Data & Knowledge Engineering 90 (2014) 38–53

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

Multidimensional topic analysis in political texts

Cäcilia Zirn [*], Heiner Stuckenschmidt

Research Group Data and Web Science, University of Mannheim, B6 26, 68159 Mannheim, Germany

ARTICLE INFO

Article history:
Received 1 June 2013
Accepted 1 July 2013
Available online 6 August 2013

Keywords:
Topic models
Political science
Text Analysis

ABSTRACT

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

1. Introduction

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

[*] Corresponding author. Tel.: +49 6211812650.
E-mail addresses: caecilia@informatik.uni-mannheim.de (C. Zirn), heiner@informatik.uni-mannheim.de (H. Stuckenschmidt).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.07.003

position towards this topic. We conclude that there is a need for methods that allow for position analysis based on multi-topic documents that takes these different topics into account.

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

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

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

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

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

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

We test the capability of our approach in two different scenarios. In the first experiment described in Section 3, we show that the method can be used to predict the distribution of ministries between the parties of a winning coalition based on the distance of the positions extracted from their manifestos to the positions in the coalition agreement. We explain the rationale of this experiment in more detail later on. We also show that although of course the result of the analysis depends on the choice of the seed words, the general principle works independently from a specific set of keywords. We compare the method to a baseline that simulated a manual approach to the problem where individual sentences are assigned to a topic based on keywords and sentences assigned to the same topic are compared. We show that our method consistently outperforms this baseline with respect to the task of predicting the assignment of the ministry. We will further investigate the impact of specific Latent Dirichlet Allocation (LDA) extensions, the seed set and the words included in the analysis.

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

2. Multi-dimensional analysis

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

2.1. Data preparation

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