

Twitter in the Parliament - A Text-based Analysis of German Political Entities

Patrick Schulze, Simon Wiegrebe

Supervisors:

Prof. Dr. Christian Heumann, Prof. Dr. Paul W. Thurner

7. Juli 2020

Outline

- Introduction
- Topic Modeling: Motivation and Theory
- Data: Collection and Preprocessing
- Model Selection and Global Characteristics
- Metadata Analysis
- Causal Inference
- Discussion

Introduction

- Rise in popularity of social media is producing huge amounts of data, especially text.
- *Politics* is a field of particular interest in the context of social media and big data (Brexit, 2016 presidential election in the US, Facebook data scandal).
- Simultaneously, advances in *Natural Language Processing* (NLP) are providing tools of analysis for such data.
- One instance of such analysis is the discovery and exploration of latent thematic clusters within text - *topic analysis*.
- In this project, we apply the *Structural Topic Model* (STM) to a self-created dataset containing Twitter posts by members of the German Bundestag (and a variety of metadata)

Topic Modeling: Motivation and Theory

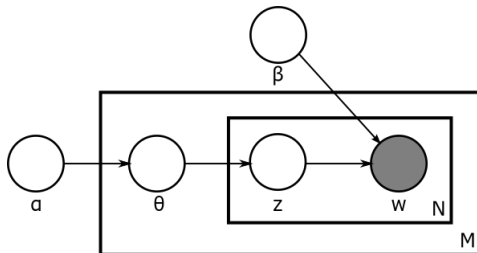
Notation and Terminology

- A *word* is an instance of a vocabulary of V unique *terms*.
- A *document* $d \in \{1, \dots, D\}$ is a sequence of words of length N_d . The n -th word of document d is denoted by $w_{d,n}$.
- A *corpus* is a collection (or set) of D documents. Therefore, $d \in \{1, \dots, D\}$ means that our corpus contains D documents.
- A *topic* $k \in \{1, \dots, K\}$ is a latent thematic cluster within a text corpus. That is, we imply a corpus can be represented by K topics.
- A *topic-word distribution* β is a probability distribution over words. We denote the word distribution corresponding to the k -th topic by β_k .
- A *topic assignment* $\mathbf{z}_{d,n}$ assign $w_{d,n}$ to a specific topic $k \in \{1, \dots, K\}$. We represent the word distribution for $w_{d,n}$ as $\beta_{d,n}$.
- *Topic proportions* θ_d are the proportions of the document d 's terms assigned to each of the topics. $\sum_{k=1}^K \theta_{d,k} = 1$, for all $d \in \{1, \dots, D\}$.

Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA)

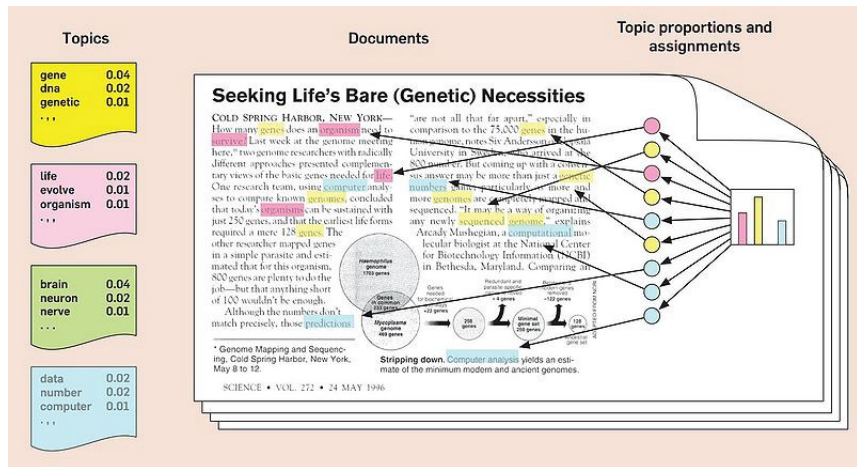
- LDA by **blei2003latent** is the first probabilistic topic model.
- Its generative process for each document $d \in \{1, \dots, D\}$ is:
 - ① Draw topic proportions $\theta_d \sim \text{Dir}_K(\alpha)$.
 - ② For each word $n \in \{1, \dots, N_d\}$:
 - a) Draw a topic assignment $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$.
 - b) Draw a word $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$.
- Graphical model representation (**roberts2016model**, p. 990):



Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA)

- The topic assignment of a document's words can be illustrated as follows:



Data

Data Collection

- MP-level data was scraped from `www.bundestag.de/abgeordnete` using Python's *BeautifulSoup* and a *selenium web driver*

Philipp Amthor, CDU/CSU
Jurist

CDU/CSU
Fraktion im Deutschen Bundestag

Abgeordnetenbüro
Deutscher Bundestag
Platz der Republik 1
11011 Berlin
Kontakt

Profil im Internet
phillip-amthor.de
Facebook

Biografie Reden Abstimmungen

Geboren am 10. November 1962 in Ueckermünde.
2011 Abitur am Greifen-Gymnasium Ueckermünde. 2012 bis 2017 Studium der Rechtswissenschaften an der Ernst-Moritz-Arndt-Universität Greifswald (Studienabschluss mit Prädikat). Stipendiat der Konrad-Adenauer-Stiftung. Kollegiat am Jungen Kolleg des Alfred Krupp Wissenschaftskollegs, nebenberuflich u.a. Mitarbeiter verschiedener Abgeordneter des Deutschen Bundestages und des Landtages Mecklenburg-Vorpommern; seit 2017 Doktorand und wissenschaftlicher Mitarbeiter an der Ernst-Moritz-Arndt-Universität Greifswald und zugleich Mitarbeiter einer internationalen Wirtschaftskanzlei in Berlin.
Seit 2008 Mitglied der CDU und der Jungen Union; seit 2010 Mitglied im Landesvorstand der Jungen Union Mecklenburg-Vorpommern; seit 2012 Kreisvorsitzender der Jungen Union Vorpommern-Greifswald; seit 2014 Mitglied des Sozialausschusses des Kreistages Vorpommern-Greifswald; seit 2017 Vorsitzender des CDU-Stadtverbandes Ueckermünde.

Direkt gewählt

Mecklenburg-Vorpommern
> Wahlkreis 016: Mecklenburgische Seenplatte I – Vorpommern-Greifswald II

Mitgliedschaften und Ämter im Bundestag

Ordentliches Mitglied
> Ausschuss für die Angelegenheiten der Europäischen Union
> Ausschuss für Innere und Heimat

Stellvertretendes Mitglied
> Ausschuss für Recht und Verbraucherschutz

Veröffentlichungspflichtige Angaben

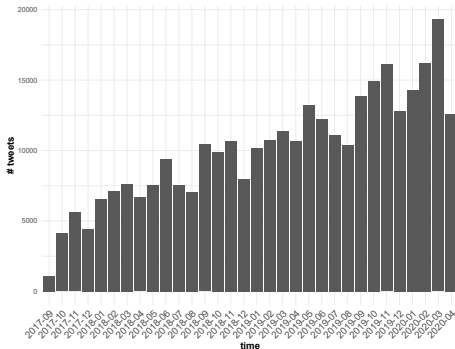
Biografie drucken

- Socioeconomic data and 2017 German federal election results were extracted from `www.bundeswahlleiter.de`.

Data

Data Collection

- Tweets (and further Twitter features) were downloaded via the official Twitter API using Python's *tweepy* library.
- Monthly tweets (after dropping MPs without electoral district) for our period of analysis, September 24, 2017 through April 24, 2020:



- Henceforth, we grouped each MP's tweets on a monthly basis.

Data

Data Preprocessing

- For preprocessing, we used the *quanteda* package in R.
- We immediately transcribed German umlauts (ä, ö, ü) and ligature (ß) and removed hyphens due to the presence of compound words in the German language (*Corona-Krise* vs *Coronakrise*).
- Next, we transformed the text data into a document-feature matrix (DFM), converted all characters to lowercase, and removed stopwords, units, interjections, etc.
- Finally, we performed word stemming, which cuts off word endings to remove discrepancies arising purely from declensions or conjugations (e.g., *politisch* → *polit*).

Results

- Hyperparameter search yields 15 distinct topics
- Topic labeling conducted manually (human judgment)
- Descriptive discussion of relationship between metadata and topics
- Causal inference: estimation of cause-effect relationships between document-specific features (e.g. political party) and topics

Bibliography