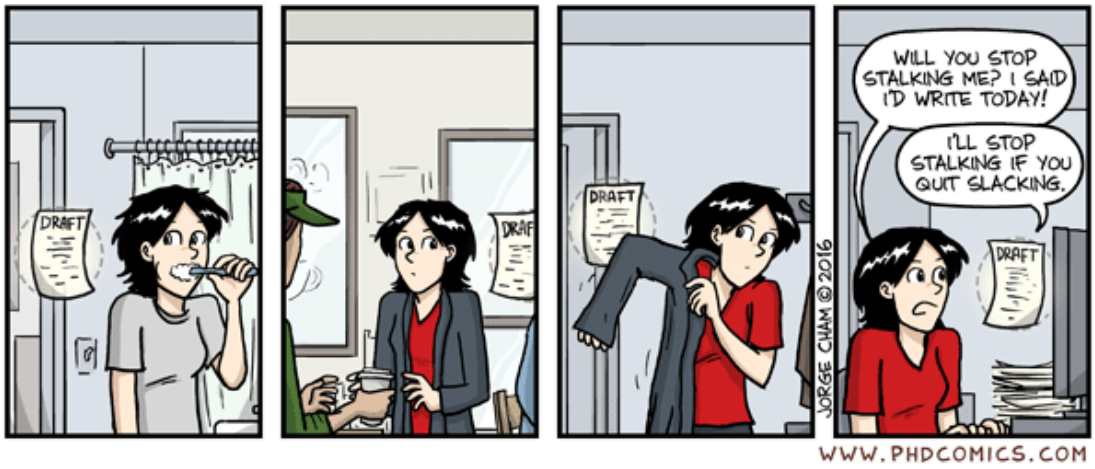
Thanks to the CoLiktiv Heidelberg, for raising the computational linguist in me and being such a great environment for studying.

Nearly last, but far from least: Mama. I thank you with all of my heart for everything that you have ever done for me and for what you are - HDL!

And furthermore, someone told me the joke: *Wenn du 30 bist, und noch keinen Doktor hast, musst du ihn halt selber machen!*
I stuck to this advice. And I'd like to extend it: if you provided to your children the possibilities to make a PhD, you deserve to be proud if they do.

Finally, I want to thank my home-office team Samson and Sushi, and the Gruffelo Catering GmbH as well as Ben&Jerry's. And thanks to all the other people who accompanied me on the way, but who I cannot list in detail - as you can see, I'm running out of space, and out of time, I have to print this thing soon. Thank you all!

... and Tim, by the way, how far is your thesis?!?



"Piled Higher and Deeper" (PhD), 04/01/2016

Well, I guess it's time to say goodbye, then ...

7
Bibliography

[Abelson and Carroll, 1965] Abelson, R. P. and Carroll, J. D. (1965). Computer simulation of individual belief systems. *The American Behavioral Scientist*, 8(9):0–24.

[Adolph, 2013] Adolph, C. (2013). *Bankers, Bureaucrats, and Central Bank Politics: The Myth of Neutrality*. Cambridge University Press.

[Altman, 2009] Altman, R. C. (2009). The great crash, 2008: a geopolitical setback for the west. *Foreign Affairs*, 88(1):2–14.

[Anand et al., 2011] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats Rule and Dogs Drool!: Classifying Stance in Online Debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA'11, pages 1–9. Association for Computational Linguistics.

[Andrzejewski et al., 2011] Andrzejewski, D., Zhu, X., Craven, M., and Recht, B. (2011). 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'11, pages 1171–1177. AAAI Press.

[Balasubramanyan et al., 2012] Balasubramanyan, R., Cohen, W. W., Pierce, D., and Redlawsk, D. P. (2012). Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News? In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM'12, pages 18–25. AAAI Press.

[Benoit et al., 2009] Benoit, K., Mikhaylov, S., and Laver, M. (2009). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, 53(2):495–513.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

[Blinder et al., 2008] Blinder, A. S., Ehrmann, M., Fratzscher, M., Haan, J. D., Blinder, A. S., Ehrmann, M., Fratzscher, M., Haan, J. D., and jan Jansen, D. (2008). Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence. *Journal of Economic Literature*, 46(4):910–945.

[Broomhead and Lowe, 1988] Broomhead, D. S. and Lowe, D. (1988). *Radial Basis Functions, Multi-variable Functional Interpolation and Adaptive Networks*. RSRE memoran-

dum / Royal Signals and Radar Establishment. Royals Signals & Radar Establishment.

[Burfoot, 2008] Burfoot, C. (2008). Using Multiple Sources of Agreement Information for Sentiment Classification of Political Transcripts. In *Proceedings of the Australasian Language Technology Association Workshop*, volume 6 of *ALTA'08*, pages 11–18.

[Caruana and Niculescu-Mizil, 2006] Caruana, R. and Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine learning*, ICML'06, pages 161–168. ACM Press.

[Casella and George, 1992] Casella, G. and George, E. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

[Demiralp and Jorda, 2002] Demiralp, S. and Jorda, O. (2002). The Announcement Effect: Evidence from Open Market Desk Data. *Economic Policy Review*, (May):29–48.

[Diermeier et al., 2007] Diermeier, D., Godbout, J.-F., Yu, B., and Kaufmann, S. (2007). Language and Ideology in Congress. In *Proceedings of the Midwest Political Science Association 65th Annual National Conference*, MPSA'07, pages 1–25.

[Ehrmann and Fratzscher, 2007] Ehrmann, M. and Fratzscher, M. (2007). Communication by Central Bank Committee Members: Different Strategies, Same Effectiveness? *Journal of Money, Credit and Banking*, 39(2-3):509–541.

[Eijffinger et al., 2015] Eijffinger, S. C., Mahieu, R., and Raes, L. (2015). Hawks and Doves at the FOMC. *CentER Discussion Paper Series*, (2015-013).

[Foot et al., 2003] Foot, K., Schneider, S. M., Dougherty, M., Xenos, M., and Larsen, E. (2003). AAnalyzing Linking Practices: Candidate Sites in the 2002 US Electoral Web Sphere. *Journal of Computer-Mediated Communication*, 8(4).

[Fracasso et al., 2003] Fracasso, A., Genberg, H., and Wyplosz, C. (2003). *How do Central Banks Write?: An Evaluation of Inflation Reports by Inflation Targeting Central Banks*, volume 2. Centre for Economic Policy Research.

[Galley et al., 2004] Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL'04, pages 669–676. Association for Computational Linguistics.

[Gottipati et al., 2013] Gottipati, S., Qiu, M., Sim, Y., Jiang, J., and Smith, N. A. (2013). Learning Topics and Positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 1858–1868. Association for Computational Linguistics.

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

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

[Gross et al., 2013] Gross, J., Acree, B., Sim, Y., and Smith, N. A. (2013). Testing the Etch-a-Sketch Hypothesis: A Computational Analysis of Mitt Romney's Ideological Makeover During the 2012 Primary vs. General Elections. In *Proceedings of the 2013 Annual Meeting of the American Political Science Association*, APSA'13.

[Hasan and Ng, 2013] Hasan, K. S. and Ng, V. (2013). Extra-Linguistic Constraints on Stance Recognition in Ideological Debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL'13, pages 816–821. Association for Computational Linguistics.

[Havrilesky and Gildea, 1991] Havrilesky, T. and Gildea, J. (1991). The Policy Preferences of FOMC Members as Revealed by Dissenting Votes: Comment. *Journal of Money, Credit and Banking*, 23(1):130–138.

[Hearst, 1997] Hearst, M. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, 23(1):33–64.

[Hillard et al., 2007] Hillard, D., Purpura, S., and Wilkerson, J. (2007). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology and Politics*, 4(4):31–46.

[Hirst et al., 2014] Hirst, G., Riabinin, Y., Graham, J., and Boizot-Roche, M. (2014). Text to Ideology or Text to Party Status? In Kaal, B., Maks, E. I., and van Elfrinkhof, A. M., editors, *From Text to Political Positions: Text analysis across disciplines*, pages 93–115. John Benjamins Publishing Company.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 289–296. Morgan Kaufmann Publishers Inc.

[Høyland et al., 2014] Høyland, B., Godbout, J.-F., Lapponi, E., and Velldal, E. (2014). Predicting Party Affiliations from European Parliament Debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, LTCSS'14, pages 56–60. Association for Computational Linguistics.

[Hsu et al., 2003] Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A Practical Guide to Support Vector Classification. Technical report, Department of Computer Science National Taiwan University, Taipei.

[Jansen, 2011] Jansen, D.-J. (2011). Does the Clarity of Central Bank Communication Affect Volatility in Financial Markets? Evidence from Humphrey-Hawkins Testimonies. *Contemporary Economic Policy*, 29(4):494–509.

[Jansen and de Haan, 2004] Jansen, D.-J. and de Haan, J. (2004). Look who's talking: Ecb communication during the first years of emu. DNB Working Papers, Netherlands Central

Bank, Research Department.

[Jansen and de Haan, 2005] Jansen, D.-J. and de Haan, J. (2005). Talking heads: the effects of ecb statements on the euro-dollar exchange rate. *Journal of International Money and Finance*, 24(2):343–361.

[Jiang and Silver, 2014] Jiang, X. and Silver, D. L. (2014). A Survey of Transfer Learning in Deep Learning Architectures. In *Proceedings of the Science Atlantic Mathematics, Statistics and Computer Science Conference 2014*.

[Joachims, 2002] Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines*. Kluwer Academic Publishers.

[Keman, 2007] Keman, H. (2007). Experts and manifestos: Different sources – Same results for comparative research? *Electoral Studies*, 26(1):76–89.

[Klüver, 2009] Klüver, H. (2009). Measuring Interest Group Influence Using Quantitative Text Analysis. *European Union Politics*, 10(4):535–549.

[Knabb et al., 2005] Knabb, R. D., Rhome, J. R., and Brown, D. P. (2005). *Tropical Cyclone Report Hurricane Katrina 23-30 August 2005*. National Hurricane Center.

[Kogan et al., 2009] Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression . In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL'09, pages 272–280. ACM Press.

[Kohn and Sack, 2003] Kohn, D. L. and Sack, B. P. (2003). Central Bank Talk: Does It Matter and Why?*. Finance and Economics Discussion Series 2003-55, Board of Governors of the Federal Reserve System (U.S.).

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

[Laver and Garry, 2000] Laver, M. and Garry, J. (2000). Estimating Policy Positions from Political Texts. *American Journal of Political Science*, 44(3):619–634.

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

[Lowe, 2011] Lowe, W. (2011). 'JFreq: Count words, quickly'. Java software version 0.5.4.

[Lowe, 2015] Lowe, W. (2015). Austin: Do things with words. Version 0.2.2.

[Lu et al., 2012] Lu, Y., Wang, H., Zhai, C., and Roth, D. (2012). Unsupervised Discovery of Opposing Opinion Networks From Forum Discussions . In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM'12, pages 1642–1646. ACM Press.

[Malouf and Mullen, 2008] Malouf, R. and Mullen, T. (2008). Taking sides: User classification for informal online political discours. *Internet Research*, 18(2):177–190.

[Marschall and McKee, 2002] Marschall, M. J. and McKee, R. J. (2002). From Campaign Promises to Presidential Policy: Education Reform in the 2000 Election. *Educational Policy*, 16(1):96–117.

[Masket, 2009] Masket, S. (2009). *No Middle Ground: How Informal Party Organizations Control Nominations and Polarize Legislatures*. University of Michigan Press.

[Mcauliffe and Blei, 2008] Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Proceedings of the 20th Advances in Neural Information Processing Systems Conference*, NIPS'07, pages 121–128. Curran Associates, Inc.

[Meade, 2005] Meade, E. E. (2005). The FOMC: Preferences, Voting, and Consensus. *Federal Reserve Bank of St. Louis Review*, 87(2):93–101.

[Medzihorsky et al., 2014] Medzihorsky, J., Littvay, L., and Jenne, E. K. (2014). Has the Tea Party Era Radicalized the Republican Party? Evidence from Text Analysis of the 2008 and 2012 Republican Primary Debates. *PS: Political Science & Politics*, 47(04):806–812.

[Musard-Gies, 2006] Musard-Gies, M. (2006). Do ECB's Statements Steer Short-Term and Long-Term Interest Rates in the Euro Zone? Money macro and finance (mmf) research group conference 2005, Money Macro and Finance Research Group.

[Nanni and Fabo, 2016] Nanni, F. and Fabo, P. R. (2016). Entities as topic labels: Improving topic interpretability and evaluability combining Entity Linking and Labeled LDA. In *Proceedings of the 27th Digital Humanities Conference*, DH'16.

[Nanni et al., 2016] Nanni, F., Zirn, C., Glavaš, G., Eichorst, J., and Ponzetto, S. (2016). Topfish: Topic- based analysis of political position in us electoral campaigns. In *International Conference on the Advances in Computational Analysis of Political Text*.

[Noessner et al., 2013] Noessner, J., Niepert, M., and Stuckenschmidt, H. (2013). RockIt: Exploiting Parallelism and Symmetry for MAP Inference in Statistical Relational Models. In *Proceedings of the AAAI Workshop: Statistical Relational Artificial Intelligence*, StaRAI'13. AAAI Press.

[Pan and Yang, 2010] Pan, S. J. and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, EMNLP'02, pages 79–86. Association for Computational Linguistics.

[Pappi et al., 2011] Pappi, F., Seher, N., and Kurella, A.-S. (2011). 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).

[Pappi and Seher, 2009] Pappi, F. U. and Seher, N. M. (2009). Party election programmes, signalling policies and salience of specific policy domains: The german parties from 1990

to 2005. *German Politics*, 18(3):403–425.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, pages 1532–1543. Association for Computational Linguistics.

[Poole, 2001] Poole, W. (2001). Expectations : The Twenty-Second Henry Thornton Lecture, Department of Banking and Finance, City University Business School, London, England. *Federal Reserve Bank of St. Louis Review*, 83(2):1–10.

[Prabhakaran et al., 2014] Prabhakaran, V., Arora, A., and Rambow, O. (2014). Staying on Topic: An Indicator of Power in Political Debates. In *Proceedings of the 19th Conference on Empirical Methods in Natural Language Processing*, EMNLP'14, pages 1481–1486.

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

[Ramage et al., 2009] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 248–256. Association for Computational Linguistics.

[Ranade et al., 2013] Ranade, S., Sangal, R., and Mamidi, R. (2013). Stance Classification in Online Debates by Recognizing Users' Intentions. In *Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue*, SIGDIAL'13, pages 61–69. Association for Computational Linguistics.

[Rao et al., 2013] Rao, D., McNamee, P., and Dredze, M. (2013). Entity Linking: Finding Extracted Entities in a Knowledge Base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer.

[Raschka, 2016] Raschka, S. (2016). When Does Deep Learning Work Better Than SVMs or Random Forests?

[Richardson and Domingos, 2006] Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1-2):107–136.

[Ringquist and Dasse, 2004] Ringquist, E. J. and Dasse, C. (2004). Lies, Damned Lies, and Campaign Promises? Environmental Legislation in the 105th Congress. *Social Science Quarterly*, 85(2):400–419.

[Rosa and Verga, 2007] Rosa, C. and Verga, G. (2007). On the consistency and effectiveness of central bank communication: Evidence from the ECB. *European Journal of Political Economy*, 23(1):146–175.

[Rosen-Zvi et al., 2004] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI'04, pages 487–494. AUAI Press.

[Rozkrut et al., 2007] Rozkrut, M., Rybiński, K., Sztaba, L., and Szwaja, R. (2007). Quest for Central Bank Communication. Does It Pay To Be "Talkative"? *European Journal of Political Economy*, 23(1):176–206.

[Scharl and Weichselbraun, 2008] Scharl, A. and Weichselbraun, A. (2008). An Automated Approach to Investigating the Online Media Coverage of US Presidential Elections. *Journal of Information Technology & Politics*, 5(1):121–132.

[Schmid, 1999] Schmid, H. (1999). *Improvements in Part-of-Speech Tagging with an Application to German*, pages 13–25. Springer Netherlands, Dordrecht.

[Schölkopf et al., 2000] Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New Support Vector Algorithms. *Neural computation*, 12(5):1207–1245.

[Schonhardt-Bailey, 2013] Schonhardt-Bailey, C. (2013). *Deliberating American Monetary Policy: A Textual Analysis*. MIT Press.

[Seher and Pappi, 2011] Seher, N. and Pappi, F. (2011). Politikfeldspezifische positionen der landesverbände der deutschen parteien. Working Paper 139, Mannheimer Zentrum für Europäische Sozialforschung (MZES).

[Silver, 2012] Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't*. Penguin Press.

[Sim et al., 2013] Sim, Y., Acree, B. D. L., Gross, J. H., and Smith, N. A. (2013). Measuring Ideological Proportions in Political Speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP'13, pages 91–101, Seattle, Washington, USA. Association for Computational Linguistics.

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

[Smola, 1996] Smola, A. J. (1996). Regression estimation with support vector learning machines. Master's thesis, Technische Universität München.

[Somasundaran and Wiebe, 2009] Somasundaran, S. and Wiebe, J. (2009). Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, ACL'09, pages 226–234. Association for Computational Linguistics.

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

[Thomas et al., 2006] Thomas, M., Pang, B., and Lee, L. (2006). Get out the Vote: Determining Support or Opposition from Congressional Floor-debate Transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP'06, pages 327–335. Association for Computational Linguistics.

[Trabelsi and Zaıane, 2014] Trabelsi, A. and Zaıane, O. R. (2014). Finding Arguing Expressions of Divergent Viewpoints in Online Debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media*, LASM'14, pages 35–43. Association for Computational Linguistics.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc.

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

[Volkens et al., 2015] Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., and Werner, A. (2015). *The Manifesto Data Collection Version 2015a*. Manifesto Project (MRG/CMP/MARPOR), Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin.

[Wang and Cardie, 2014] Wang, L. and Cardie, C. (2014). Improving Agreement and Disagreement Identification in Online Discussions with A Socially-Tuned Sentiment Lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA'14. Association for Computational Linguistics.

[Werner et al., 2014] Werner, A., Lacewell, O., and Volkens, A. (2014). Manifesto Coding Instructions: 5th fully revised edition.

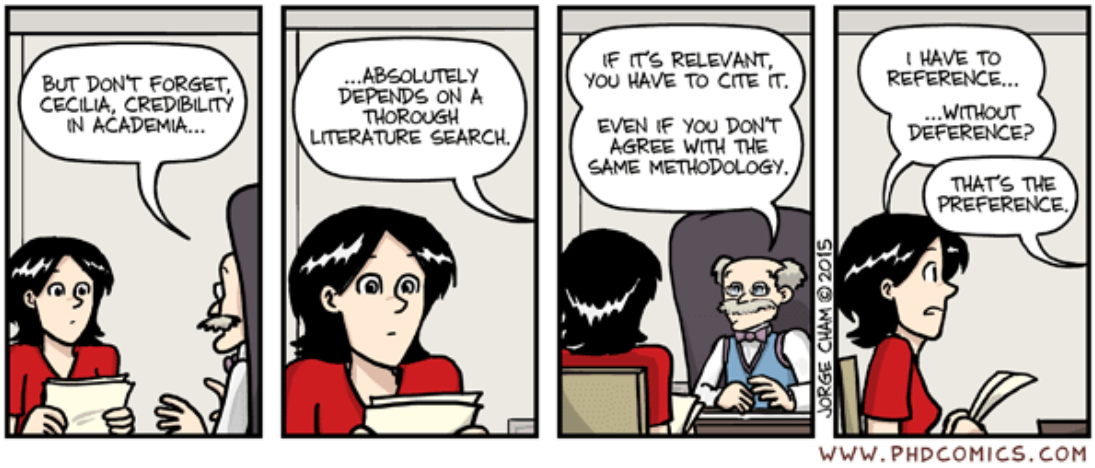
[Yano et al., 2012] Yano, T., Smith, N. A., and Wilkerson, J. D. (2012). Textual Predictors of Bill Survival in Congressional Committees. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT'12, pages 793–802. Association for Computational Linguistics, Association for Computational Linguistics.

[Zirn, 2014] Zirn, C. (2014). Analyzing positions and topics in political discussions of the german bundestag. In *ACL (Student Research Workshop)*, pages 26–33.

[Zirn et al., 2016] Zirn, C., Glavaš, G., Nanni, F., Eichorst, J., and Stuckenschmidt, H. (2016). Classifying topics and detecting topic shifts in political manifestos. In *International Conference on the Advances in Computational Analysis of Political Text*. PolText.

[Zirn et al., 2014] Zirn, C., Schäfer, M., Strube, M., Ponzetto, S. P., and Stuckenschmidt, H. (2014). Exploring structural features for position analysis in political discussions. In *Paper entry to the 2014 NLP Unshared Task in PoliInformatics*. PoliInformatics Research Coordination Network (PInet).

[Zirn and Stuckenschmidt, 2014] Zirn, C. and Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53.



"Piled Higher and Deeper" (PhD), 10/05/2015