

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

Introduction

- Huge amounts of data, especially text, produced by social media
- Field of particular interest in the context of social media and big data: *Politics* (e.g., Brexit, 2016 presidential election in the US, Facebook data scandal).
- Tools of analysis for such data simultaneously provided by advances in *Natural Language Processing* (NLP)
- *topic analysis*: analytical tool for discovery and exploration of latent thematic clusters within text
- In this project: application of the *Structural Topic Model* (STM) **roberts2016model** 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 (I)

- *Words* w : instances of a vocabulary of V unique *terms*.
- *Documents* $d \in \{1, \dots, D\}$: sequences of words of length N_d ; $w_{d,n}$ denoting n -th word of document d
- *Corpus*: collection (or set) of D documents
- *Topics* $k \in \{1, \dots, K\}$: latent thematic clusters within a text corpus; (implicit) representation of a corpus
- *Topic-word distributions* β : probability distributions over words; β_k denoting the word distribution corresponding to the k -th topic

Topic Modeling: Motivation and Theory

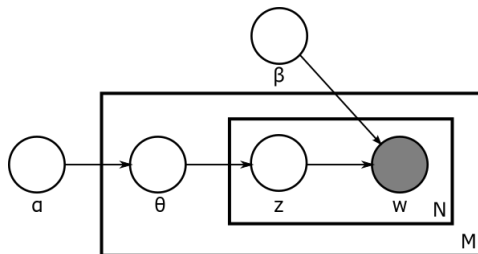
Notation and Terminology (II)

- *Topic assignments* $\mathbf{z}_{d,n}$: assignment of $w_{d,n}$ to a specific topic $k \in \{1, \dots, K\}$; $\beta_{d,n}$ representing the (assigned) word distribution for $w_{d,n}$
- *Topic proportions* θ_d : proportions of document d 's terms assigned to each of the topics; $\sum_{k=1}^K \theta_{d,k} = 1$, for all $d \in \{1, \dots, D\}$
- *Bag-of-words* assumption: only words themselves meaningful, unlike word order or grammar; equivalent to assuming *exchangeability* **aldous1985exchangeability**.

Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA) (I)

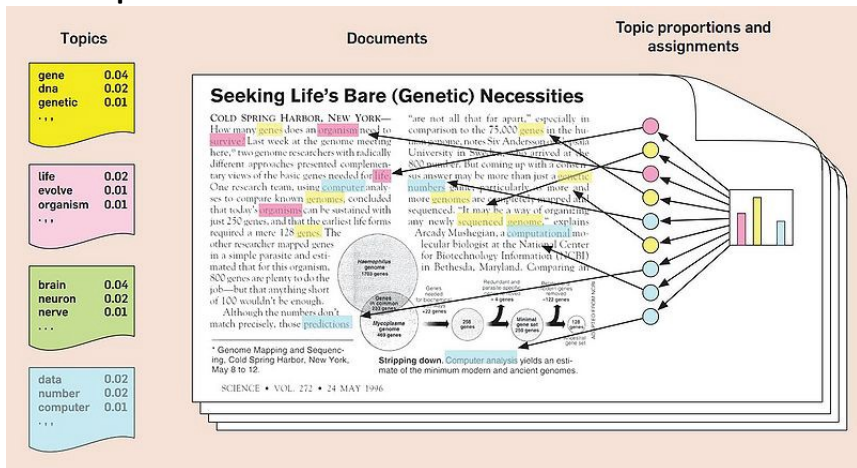
- First topic model with entirely probabilistic generating process: LDA **blei2003latent**
- Generative process for each document $d \in \{1, \dots, D\}$:
 - 1 Draw topic proportions $\theta_d \sim \text{Dir}_K(\alpha)$.
 - 2 For each word $n \in \{1, \dots, N_d\}$:
 - a Draw a topic assignment $z_{d,n} \sim \text{Multinomial}_K(\theta_d)$.
 - b Draw a word $w_{d,n} \sim \text{Multinomial}_V(\beta_{d,n})$.
- Graphical model representation of LDA: **blei2003latent**



Topic Modeling: Motivation and Theory

Latent Dirichlet Allocation (LDA) (II)

- Illustration of topic assignment for the words of a document:
blei2012probabilistic



Data

Data Collection (I)

- MP-level data: from www.bundestag.de/abgeordnete using Python's *BeautifulSoup* and a *selenium* web driver **van1995python richardson2007beautiful**

The screenshot shows the official profile of Philipp Amthor, a Member of Parliament (MP) for the CDU/CSU. The page includes a portrait photo, his name and affiliation, and contact information. It also features a 'Biografie' (Biography) section with details about his education and career, a map of his constituency (Mecklenburg-Vorpommern), and a list of his committee assignments in the Bundestag.

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 Greifenhof-Gymnasium Ueckermünde; 2012 bis 2017 Studium der Rechtswissenschaften an der Ernst-Moritz-Arzt Universität Greifswald (Studienabschluss mit Prädikat); Stipendiat der Konrad-Adenauer-Stiftung; Kollegat 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-Arzt-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

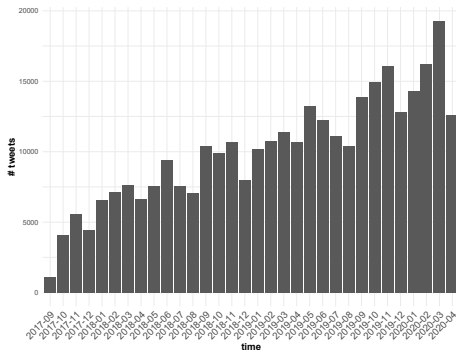
Biografie drucken

- Socioeconomic data and 2017 German federal election results: from www.bundeswahlleiter.de.

Data

Data Collection (II)

- Tweets (and further Twitter features): 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:



- In the following: grouping each MP's tweets on a monthly basis

Data

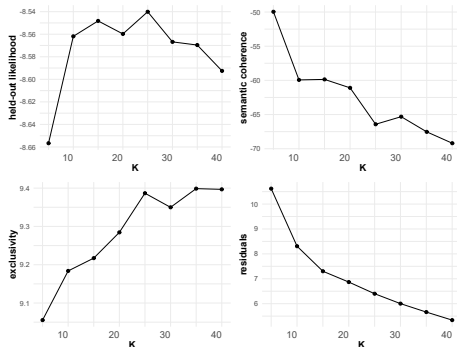
Data Preprocessing

- Preprocessing: in R **R**, using the *quanteda* package **quanteda**
- Transcription of German umlauts (ä/Ä, ö/Ö, ü/Ü) and ligature (ß)
- Removal of hyphens: relevant for compound words (e.g., *Corona-Krise* vs *Coronakrise*)
- Transformation of text data into document-feature matrix (DFM); conversion to lowercase; removal of stopword, units (*kg*, *uhr*), interjections (*aaahhh*, *ufff*), etc.
- Word stemming, i.e., cutting off word endings (e.g., *politisch* → *polit*) **lucas2015computer**

Model Selection and Global Characteristics

Model Selection

- Model evaluation metrics for hyperparameter K (number of topics):



- Best trade-off: $K = 15$

Model Selection and Global Characteristics

Labeling (I)

- First step: inspection of top words

Topic 1 Top Words:

Highest Prob: buerg, link, merkel, frau, sich

FREX: altpartei, islam, linksextremist, asylbewerb, linksextrem

Lift: eitan, 22jaehrig, abdelsamad, abgehalftert, afdforder

Score: altpartei, linksextremist, frauenkongress, islamist, boehring

Topic 3 Top Words:

Highest Prob: brauch, wichtig, leid, dank, klar

FREX: emissionshandel, soli, marktwirtschaft, feedback, co2steu

Lift: aequivalenz, altersvorsorgeprodukt, bildungsqualitaet, co2limit, co2meng

Score: emissionshandel, co2limit, basisrent, euert, technologieoff

Topic 4 Top Words:

Highest Prob: sozial, miet, kind, arbeit, brauch

FREX: mindestlohn, miet, wohnungsbau, mieterinn, loehn

Lift: auseinanderfaellt, baugipfel, bestandsmiet, billigflieg, binnennachfrag

Score: miet, mieterinn, mietendeckel, grundsicher, bezahlbar

Topic 6 Top Words:

Highest Prob: gruen, klimaschutz, brauch, klar, euro

FREX: fossil, erneuerbar, kohleausstieg, verkehrsminist, verkehrsw

Lift: abgasbetrug, abgebagert, abschaltleinricht, abschaltet, ammoniak

Score: erneuerbar, fossil, zdebel, verkehrsminist, klimaschutz

Model Selection and Global Characteristics

Labeling (II)

- word cloud
- Twitter excerpt!!

Model Selection and Global Characteristics

Labeling (III)

- Final labels
- XXX

Model Selection and Global Characteristics

Global Topic Proportions

- graph: weighted and unweighted props
- XXX

Model Selection and Global Characteristics

Global Topic Correlations

- topic 3 and 6 overlap
- topic correlation graph

Bibliography