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

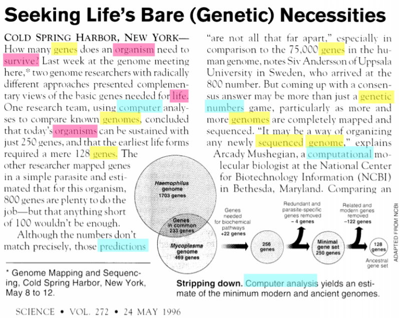
Introduction

- In this project: application of the **roberts2016model** to a self-created dataset
- Key contributions of this project:
  - Construction of dataset containing Twitter posts by members of the German Bundestag and a variety of metadata
  - Application of the *Structural Topic Model* (STM) to German MPs' Twitter communication
  - Development of new tools for estimation of relationship between topic proportions and metadata
  - Development and application of STM-specific train-test split to enable causal inference

Topic Modeling: Motivation and Theory

Motivation

- Motivating example: excerpt from a scientific article **blei2012presentation**



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Question at hand: how to assign colored words to topics?

Topic Modeling: Motivation and Theory

Notation and Terminology (I)

- *Words* w: instances of a vocabulary of V unique *terms*.
- *Documents* $d \in \{1, \ldots, D\}$: 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* $k \in \{1, \ldots, K\}$: latent thematic clusters within a text corpus; (implicit) representation of a corpus
- *Topic-word distributions* $\boldsymbol{\beta}$: probability distributions over words; $\boldsymbol{\beta}_k$ denoting the word distribution corresponding to the k-th topic

Topic Modeling: Motivation and Theory

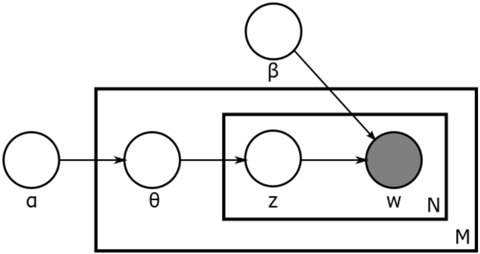
Notation and Terminology (II)

- *Topic assignments* $\boldsymbol{z}_{d,n}$: assignment of $w_{d,n}$ to a specific topic $k \in \{1, \ldots, K\}$; $\boldsymbol{\beta}_{d,n}$ representing the (assigned) word distribution for $w_{d,n}$
- *Topic proportions* $\boldsymbol{\theta}_d$: proportions of document d's words assigned to each of the topics; $\sum_{k=1}^{K} \theta_{d,k} = 1$, for all $d \in \{1, \ldots, D\}$
- *Bag-of-word* 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)

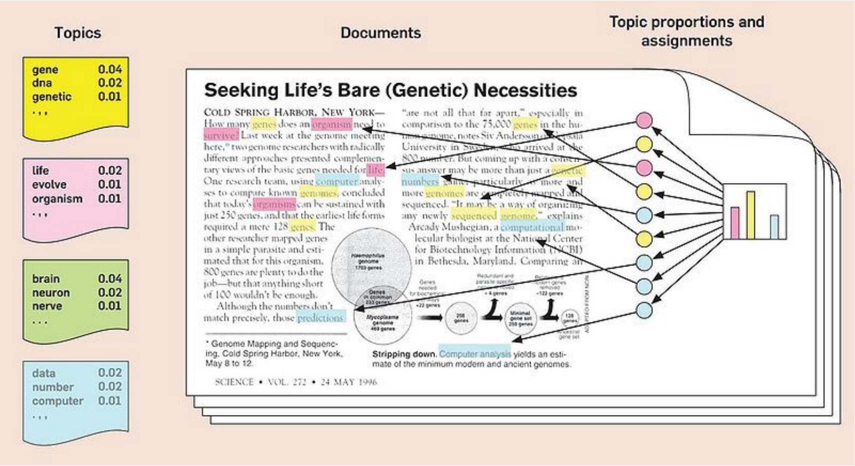
- First topic model with entirely probabilistic generating process: LDA **blei2003latent**
- Generative process for each document $d \in \{1, \ldots, D\}$:
  1) Draw topic proportions $\boldsymbol{\theta}_d \sim \mathrm{Dir}_K(\boldsymbol{\alpha})$.
  2) For each word $n \in \{1, \ldots, N_d\}$:
     a) Draw a topic assignment $\boldsymbol{z_{d,n}} \sim \mathrm{Multinomial}_K(\boldsymbol{\theta}_d)$.
     b) Draw a word $w_{d,n} \sim \mathrm{Multinomial}_V(\boldsymbol{\beta}_{d,n})$.
- 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**



Topic Modeling: Motivation and Theory

Inference and Parameter Estimation

- (Hierarchical) Bayesian model $\rightarrow$ exact inference impossible due to marginal distributions in the denominator of posterior distribution p
- Variational inference: positing a simple distribution family q
- Mean-field variational inference: positing full factorizability of approximating posterior q
- Then: minimizing Kullback-Leibler divergence between q and p

Data

Data Collection (I)

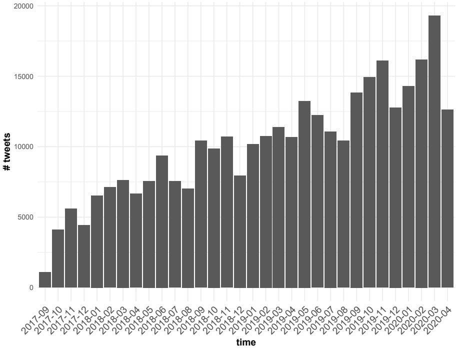
- MP-level data: from `www.bundestag.de/abgeordnete` using Python's *BeautifulSoup* and a *selenium web driver* **van1995python richardson2007beautiful**

- Twitter profiles: from official party homepages
- 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**roesslein2020tweepy**
- 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

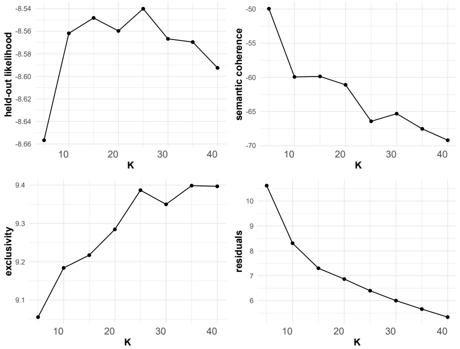
Data Preprocessing

- Preprocessing: in R **R**, using the *quanteda* package **quanteda**
- Transcription of German umlauts (e.g. ä $\rightarrow$ a) and ligature (ß$\rightarrow$ 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* $\rightarrow$ *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)

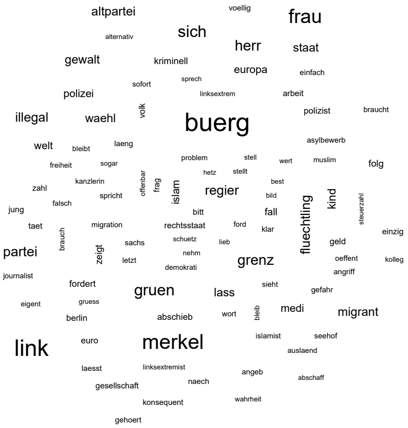
- Three-step procedure for labeling
- First step: top words for different weighting methodologies

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, euet, 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, abgebaggert, abschalteinricht, abschaltet, ammoniak
Score: erneuerbar, fossil, zdebel, verkehrsminist, klimaschutz

Model Selection and Global Characteristics

Labeling (II)

- Word cloud of **Highest Prob** top words (for topic 1):



- Word size corresponding to word frequency in topic 1

Model Selection and Global Characteristics

Labeling (III)

- Second step: look at documents (i.e., original tweets) with highest proportion of topic 1

Martin Hess
@Martin_Hess_AfD

Ehem. Verfassungsrichter bestätigt AfD-Forderung: Zurückweisung illegaler Migranten dringend geboten. Gegenwärtige Politik widerspricht dem Verstand und auch der Verfassung. Wir müssen zurück zu Recht & Ordnung, wie die #AfD seit fast 3 Jahren fordert!

Hans-Jürgen Papier hält Zurückweisung von Migranten an deutscher Grenze für ...
Im Asylstreit meldet sich nun Ex-Verfassungsrichter Papier zu Wort. Die Zurückweisung von Migranten an den Grenzen sei zwingend nötig, schreibt er in...
welt.de

9:47 AM · Jun 30, 2018 · Twitter for iPhone

Model Selection and Global Characteristics

Labeling (IV)

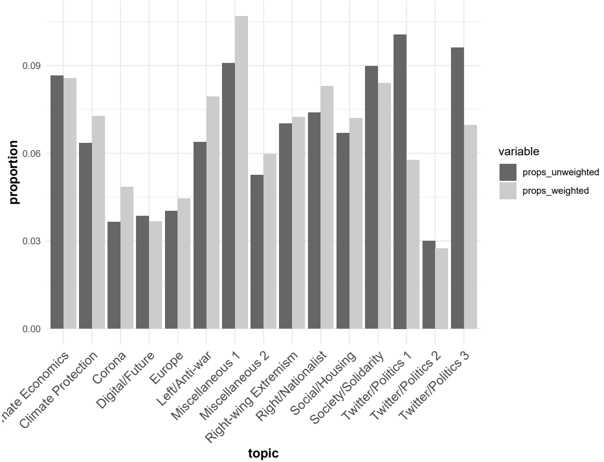
- Third step: assigning labels

Topic 1	Right/Nationalist
Topic 2	Miscellaneous 1
Topic 3	Climate Economics
Topic 4	Social/Housing
Topic 5	Digital/Future
Topic 6	Climate Protection
Topic 7	Europe
Topic 8	Corona
Topic 9	Left/Anti-war
Topic 10	Twitter/Politics 1
Topic 11	Twitter/Politics 2
Topic 12	Miscellaneous 2
Topic 13	Twitter/Politics 3
Topic 14	Right-wing Extremism
Topic 15	Society/Solidarity

Model Selection and Global Characteristics

Global Topic Proportions

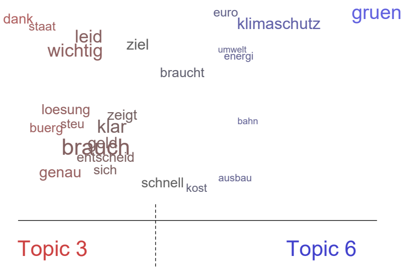
- Illustration of **global** topic proportions:

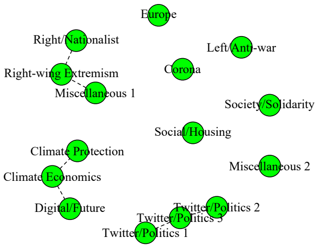


Model Selection and Global Characteristics

Global Topic Correlations

- Vocabulary overlap (left) and topic correlations (right):





Bibliography