

Session 4: **Topic Modeling**



„Magie“ verstehen: Klassische Schritte

1. Preprocessing

2. Analyse

3. Test auf
Qualitätskriterien



Session 1
(Einführung &
Preprocessing)



Session 2-4
(Co-Occurrence,
Diktionäre,
Topic Modeling)



Session 5
(Qualitätskriterien)



Agenda

1. Einführung
2. Model-Einstellungen
3. Analyse
4. Interpretation
5. Outro

1. Einführung

Alles, was heute gesagt wird – nur besser:



COMMUNICATION METHODS AND MEASURES

Communication Methods and Measures

Routledge
Taylor & Francis Group

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: <http://www.tandfonline.com/loi/hcms20>

Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology

Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam

To cite this article: Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam (2018) Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, *Communication Methods and Measures*, 12:2-3, 93-116, DOI: [10.1080/19312458.2018.1430754](https://doi.org/10.1080/19312458.2018.1430754)

To link to this article: <https://doi.org/10.1080/19312458.2018.1430754>

Maier et al., [2018](#)

COMMUNICATION METHODS AND MEASURES
2023, VOL. 17, NO. 2, 111–130
<https://doi.org/10.1080/19312458.2023.2167965>



Check for updates

What We Can Do and Cannot Do with Topic Modeling: A Systematic Review

Yingying Chen ^a, Zhao Peng ^b, Sei-Hill Kim^c, and Chang Won Choi ^d

^aSchool of Journalism and Communication, Renmin University of China, Beijing, China; ^bSchool of Communication, Emerson College, Massachusetts, USA; ^cSchool of Journalism and Mass Communications, University of South Carolina, South Carolina, USA; ^dSchool of Journalism and New Media, University of Mississippi, Mississippi, USA

Chen et al., [2023](#)



Sagt euch Clusteranalyse noch etwas?

- Wir wollen „übergreifende Muster“/Typen/Cluster finden
- Wir wollen unsere Daten explorativ bzw. induktiv erforschen, d.h. ohne vorab zu wissen, welche Kategorien wir identifizieren wollen
- Topic Modeling funktioniert relativ ähnlich...

Topic Modeling: Definition

„computational content-analysis technique [...] used to investigate the “hidden” thematic structure of [...] texts” (Maier et al., [2018](#), S. 93)

- Methode: Unüberwachtes maschinelles Lernen
- Vorgehen: wir identifizieren vorab **unbekannte, latente Themen** auf Basis häufig **gemeinsamer vorkommender, manifester Features** (s. **Sitzung 2: Co-Occurrence!**)
- Entsprechend gut kombinierbar mit z. B. qualitativen Methoden

Ein Beispiel aus der New York Times

What Happens When a Defendant Gets Covid-19 During Trial?

Fred Daibes, a real estate developer charged with Senator Robert Menendez, began feeling sick during the fifth week of the corruption trial, delaying it for at least a few days.

Was ist das
Thema dieses
Artikels?



Quelle: NYT, 14.06.2024

Ein Beispiel aus der New York Times

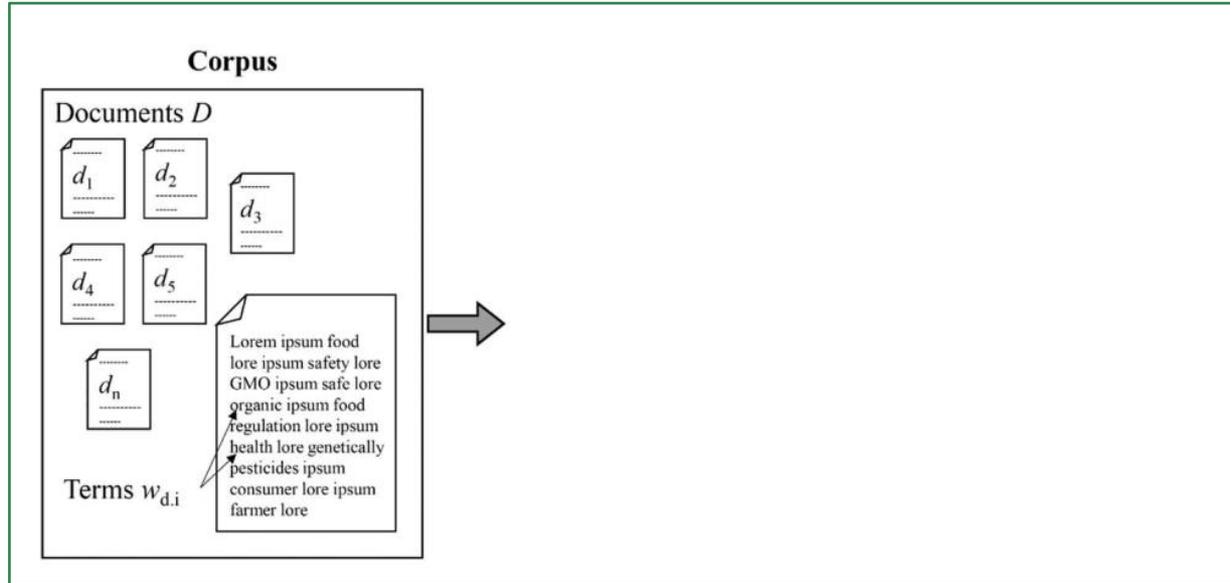
What Happens When a Defendant Gets Covid-19 During Trial?

Fred Daibes, a real estate developer charged with Senator Robert Menendez, began feeling sick during the fifth week of the corruption trial, delaying it for at least a few days.

Gerichtsverfahren?
Gesundheit?
Kriminalität?

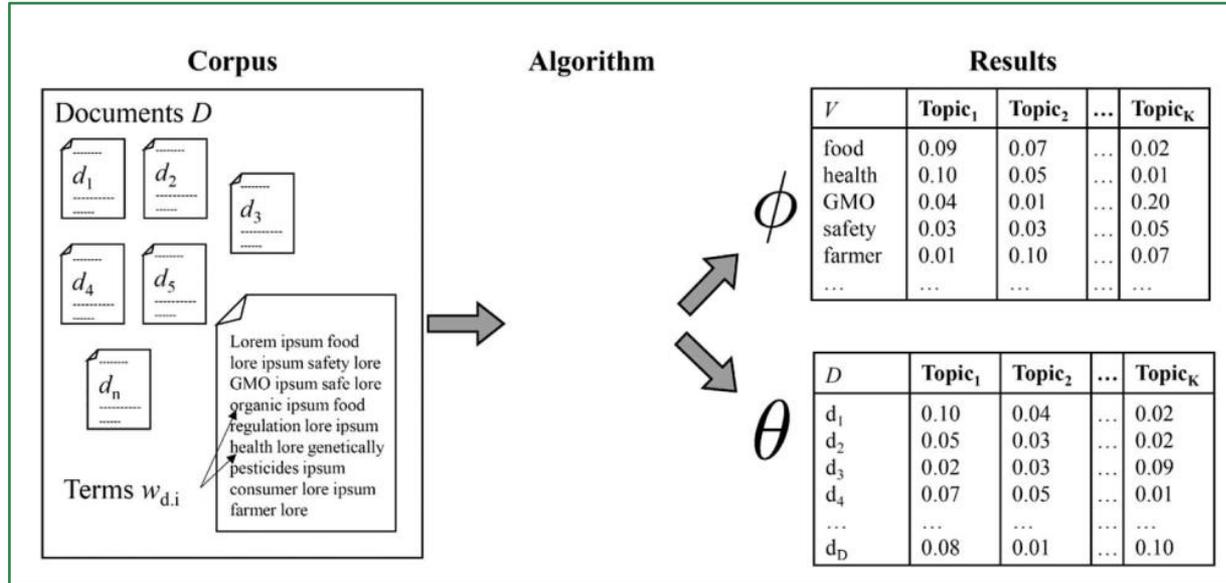
Quelle: NYT, 14.06.2024

Topic Modeling: Definition



(Maier et al., [2018](#), S. 94)

Topic Modeling: Definition



(Maier et al., [2018](#), S. 94)

Zentrale Matrizen

Word-topic oder Phi-Matrix:

- Bedingte Wahrscheinlichkeit, mit der Features in Themen prävalent sind
- Wortlisten, die Themen beschreiben ("**Top Features**")

V	Topic ₁	Topic ₂	...	Topic _K
food	0.09	0.07	...	0.02
health	0.10	0.05	...	0.01
GMO	0.04	0.01	...	0.20
safety	0.03	0.03	...	0.05
farmer	0.01	0.10	...	0.07
...

(Maier et al., [2018](#), S. 94)

Zentrale Matrizen

Document-topic oder Theta-Matrix:

- Bedingte Wahrscheinlichkeit, mit der Themen in Dokumenten prävalent sind
- Dokumentlisten, die Themen beschreiben ("**Top Documents**")

D	Topic ₁	Topic ₂	...	Topic _K
d ₁	0.10	0.04	...	0.02
d ₂	0.05	0.03	...	0.02
d ₃	0.02	0.03	...	0.09
d ₄	0.07	0.05	...	0.01
...
d _D	0.08	0.01	...	0.10

(Maier et al., [2018](#), S. 94)

Topic Modeling: Definition

- **Probabilistisches Modell:** Zuordnung von Wahrscheinlichkeiten, nicht eindeutigen Klassen
 - Features haben eine Wahrscheinlichkeit von kleiner als 1 je Thema (ϕ -matrix)
 - Themen haben eine Wahrscheinlichkeit von kleiner als 1 je Dokument (θ -matrix)
- **Das heißt praktisch:**
 - Das Modell sagt euch nicht eindeutig, welches das „eine“ Thema je Dokument ist oder wie ein Thema zu interpretieren ist – es gibt euch nur (probabilistische) Hinweise.

Topic Modeling: Definition

- **Generatives Modell:** Wir finden das statistische „passendste“ Modell, um unseren Korpus zu „generieren“
 - Gemeinsame Modellierung der beobachteten Variablen (Features i in den Dokumenten d) & der latenten Variablen (ϕ, θ)
- **Das heißt praktisch:**
 - Das Modell läuft in iterativen Schleifen immer und immer wieder (oft lange) durch, bis eine gute Lösung gefunden wurde.
 - Aber: Es gibt **z.T. nicht-deterministische** (d.h. je nach Einstellungen unterschiedliche) Lösungen.

Beispiel-Studie: News über Klimawandel

Global Environmental Change 70 (2021) 102353

Contents lists available at [ScienceDirect](#)

 **Global Environmental Change**

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



Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018)

Valerie Hase^{a,*}, Daniela Mahl^a, Mike S. Schäfer^a, Tobias R. Keller^b

^a Department of Communication and Media Research, University of Zurich, Switzerland
^b gfs.bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Climate Change
Media Coverage
Longitudinal Analysis
Computational Social Science
Automated Content Analysis
Comparative Research

ABSTRACT

Climate change poses a challenge to countries across the world, with news media being an important source of information on the issue. To understand how and how much news media cover climate change, this study compares coverage in ten countries from the Global North and the Global South between 2006 and 2018 ($N = 71,674$). Based on a panel analysis, we illustrate that news media attention varies across countries and is often associated with political, scientific, and (partly) societal focusing events. Based on an automated content analysis, we also find that news media do not only cover ecological changes or climate science, but that they focus predominantly on the societal dimension of climate change: They emphasize how humans are aware of, affected by, battle, or cause climate change. Overall, the study illustrates important differences between the Global North and the Global South. While countries from the Global North cover climate change more frequently, countries from the Global South focus more on its challenges and implications for society at large, i.e., the societal dimension of climate change.

(Hase et al., [2021](#))

Beispiel-Studie: News über Klimawandel



Korpus

10 Länder*, 2006-2018 ($N = 71,674$ Artikel)

*Australia, Canada, Germany, India, Namibia, New Zealand, South Africa, Thailand, UK, USA

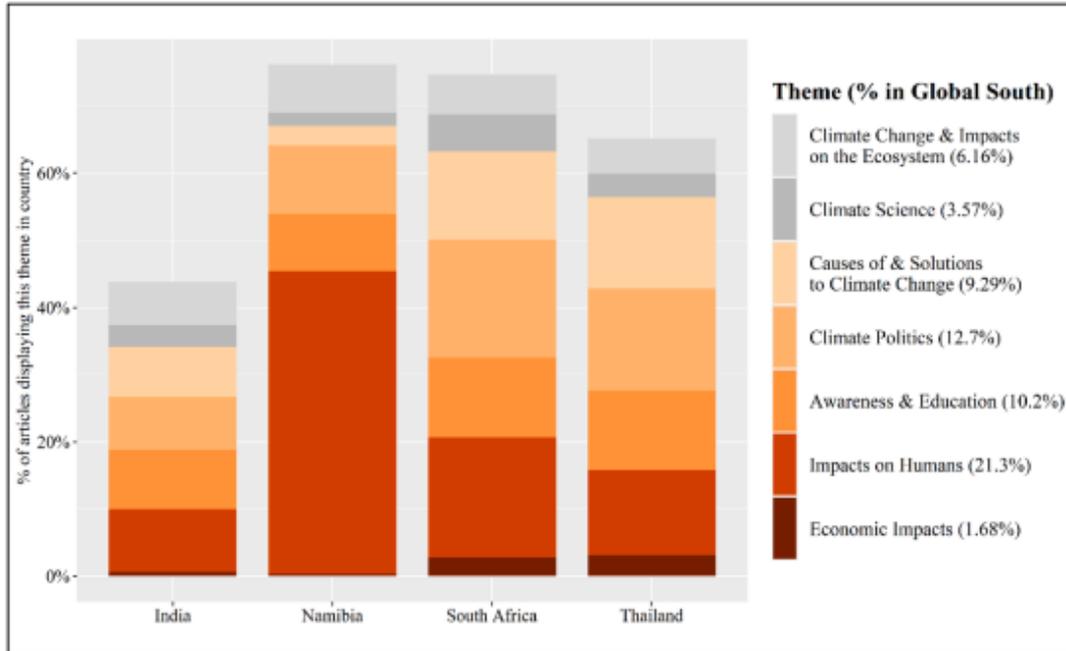


Method

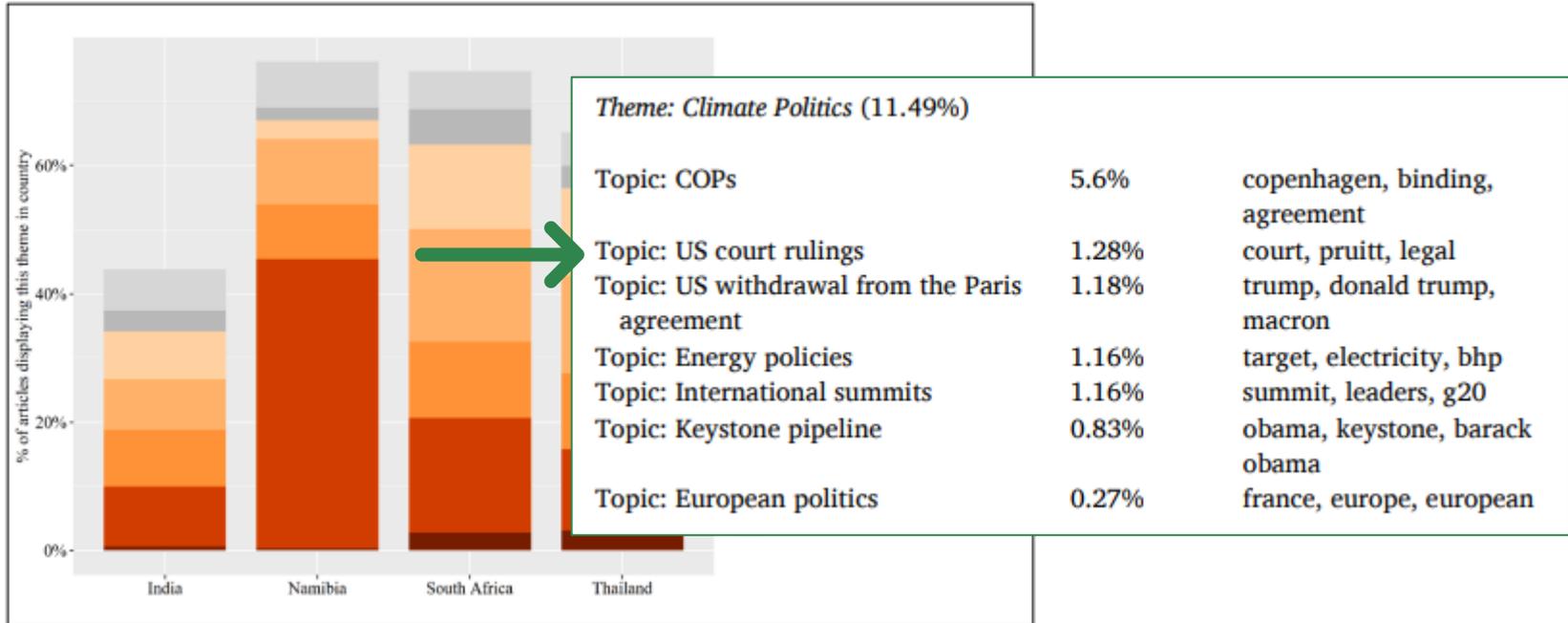
Structural topic modeling mit $K = 46$ relevanten Themen*,
die in 7 thematische Kontexte sortiert wurden

*(dazu später mehr)

Beispiel-Studie: News über Klimawandel



Beispiel-Studie: News über Klimawandel



Beispiel-Studie: *Word-Topic-Matrix*

feature	Topic1	Topic2	Topic3	Topic4	Topic5
is	0.046554314	1.285196e-02	1.234639e-02	2.357912e-07	2.578756e-02
are	0.023774019	4.356566e-03	3.769127e-04	2.868092e-04	8.405222e-03
be	0.019577716	7.653390e-04	4.580330e-03	2.589806e-05	4.367386e-03
percent	0.016751425	2.505350e-26	3.738808e-58	1.368502e-30	6.345623e-13
can	0.015502349	3.181028e-03	4.834248e-05	1.635215e-05	3.727694e-04
climate_change	0.014968295	1.699731e-02	1.064793e-02	1.667098e-03	1.909034e-03
german	0.014797019	4.410063e-27	3.428276e-38	6.761932e-04	5.556742e-19



Wahrscheinlichkeit, mit der das Feature „German“ in Topic 1 vorkommt: 1.5%

Beispiel-Studie: *Document-Topic-Matrix*

docnum	Topic1	Topic2	Topic3	Topic4	Topic5
110	0.8815214	0.0003408218	2.501259e-06	2.969106e-04	0.0015961312
26190	0.8331462	0.0014109805	8.691409e-06	1.469968e-02	0.0056127009
26464	0.7489909	0.0003700080	1.326193e-05	2.335178e-04	0.0026629212
26038	0.7379161	0.0026914179	1.582467e-05	1.180497e-01	0.0018521122
26342	0.7253922	0.0041106925	2.544697e-05	5.572124e-03	0.0015391569



Wahrscheinlichkeit, mit der Topic 1 in Artikel 26.342 vorkommt: 72.5%

Ein zweites Beispiel

What Communication Scholars Write About: An Analysis of 80 Years of Research in High-Impact Journals

ELISABETH GÜNTHER

University of Münster, Germany

EMESE DOMAHIDI

Leibniz-Institut für Wissensmedien, Germany

Research topics, as indicators of the profession's development, are central to the evaluation of academic practices in communication research. To investigate the main topics in our field, we trace the development of research topics since the 1930s by evaluating more than 15,000 articles from 19 academic journals based on an automated content analysis. Topic modeling reveals a high diversity from the early years on. Only a few journals show the tendency to focus on one topic only, whereas most outlets cover a broad variety and thus represent the field as a whole. Although our discipline is strongly interconnected with the changing media landscape, results show that communication research is characterized by high consistency. Although they have not provoked a revolutionary change, Internet and social media have become the most monitored media, parallel to—not displacing—classic media such as newspapers and TV.

Günther & Domahidi, [2017](#)

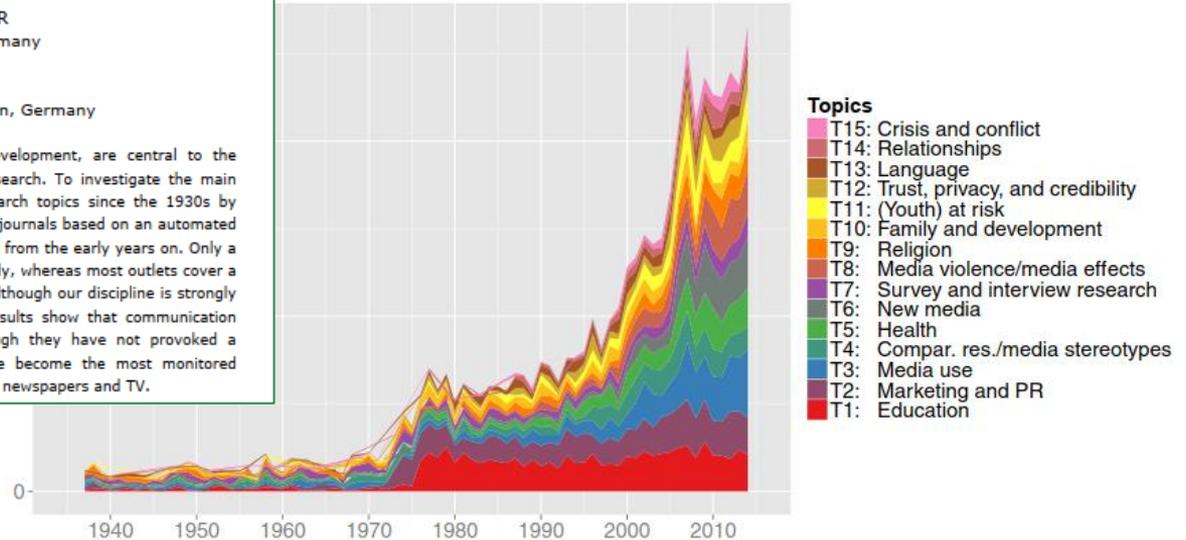


Figure 3. Topic attention over time (core topics, $n = 10,017$, maximum of two topics per abstract).

Topic Modeling

1. **Preprocessing**
2. Modell-Einstellungen
3. Analyse & Interpretation
4. Validierung



Wichtige Fragen u. a.:

1. Welche Preprocessing-Schritte wähle ich? (Denny & Spirling, [2018](#); Maier et al., [2020](#))

Topic Modeling

1. Preprocessing
2. **Modell-Einstellungen**
3. Analyse & Interpretation
4. Validierung



Wichtige Fragen u. a.:

1. Welches Verfahren bzw. welchen Algorithmus wähle ich?
(z. B. Churchill et al., [2020](#); Eshima et al., [2023](#); Roberts et al., [2014](#))
2. Welche Anzahl Topics K wähle ich?
3. Wie setze ich den Hyperparameter α als „prior“ für die θ -Matrix?
4. Wie setze ich den Hyperparameter β als „prior“ für die φ -Matrix?

Topic Modeling

1. Preprocessing
2. Modell-Einstellungen
3. **Analyse & Interpretation**
4. Validierung



Wichtige Fragen u. a.:

1. Welche Themen „behalte“ ich – und welche ignoriere ich als sog. „Background“-Themen?
2. Wie labelle & interpretiere ich Themen?
3. Wie ordne ich Dokumente Themen zu? Mache ich das?

Topic Modeling

1. Preprocessing
2. Modell-Einstellungen
3. Analyse & Interpretation
4. **Validierung**



Wichtige Fragen u. a.:

1. Wie evaluiere ich, ob die identifizierten „Themen“ mein latentes theoretisches Konstrukt valide abbilden?

(Bernhard et al., [2023](#); Quinn et al., [2010](#))

Topic Modeling

In welchem Zusammenhang bzw. für welche theoretischen Fragen seid ihr bisher mit Topic Modeling in Berührung gekommen?





Was sind „Themen“?

- Events?
- Themen wie Sport, Politik?
- Frames?

(Eisele et al., [2023](#); Günther et al., [2022](#); Nicholls et al., [2021](#))

Topic Modeling für Frame-Analysen?

- **Befürworter:innen** (u.a. Walter & Ophir, [2019](#)):
 - Topic Modeling zur Identifikation von Themen
 - Netzwerkanalyse & Community Detection zur Frame-Identifikation

COMMUNICATION METHODS AND MEASURES
2019, VOL. 13, NO. 4, 248–266
<https://doi.org/10.1080/19312458.2019.1639145>

 **Routledge**
Taylor & Francis Group

 Check for updates

News Frame Analysis: An Inductive Mixed-method Computational Approach

Dror Walter  and Yotam Ophir 

^aCommunication, Georgia State University, Atlanta, USA; ^bCommunication, University at Buffalo, State University of New York, Buffalo, USA

ABSTRACT

Framing is one of the most central, applicable, and contested theories in communication research. At the heart of the debate on framing is the question of operationalizing and measuring emphasis frames. We harness novel computational tools to propose a new method for inductive identification of frames. We argue and demonstrate that frame elements could be identified using topic modeling, and that frame elements can then be automatically grouped into frame “packages” using community detection techniques applied to the topic network. Building upon recent conceptual and methodological developments in framing research, we introduce a new approach, the Analysis of Topic Model Networks (ANTMN). We demonstrate the applicability of our method in case studies where framing theory is developed and fairly consistent, and in exploratory ones where it is not, using three diverse U.S. news corpora: the coverage of political candidates in Senate races ($n = 8,337$ articles), foreign nations ($n = 18,216$), and infectious diseases and epidemics ($n = 5,005$). We conclude by discussing the theoretical, methodological, and practical implications of ANTMN.

Topic Modeling für Frame-Analysen?

- Kritischere Perspektiven

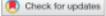
(u.a. Eisele et al., [2023](#); Nicholls & Culpepper, [2021](#)):

- Funktioniert wenn überhaupt (!) nur bei thematisch engen Korpora
- Andere Verfahren, etwa überwacht maschinelles Lernen, besser

Insgesamt: Don't do it!

POLITICAL COMMUNICATION
2021, VOL. 38, NOS. 1-2, 159-181
<https://doi.org/10.1080/10584609.2020.1812777>

 Routledge
Taylor & Francis Group



Computational Identification of Media Frames: Strengths, Weaknesses, and Opportunities

Tom Nicholls  and Pepper D. Culpepper 

Blavatnik School of Government, University of Oxford, Oxford, UK

ABSTRACT
With the availability of large volumes of electronic communications data and the increasing sophistication of computational techniques, the development of automated approaches for different kinds of framing analysis is an important goal of researchers. There is as yet no standard method for the “unsupervised” inductive identification of frames based upon the content of articles. Three groups of core approaches underlie a wide range of work in this area, and we compare three techniques based on these approaches against each other and against manual human analysts. The three techniques are a k-means clustering algorithm together with a sophisticated natural language processing (NLP)-based feature selection process; evolutionary factor analysis (EFA), a factor analysis approach; and the structural topic model (STM). We use two datasets – one very focused and one extremely broad – as examples of the kind of frame analysis problems readers may wish to attempt. Even in a highly targeted dataset, we find some distance between the frames generated by computational analysis and those manually produced by our human analysts. The details of each method have a substantial impact on frame quality and interpretation. We find that the STM approach is the most effective for our narrow-scope dataset, but that it returns definite topics, and not frames, when working on our very broad dataset. We also show that we can get results emphasizing different parts of the framing problem by combining parts of different methods as a multi-stage process, rather than viewing available methods as simple plug-and-play models.

KEYWORDS
Framing; computational communication science; political communication; methods



Wie kann ich diese Analysen in R anwenden?

- Kleine Auswahl möglicher R-Pakete
 - „*stm*“ (für Structural Topic Modeling)
 - „*keyATM*“ (für Keyword Assisted Topic Modeling)
 - „*topicmodels*“ (für LDA basiertes Verfahren)
 - „*tidytext*“ (für Extraktion z.B. der Theta- oder Phi-Matrix)
 - „*LDAvis*“ (zur Visualisierung)
 - „*stm insights*“ (zur Visualisierung)
 - „*oolong*“ (für Validierungen)

Take-Aways



- **Topic Modeling:** Explorative Identifikation unbekannter, latenter Themen auf Basis häufig gemeinsamer vorkommender, manifester Features mittels unüberwachten maschinellen Lernens
- **Wichtigste Schritte:**
 - Preprocessing
 - Modell-Einstellungen
 - Analyse & Interpretation
 - Validierung

2. Modell-Einstellungen



Modell-Einstellungen

1. Welches Verfahren bzw. welchen Algorithmus wähle ich?
(z. B. Churchill et al., [2020](#); Eshima et al., [2023](#); Roberts et al., [2014](#))
2. Welche Anzahl Topics K wähle ich?
3. Wie setze ich den Hyperparameter α als „prior“ für die θ -Matrix?
4. Wie setze ich den Hyperparameter β als „prior“ für die φ -Matrix



Modell-Einstellungen

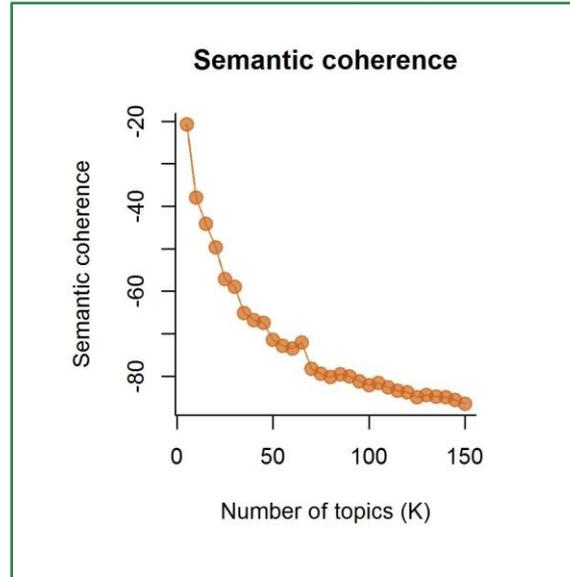
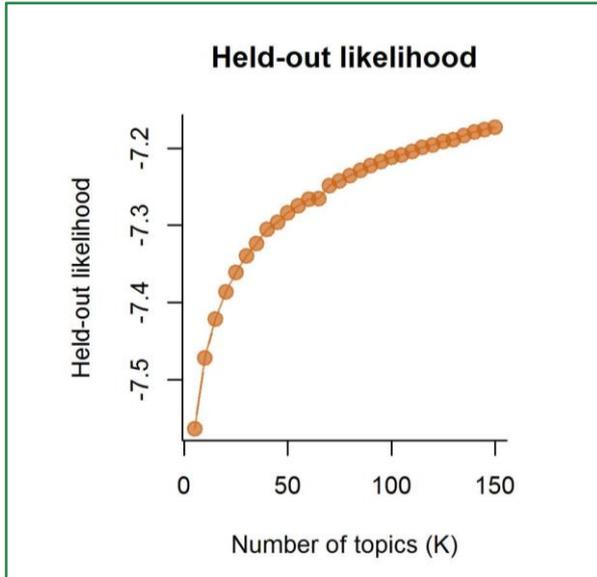
1. Welches Verfahren bzw. welchen Algorithmus wähle ich?
(z. B. Churchill et al., [2020](#); Eshima et al., [2023](#); Roberts et al., [2014](#))
2. **Welche Anzahl Topics K wähle ich?**
3. Wie setze ich den Hyperparameter α als „prior“ für die θ -Matrix?
4. Wie setze ich den Hyperparameter β als „prior“ für die φ -Matrix



Anzahl von Themen K

- Forschende müssen vorab bestimmen, welche Anzahl Themen sie identifizieren wollen: 5? 20? 200?
- Es gibt keine einzelne richtige Lösung. Diese kann davon abhängen...
 - was Themen für uns theoretisch bedeuten
 - wie sinnvoll welche Anzahl von Themen für den spezifischen Korpus ist
- Entscheidung basiert u.a. auf
 - Statistischem Fit (z.B. Coherence, Perplexity)
 - Interpretierbarkeit (z.B. Top Features, Top Documents)
 - Rank-1 Metrik (z.B. Häufigkeit bestimmter Themen)

Statistischer Fit



Achtung: Statistischer Fit korreliert kaum mit manuellen Einschätzungen! (Chang et al., [2009](#))

(Hase et al., [2021](#))



Interpretierbarkeit

Bei Lösungen mit unterschiedlichem K:

- **Top Features:** Ergeben Features, die ein Thema beschreiben, eine sinnvolle Interpretation des Themas?
- **Top Documents:** Passen Dokumente, die ein Thema beschreiben, zum Thema?



Rank-1 Metrik

- Zeigt an, wie häufig jedes Thema das **Hauptthema** (d.h. prävalenteste Thema) über alle Dokumente hinweg ist
- Kleine Themen = ggf. irrelevante Themen?
- Wichtig: Eindeutige Zuordnung widerspricht eigentlich dem probabilistischen Ansatz von Topic Modeling – können nicht auch mehrere Themen vorkommen?

```
function(scope, element, attr, ngSwitchController) {
  var watchExpr = attr.ngSwitch || attr.on,
      selectedTranscludes = [],
      selectedElements = [],
      previousElements = [],
      selectedScopes = [];

  scope.$watch(watchExpr, function ngSwitchWatchAction(value) {
    var i, ii;
    for (i = 0, ii = previousElements.length; i < ii; ++i) {
      previousElements[i].remove();
    }
    previousElements.length = 0;

    for (i = 0, ii = selectedScopes.length; i < ii; ++i) {
      var selected = selectedElements[i];
      selectedScopes[i].$destroy();
    }
  });
}
```

```
selectedElements.length = 0;
selectedScopes.length = 0;
```

```
if ((selectedTranscl...
```

Zeit für R!

Pakete installieren & aktivieren

```
#install.packages("tidyverse")
#install.packages("RCurl")
#install.packages("quanteda")
#install.packages("stm")
#install.packages("reshape2")
```

```
library("tidyverse")
library("RCurl")
library("quanteda")
library("stm")
library("reshape2")
```

```
install.packages("stminsights")
library("stminsights")
```

Preprocessing

```
# Daten laden
url <- getURL("https://raw.githubusercontent.com/valeriehase/textasdata-ms/main/data")
data <- read.csv2(text = url)

# Preprocessing
tokens <- tokens(data$Description,
                 what = "word", #Tokenisierung, hier zu Wörtern als Analyseeinheit
                 remove_punct = TRUE, #Entfernung von Satzzeichen
                 remove_numbers = TRUE) %>% #Entfernung von Zahlen

# Kleinschreibung
tokens_tolower() %>%

# Entfernung von Stopwörtern
tokens_remove(stopwords("english")) %>%

# Stemming
tokens_wordstem()
```

Preprocessing

```
# Text-as-Data Repräsentation als Document-Feature-Matrix
dfm <- tokens %>%
  dfm() %>%

# Relative pruning
dfm_trim( min_docfreq = 0.005,
          max_docfreq = 0.99,
          docfreq_type = "prop",
          verbose = TRUE)
```

```
# DFM in STM-Objekt umwandeln
dfm_stm <- convert(dfm, to = "stm")
```

Statistischer Fit

```
# dfm_stm$documents: Welche Dokumente nutzen wir?
# dfm_stm$vocab: Welche Features nutzen wir?
stat_fit <- searchK(dfm_stm$documents, dfm_stm$vocab, K = c(4,6), verbose = TRUE)

# Wir speichern die Ergebnisse im Objekt "Plot" ab
plot <- data.frame("K" = c(4, 6),

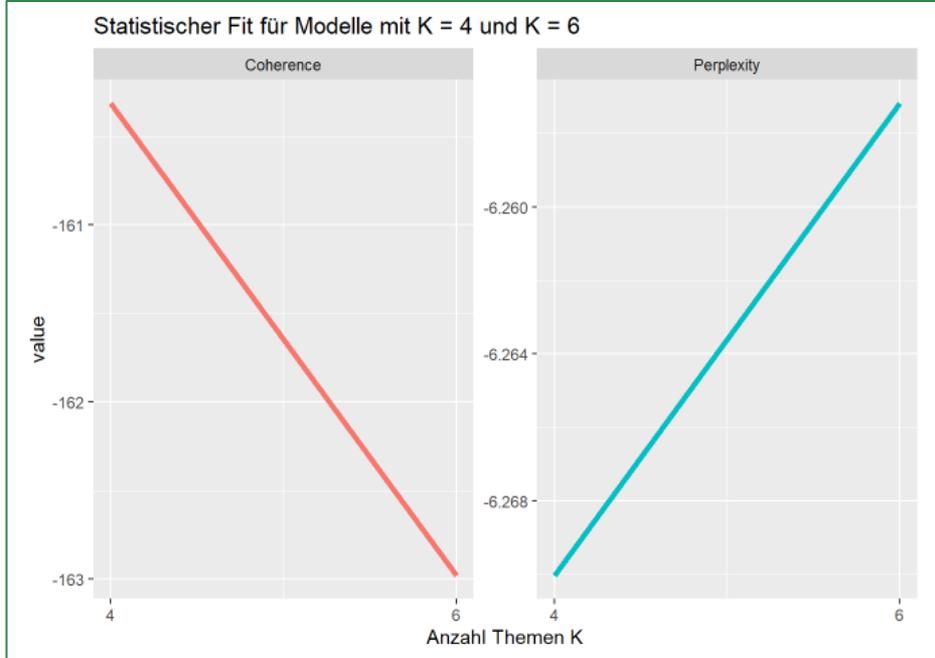
                  #Kohärenz: Je höher, desto besser
                  "Coherence" = unlist(stat_fit$results$semcoh),

                  #Perplexität: Je niedriger, desto besser
                  "Perplexity" = unlist(stat_fit$results$heldout))

# Wir wandeln das Format zu einem "long format" um
plot <- melt(plot, id = c("K"))

# Plot erstellen
ggplot(plot, aes(K, value, color = variable)) +
  geom_line(linewidth = 1.5, show.legend = FALSE) +
  scale_x_continuous(breaks = c(4, 6)) +
  facet_wrap(~ variable, scales = "free_y") +
  labs(x = "Anzahl Themen K",
       title = "Statistischer Fit für Modelle mit K = 4 und K = 6")
```

Statistischer Fit



Kohärenz:
sollte hoch
sein

Perplexity:
sollte niedrig
sein

Interpretierbarkeit: Top Features

```
#Model mit K = 4 berechnen
model_4K <- stm(documents = dfm_stm$documents,
                vocab = dfm_stm$vocab,
                K = 4)

#Model mit K = 6 berechnen
model_6K <- stm(documents = dfm_stm$documents,
                vocab = dfm_stm$vocab,
                K = 6)
```

Interpretierbarkeit: Top Features

```
# Top Features für K = 4
topics_4K <- labelTopics(model_4K, n = 10)

# Nur Top-10 Features nach Frex-Gewichtung, welche
# Gewichtet Features nach Kohärenz und Exklusivität
topics_4K <- data.frame("features" = t(topics_4K$fr

# Benennung & Ausgabe
colnames(topics_4K) <- paste("Topics", c(1:4))
topics_4K
```

	Topics 1	Topics 2	Topics 3	Topics 4
1	crime	find	life	mysteri
2	human	seri	live	discov
3	detect	struggl	new	boy
4	adventur	chang	friend	lead
5	crimin	name	citi	agent
6	case	everi	school	angel
7	solv	polit	york	dark
8	use	head	togeth	mother
9	futur	seem	high	surviv
10	alien	embark	person	children

```
# Top Features für K = 6
topics_6K <- labelTopics(model_6K, n = 10)

# Nur Top-10 Features nach Frex-Gewichtung, welche besser i
# Gewichtet Features nach Kohärenz und Exklusivität
topics_6 <- data.frame("features" = t(topics_6K$fr

#Benennung & Ausgabe
colnames(topics_6) <- paste("Topics", c(1:6))
topics_6
```

	Topics 1	Topics 2	Topics 3	Topics 4	Topics 5	Topics 6
1	polic	best	vampir	human	comedi	american
2	crime	school	famili	alien	show	war
3	detect	new	teen	earth	togeth	serial
4	profession	boy	brother	planet	various	dream
5	investig	hero	differ	futur	work	killer
6	stori	york	son	struggl	anim	plan
7	person	mother	epic	fight	name	cia
8	case	return	demon	galaxi	featur	turn
9	special	high	world	space	everyth	drug
10	polit	troubl	navig	last	lead	begin

Interpretierbarkeit: Top Documents

```
findThoughts(model_4K, data$Description, topics = 1, n = 3)
```

Topic 1:

```
Detective Jane Rizzoli and Chief Medical Examiner Dr. Maura Isles team up to solve  
An elite unit, led by an ex-homicide cop, which is linked to the Miami-Dade Police  
The further adventures in time and space of the alien adventurer known as the Doctc
```

Bitte lest euch mind. drei Top Documents je Thema durch –
wie würdet ihr diese Themen nun beschreiben?



Rank-1 Metrik

```
theta_4K <- make.dt(model_4K)
theta_6K <- make.dt(model_6K)

#Schauen wir uns kurz beispielhaft die Matrix an:
theta_4K %>%
  head()
```

	docnum	Topic1	Topic2	Topic3	Topic4
	<int>	<num>	<num>	<num>	<num>
1:	1	0.2663954	0.2620603	0.2840283	0.1875160
2:	2	0.3234499	0.1736692	0.2082024	0.2946784
3:	3	0.1401247	0.1727312	0.1448235	0.5423207
4:	4	0.1237850	0.1137826	0.6349453	0.1274870
5:	5	0.1694126	0.1476420	0.3017674	0.3811779
6:	6	0.2748651	0.1791878	0.2271607	0.3187864

Rank-1 Metrik

```
# Zuerst erstellen wir zwei leere Spalten in unserem Dataframe data
data <- data %>%
```

```
  # Leere Variable für Hauptthema, wird später "aufgefüllt"
  mutate(Rank1_K4 = NA,
         Rank1_K6 = NA)
```

```
# Berechnung von Rank-1 Metrik
for (i in 1:nrow(data)){ # Schleife: Für jede nachfolgende Zeile...
```

```
  # Bestimme Hauptthema für K = 4
```

```
  # Wähle alle Spalten der Document-Topic-Matrix aus (ohne die erste, die nur doc_id)
  column <- theta_4K[i,-1]
```

```
  # Bestimmung des Hauptthemas (Spalte mit dem höchsten Wert)
  maintopic <- colnames(column)[which(column == max(column))]
```

```
  # Zuweisung des Hauptthemas zur entsprechenden Zeile
  data$Rank1_K4[i] <- maintopic
  rm(column, maintopic)
```

```
  # Bestimme Hauptthema für K = 6
```

```
  # Wähle alle Spalten der Document-Topic-Matrix aus (ohne die erste, die nur doc_id)
  column <- theta_6K[i,-1]
```

```
  # Bestimmung des Hauptthemas (Spalte mit dem höchsten Wert)
  maintopic <- colnames(column)[which(column == max(column))]
```

```
  # Zuweisung des Hauptthemas zur entsprechenden Zeile
  data$Rank1_K6[i] <- maintopic
  rm(column, maintopic)
}
```

Rank-1 Metrik

Schauen wir uns an, wie häufig jedes Thema bei $K = 4$ vorkommt!

```
# Erzeugung einer Häufigkeitstabelle für Rank-1
data %>%

# absolute Anzahl jedes Themas
count(Rank1_K4) %>%

# Ausgabe in Prozent (perc)
mutate(perc = prop.table(n)*100,
       perc = round(perc, 2))
```

	Rank1_K4	n	perc
1	Topic1	190	21.11
2	Topic2	40	4.44
3	Topic3	311	34.56
4	Topic4	359	39.89

```
# Erzeugung einer Häufigkeitstabelle für Rank-1 Themen bei K = 6
data %>%

# absolute Anzahl jedes Themas
count(Rank1_K6) %>%

# Ausgabe in Prozent (perc)
mutate(perc = prop.table(n)*100,
       perc = round(perc, 2))
```

	Rank1_K6	n	perc
1	Topic1	147	16.33
2	Topic2	116	12.89
3	Topic3	189	21.00
4	Topic4	197	21.89
5	Topic5	139	15.44
6	Topic6	112	12.44

Take-Aways



- **Modell-Einstellungen:** Konfigurationen, die Forschende festlegen müssen, bevor sie ihr Topic Model berechnen können (z.B. Hyperparameter)
- **K:** Die Anzahl an Themen, die das Modell finden soll. Lässt sich entscheiden auf Basis von...
 - Statistischem Fit
 - Interpretierbarkeit
 - Rank-1 Metrik



Pause

3. Analyse & Interpretation

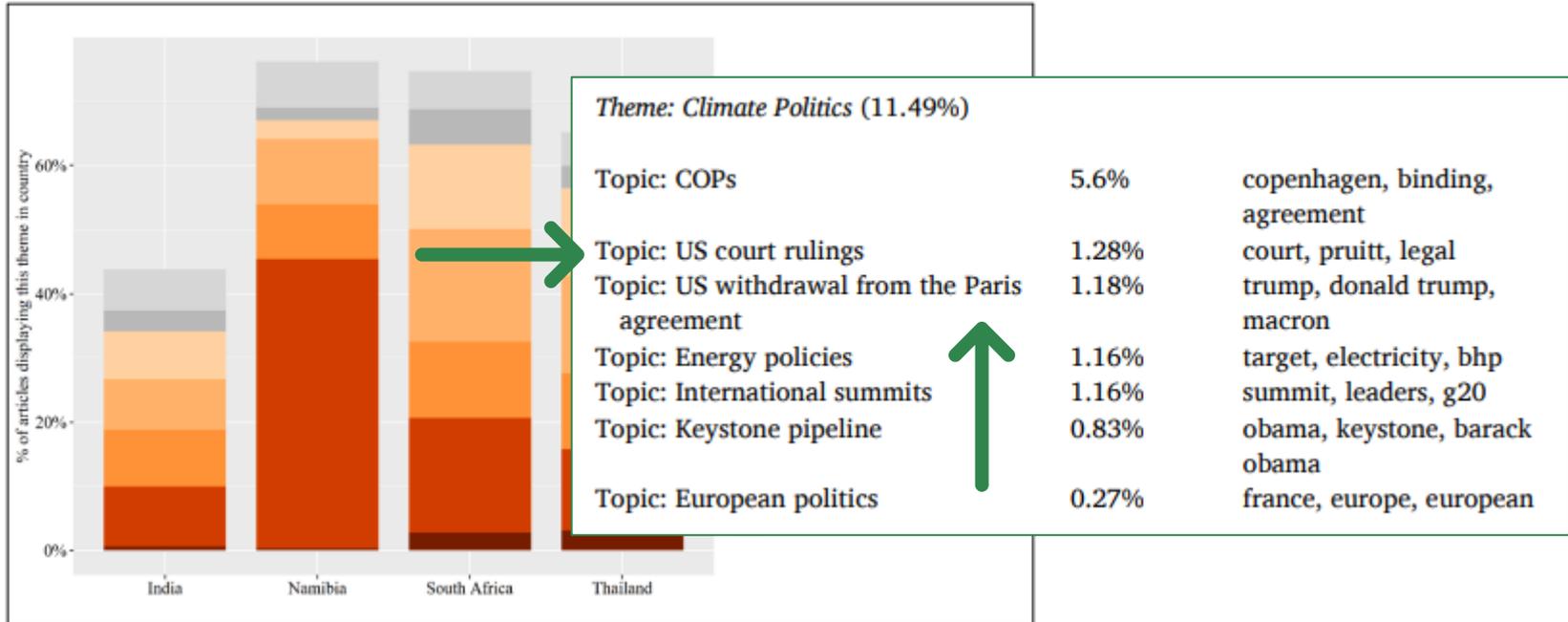


Analyse

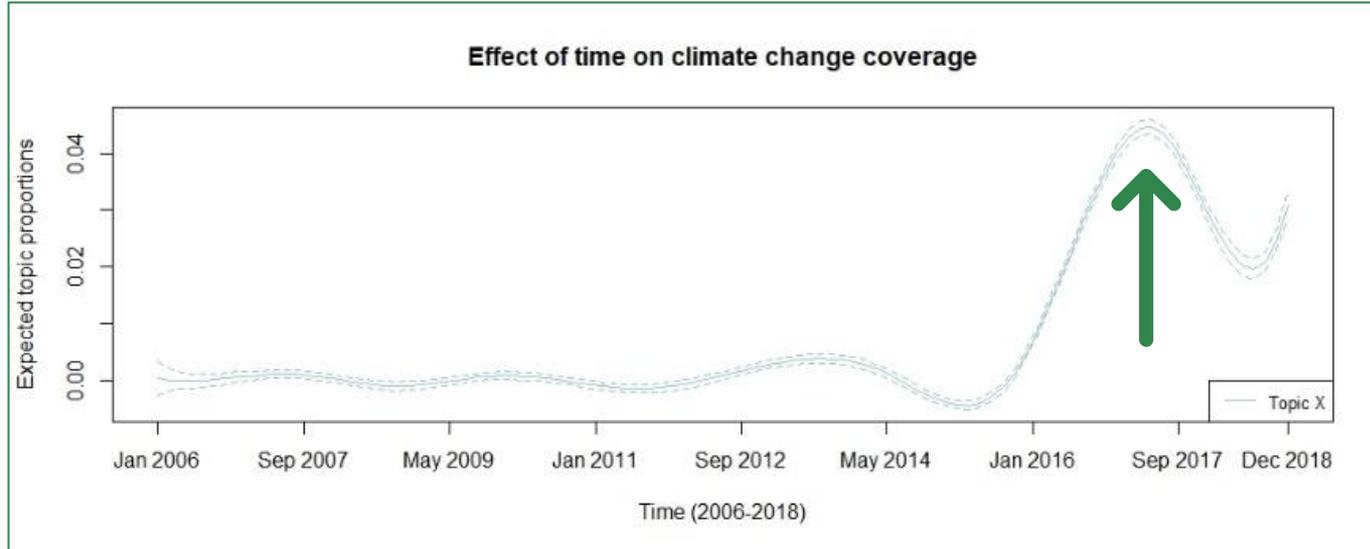
Structural Topic Modeling (eine Variante von Topic Modeling, beliebt in R!) ermöglicht es, den Einfluss unabhängiger Variablen zu modellieren, genauer auf:

- die Prävalenz von Themen (`prevalence`-Argument)
- den Inhalt von Themen (`content`-Argument)

Beispiel-Studie: News über Klimawandel



Beispiel-Studie: News über Klimawandel



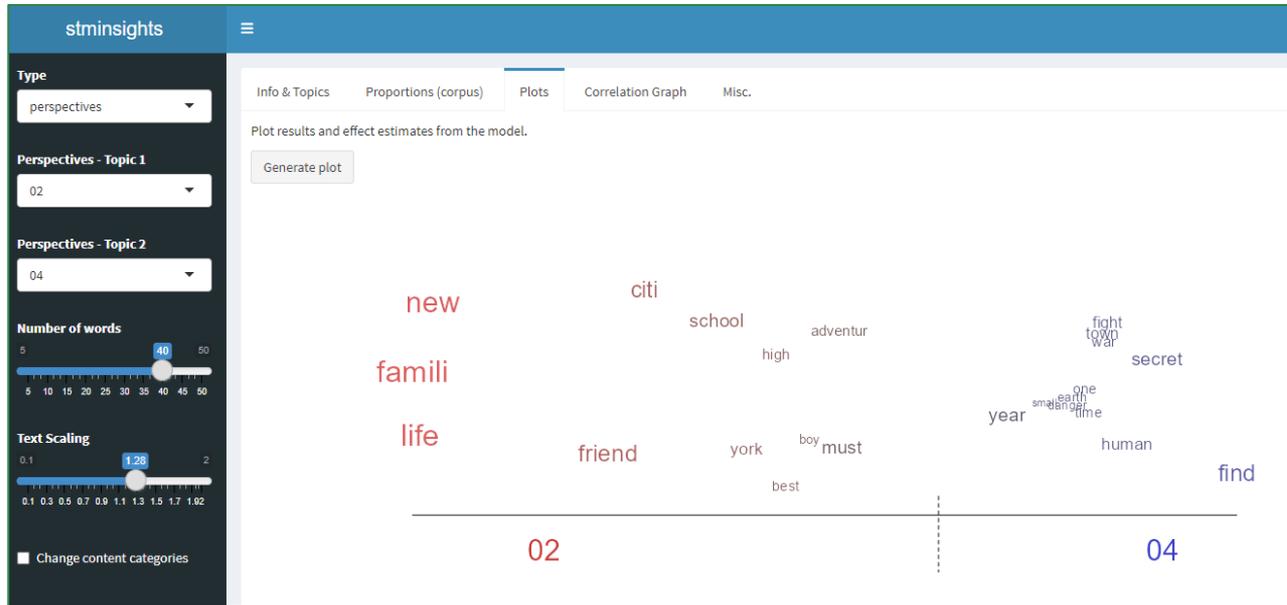


Interpretation

- Identifikation & Ausschluss von „Background“-Topics
- Identifikation & Labelling von relevanten Topics
- Ggf. Gruppierung in übergreifende Kontexte (z.B. „politische Themen“)
- Nutzung für deskriptive oder inferenzstatistische Verfahren

Visualisierung mit stminsights

- Paket von Carsten Schwemmer, das Visualisierung von Themenmodellen ermöglicht



```
function(scope, element, attr, ngSwitchController) {
  var watchExpr = attr.ngSwitch || attr.on,
      selectedTranscludes = [],
      selectedElements = [],
      previousElements = [],
      selectedScopes = [];

  scope.$watch(watchExpr, function ngSwitchWatchAction(value) {
    var i, ii;
    for (i = 0, ii = previousElements.length; i < ii; ++i) {
      previousElements[i].remove();
    }
    previousElements.length = 0;

    for (i = 0, ii = selectedScopes.length; i < ii; ++i) {
      var selected = selectedElements[i];
      selectedScopes[i].$destroy();
    }
  });
}
```

```
selectedElements.length = 0;
selectedScopes.length = 0;
```

```
if ((selectedTranscl...
```

Zeit für R!

Einfluss unabhängiger Variablen

Hypothese: Das Erscheinungsjahr hat einen Effekt auf das Thema einer Serie

```
data <- data %>%  
  
# Wir entfernen alle nicht-numerische Zeichen, um "-" zu entfernen  
mutate(Year_Start = gsub("[^0-9]", "", Year),  
  
# Wir beschränken uns nur auf die ersten 4 Jahre  
Year_Start = substr(Year_Start, 1, 4),  
  
# Wir verwandeln das ganze in eine numerische Variable  
Year_Start = as.numeric(Year_Start),  
  
#Wir ersetzen fehlende Werte mit dem Mittelwert (2010)  
Year_Start = replace(Year_Start,  
                      is.na(Year_Start),  
                      2010))
```

```
#Ausgabe der ersten Zeilen  
data %>%  
  
#Reduktion auf weniger Variablen  
select(Title, Year, Year_Start) %>%  
  
#Ausgabe der ersten Zeilen  
head()
```

	Title	Year	Year_Start
1	1. Game of Thrones	2011-2019	2011
2	2. Breaking Bad	2008-2013	2008
3	3. Stranger Things	2016-2025	2016
4	4. Friends	1994-2004	1994
5	5. The Walking Dead	2010-2022	2010
6	6. Sherlock	2010-2017	2010

Einfluss unabhängiger Variablen

```
# Wir lassen das angepasste Modell laufen
model_6K_year <- stm(documents = dfm_stm$documents,
                     vocab = dfm_stm$vocab,
                     K = 6,
                     prevalence = ~ Year_Start, #neu!
                     data = data)
```

```
effect <- estimateEffect(formula = ~ Year_Start,
                          stmobj = model_6K_year,
                          metadata = data)
```

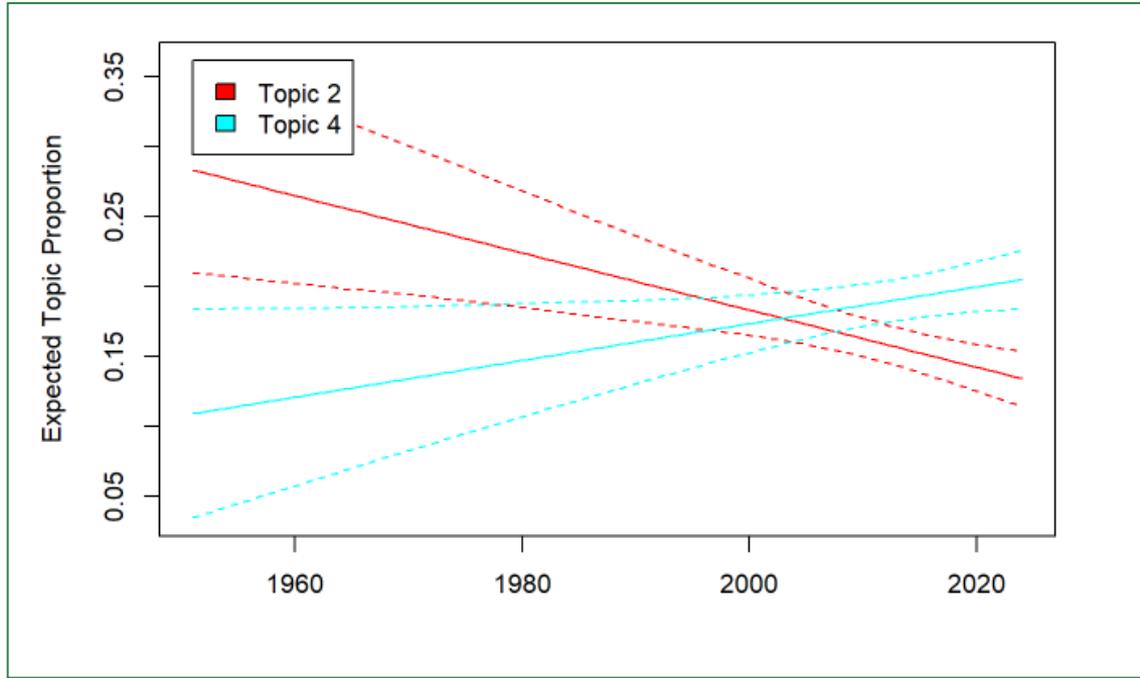
Einfluss unabhängiger Variablen

```
topics_6 %>%  
  select(`Topics 2`, `Topics 4`)
```

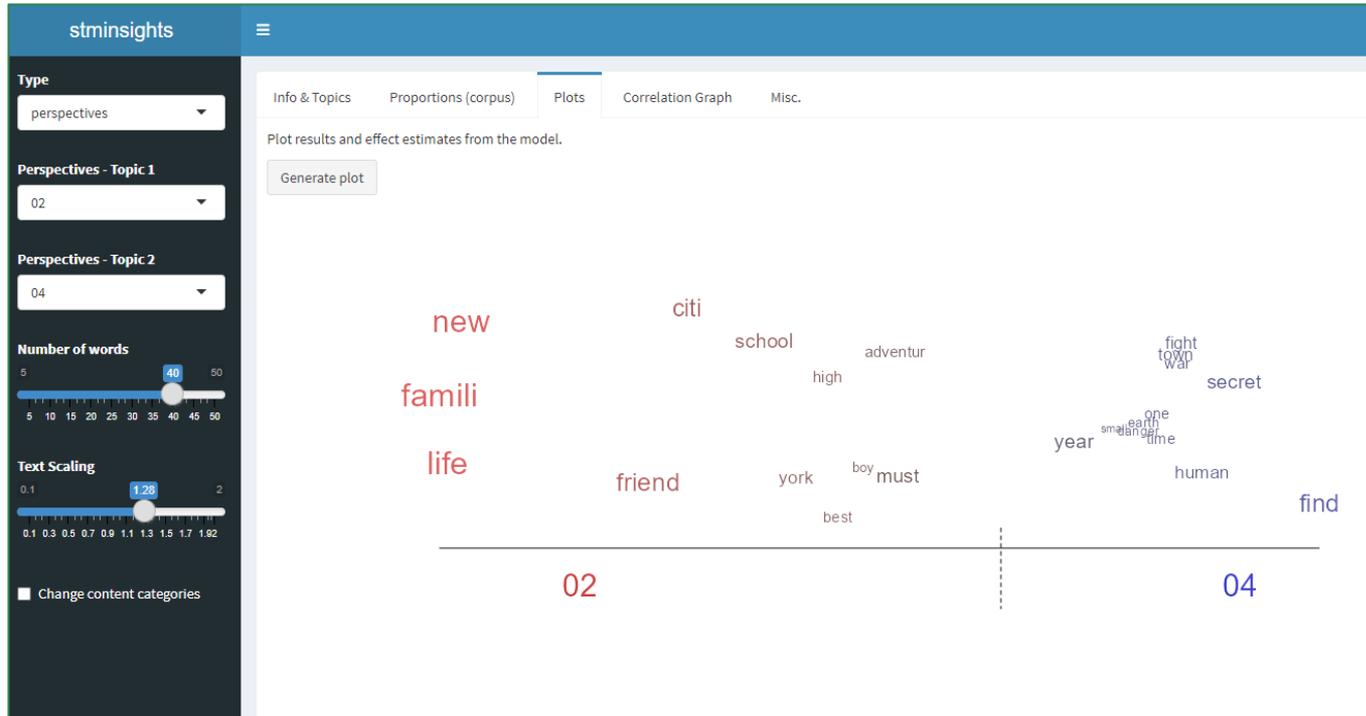
	Topics 2	Topics 4
1	best	human
2	school	alien
3	new	earth
4	boy	planet
5	hero	futur
6	york	struggl
7	mother	fight
8	return	galaxi
9	high	space
10	troubl	last

```
plot(effect, "Year_Start",  
      method = "continuous",  
      topics = c(2,4),  
      model = model_6K_year)
```

Einfluss unabhängiger Variablen



Weitere Visualisierung



Take-Aways



- **Background Topics:** Topics, die keine kohärente Bedeutung aufweisen oder für die eigene Studie keinen theoretischen Mehrwert bieten (z.B. Sprache statt Thema)
- **Top Features:** Features, welche ein Thema beschreiben
- **Top Documents:** Dokumente, welche ein Thema beschreiben

5. Anwendungsaufgabe



Jetzt seid ihr dran!



Könnt ihr...



Basis: Testen, wie sich das Modell verändert, wenn wir $K = 10$ Themen modellieren? Wird es besser oder schlechter?



Fortgeschritten: Mittels des Datensatzes zu Horoskopen testen, ob Zwillinge und Wassermänner andere Themen in ihren Horoskopen vorhergesagt kriegen?

Jetzt seid ihr dran!



AUFGABE 1.1 (BASIS)

Können ihr testen, wie sich das Modell verändert, wenn wir mit $K = 10$ Serien arbeiten? Wird es besser oder schlechter interpretierbar?

```
# dfm_stm$documents: Welche Dokumente nutzen wir?  
# dfm_stm$vocab: Welche Features nutzen wir?  
stat_fit <- searchK(dfm_stm$documents, dfm_stm$vocab,  
                   K = c(4, 6, 10), verbose = TRUE)
```

Jetzt seid ihr dran!

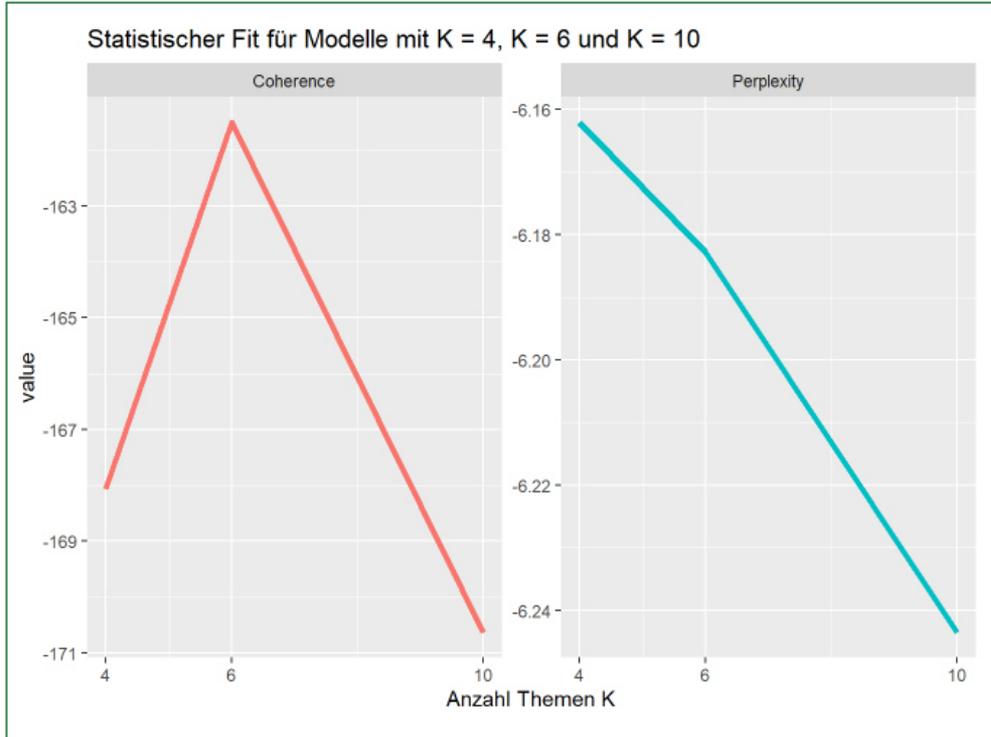


```
# Wir speichern die Ergebnisse im Objekt "Plot" ab
plot <- data.frame("K" = c(4, 6, 10),
                  "Coherence" = unlist(stat_fit$results$semcoh),
                  "Perplexity" = unlist(stat_fit$results$heldout))

# Wir wandeln das Format zu einem "long format" um
plot <- melt(plot, id = c("K"))

# Plot erstellen
ggplot(plot, aes(K, value, color = variable)) +
  geom_line(linewidth = 1.5, show.legend = FALSE) +
  scale_x_continuous(breaks = c(4, 6, 10)) +
  facet_wrap(~ variable, scales = "free_y") +
  labs(x = "Anzahl Themen K",
       title = "Statistischer Fit für Modelle mit K = 4, K = 6 und K = 10")
```

Jetzt seid ihr dran!



Jetzt seid ihr dran!



```
# Model mit K = 10 berechnen
model_10K <- stm(documents = dfm_stm$documents,
                 vocab = dfm_stm$vocab,
                 K = 10)
```

```
# Top Features
#für K = 10
topics_10K <- labelTopics(model_10K, n = 10)
topics_10 <- data.frame("features" = t(topics_10K$frex))
colnames(topics_10) <- paste("Topics", c(1:10))
topics_10
```

Jetzt seid ihr dran!



```
Topics 1 Topics 2 Topics 3 Topics 4 Topics 5 Topics 6 Topics 7
1 various mother lead find polic angel brother
2 part death human chang seri friend search
3 take suburban mission decad star los fall
4 mani unravel alien hunt small live attempt
5 trial student evil set around famili woman
6 host relationship earth peopl comic deal love
7 includ die journey supernatur comedi profession two
8 sexual murder cop futur antholog life control
9 intrigu real event career british navig becom
10 show sex hero surviv base misadventur made

Topics 8 Topics 9 Topics 10
1 use york agent
2 serial citi ident
3 solv late cia
4 killer new job
5 meet london agenc
6 name side assign
7 decid public danger
8 hospit author secret
9 uniqu drug protect
10 help save organ
```

Jetzt seid ihr dran!



AUFGABE 1.2 (FORTGESCHRITTEN)

Können Sie mittels des Datensatzes zu Horoskopen testen, ob Zwillinge und Wassermänner andere Themen in ihren Horoskopen vorhergesagt kriegen?

(Daten einlesen & Preprocessing: haben wir uns hier mal geschenkt)

Jetzt seid ihr dran!



```
# Berechnung eines Modells, hier probeweise mit K = 5 Themen  
# Wir nutzen das Sternzeichen als unabhängige Variable
```

```
model_horoscope <- stm(documents = dfm,  
                        vocab = dfm_stm  
                        K = 4,  
                        prevalence = ~,  
                        data = data,  
                        verbose = F)
```

```
# Ein erster Blick in die Top Features  
topics_horoscope <- labelTopics(model_  
topics_horoscope <- data.frame("featur  
colnames(topics_horoscope) <- paste("T  
topics_horoscope
```

	Topics 1	Topics 2	Topics 3	Topics 4
1	person	feel	though	rather
2	financi	peopl	success	keep
3	matter	accept	just	go
4	improv	like	get	may
5	well	work	friend	sure
6	relationship	give	let	quick
7	life	ideal	strong	difficulti
8	now	although	far	busi
9	care	aquarian	next	way
10	situat	practic	away	eye

Jetzt seid ihr dran!



```
#Schauen wir uns den Effekt an
effect <- estimateEffect(formula = ~ Signs,
                        stmobj = model_horoscope,
                        metadata = data)

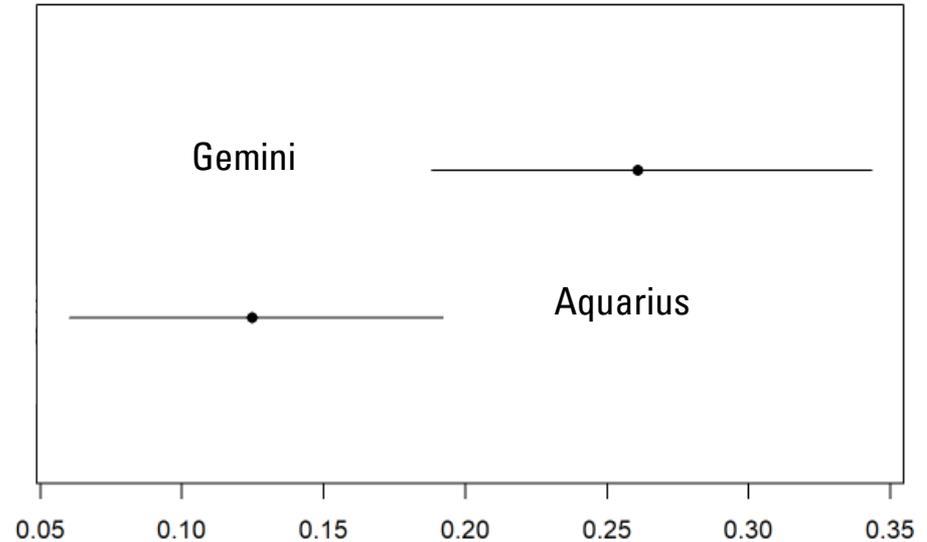
# Und visualisieren wir den Unterschied
plot(effect, "Signs",
      method = "pointestimate",
      topics = c(4),
      model = model_horoscope)
```

Jetzt seid ihr dran!



```
#Schauen wir uns den Effekt an
effect <- estimateEffect(formula = ~ Signs
                          stm
                          met

# Und visualisieren wir den
plot(effect, "Signs",
      method = "pointestimate",
      topics = c(4),
      model = model_horoscope
```



6. Outro

Chancen von Topic Models

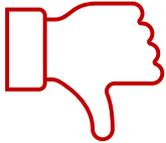


- Explorative Analyse von Themen (?) in großen Korpora
- Nutzung z.B. für nachfolgende qualitative & quantitative Methoden

Chancen von Topic Models



- Explorative Analyse von Themen (?) in großen Korpora
- Nutzung z.B. für nachfolgende qualitative & quantitative Methoden



- Gefahr der Überinterpretation & Frage nach Theorie: Was bedeuten «Themen»?
- Viele Freiheitsgrade bei methodischen Entscheidungen, die dokumentiert werden müssen
- Funktioniert weniger gut für z.B. kurze Texte
- Können extrem aufwendig sein! (Interpretation, Validierung, etc.)

Andere Verfahren: Keyword-Assisted TM



ARTICLE

Keyword-Assisted Topic Models

Shusei Eshima, Kosuke Imai ✉, Tomoya Sasaki

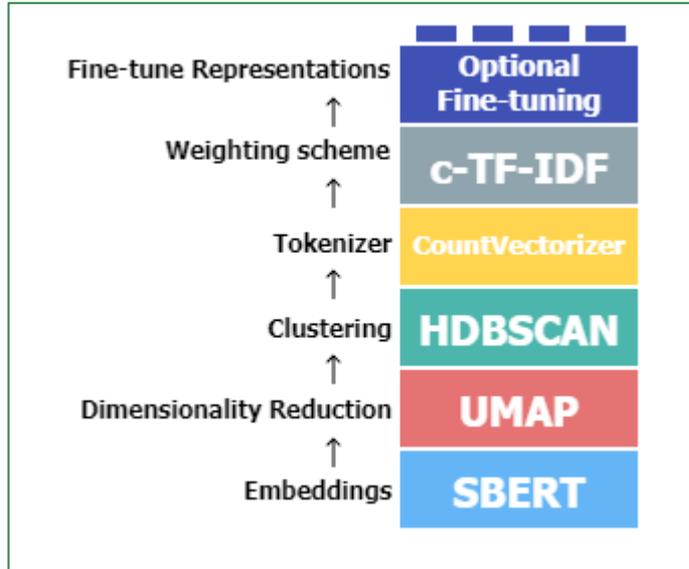
First published: 01 April 2023 | <https://doi.org/10.1111/ajps.12779> | Citations: 6

The proposed methodology is implemented via an open-source software package keyATM, which is available at <https://cran.r-project.org/package=keyATM>. We thank Doug Rice and Yutaka Shinada for sharing their data, Luwei Ying, Jacob Montgomery, and Brandon Stewart for sharing with us their experience of setting up validation exercises, and Soichiro Yamauchi for advice on methodological and computational issues. We also thank Soubhik Barari, Matthew Blackwell, Max Goplerud, Andy Halterman, Masataka Harada, Hiroto Katsumata, Gary King, Dean Knox, Shiro Kuriwaki, Will Lowe, Luke Miratrix, Hirofumi Miwa, Daichi Mochihashi, Santiago Olivella, Yon Soo Park, Reed Rasband, Hunter Rendleman, Sarah Mohamed, Yuki Shiraito, Tyler Simko, and Diana Stanescu, as well as seminar participants at the Institute for Quantitative Social Science Applied Statistics Workshop, the Japanese Society for Quantitative Political Science 2020 Winter Meeting, International Methods Colloquium, Annual Conference of the Society for Political Methodology (2020), and Annual Conference of the American Political Science Association (2020) for helpful discussions and comments on this project. Lastly, we thank the editors and our three anonymous reviewers for providing us with additional comments.

Keyword-Assisted Topic Modeling (Eshima et al., [2024](#))

- Theoretisch informierte Wahl von K mittels Seed-Wörtern
- Themen können z.B. deduktiv abgeleitet werden
- In R: keyATM Paket

Andere Verfahren: BERTopic



Quelle: Grootendorst, [2022](#)

BERTopic (Grootendorst, [2022](#))

- Basiert auf Transformer-Modellen, die bag-of-words Annahme umgehen
- Textdaten: semantische Repräsentation (im Prinzip: als Vektoren in einem n-dimensionalen Raum) → **Sitzung 6**
- Dann ähnlich: Dimensionsreduktion & Clustering für Identifikation von Topics
- Leichter in Python umzusetzen (s. [Tutorial](#))

Wie berichte ich Topic-Modeling in Papern?

Before running the model, researchers have to decide on the number of topics K that should be estimated. Models with $10 < K < 150$ in increments of $K = 5$ were evaluated concerning the suitability of topics (e. g., internal coherence, exclusivity of topics), their substantivity, and robustness. In a discussion, the research team decided on a model with $K = 85$ topics.

← Modell-
Einstellungen

Beispiel aus Hase et al. ([2021](#)). Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). *Global Environmental Change*.

Wie berichte ich Topic-Modeling in Papern?

Before running the model, researchers have to decide on the number of topics K that should be estimated. Models with $10 > K > 150$ in increments of $K = 5$ were evaluated concerning the suitability of topics (e. g., internal coherence, exclusivity of topics), their substantivity, and robustness. In a discussion, the research team decided on a model with $K = 85$ topics. Next, members of the research team were supplied with information on each topic, for example its top terms, a random sample of articles representing the topic, and its robustness (Supplementary Material, Appendix C). They then coded which topics to keep and which to exclude ($\alpha = 0.71$). As we are interested in cross-national comparisons, we excluded topics driven by a single country based on the Hirschman-Herfindahl Index ($HH > 0.8$) (Maier et al., 2018). 46 topics were kept for further analysis.

← Modell-
Einstellungen

← Identifikation
relevanter
Themen

Beispiel aus Hase et al. (2021). Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018). *Global Environmental Change*.

Wie berichte ich Topic-Modeling in Papern?

In repeated rounds of discussions, we then decided on labels describing each topic. We also discussed overarching themes/dimensions each topic could be sorted into. Discussions were informed by previous studies, for example descriptions of the societal dimension (Painter and Schäfer, 2018) or themes such as climate science or environmental impacts/changes (e.g., Boykoff, 2008; McComas and Shanahan, 1999; Hoffman, 2011). However, deduced themes/dimensions were extended and revised inductively through the material at hand in an interactive, interpretative process.



Interpretation
& Labelling

Beispiel aus Hase et al. (2021).

Wie berichte ich Topic-Modeling in Papern?

In repeated rounds of discussions, we then decided on labels describing each topic. We also discussed overarching themes/dimensions each topic could be sorted into. Discussions were informed by previous studies, for example descriptions of the societal dimension (Painter and Schäfer, 2018) or themes such as climate science or environmental impacts/changes (e.g., Boykoff, 2008; McComas and Shanahan, 1999; Hoffman, 2011). However, deduced themes/dimensions were extended and revised inductively through the material at hand in an interactive, interpretative process. Based on this process, each topic was sorted into one out of seven overarching themes and, as a more aggregated measure, one out of three dimensions: the scientific dimension consisting of one theme (*Climate Science*), the ecological dimension consisting of another (*Climate Change & Impacts on the Ecosystem*), and the societal dimension consisting of five themes (*Causes of & Solutions to Climate Change, Climate Politics, Awareness & Education, Impacts on Humans, Economic Impacts*). While some call these aggregated categories frames, we consider frames to entail more complex theoretical concepts which can often not be easily identified automatically (Nicholls and Culpepper, 2020).



Interpretation
& Labelling



Gruppierung
in Kontexte

Beispiel aus Hase et al. (2021).

Aufwendig? Ja!

Und mit Qualitätskriterien haben wir noch nicht mal angefangen

3.3.3. *Validity & replicability*

Scholars have pointed out important limitations of topic modeling (Brookes and McEnergy, 2019; Grundmann, 2021; Maier et al., 2018), for instance a lack of linguistic sensitivity. To reassure linguistic sensitivity, we followed recent recommendations (Brookes and McEnergy, 2019; Song et al., 2020). At least ten articles related to each topic were read by every member of the research team before labeling and interpretation. Moreover, results were validated manually based on two validation sets ($F_1 = 0.74$ and $F_1 = 0.76$ for classification of dimensions). Results showed not overly high, but sufficient validity scores except for the theme *Economic Impacts*, which should thus be interpreted with caution. Another limitation relates to the replicability and robustness of results, for instance models converging to different solutions. To reassure replicability, we employed spectral learning as a deterministic method for initialization (Roberts et al., 2016). We also checked the robustness of results independent of parameter settings, here topics being reproduced for other choices for K (Wilkerson and Casas, 2017). Detailed information on these tests can be found in the Supplementary Material (Appendix D). We agree that a final limitation – the theoretical underpinnings of topics – still applies (Brookes and McEnergy, 2019; Grundmann, 2021; Maier et al., 2018) as is discussed later.

Beispiel aus Hase et al. (2021).



Wie berichte ich Topic Modeling in Papern?

- **Immer:** Relevante Schritte kurz nennen & im Appendix ausführen
 - Wahl von K ?
 - Identifikation (irr-)relevanter Themen & Interpretation?
- **Noch besser:** Code (und ggf. Daten) teilen
- **Am besten:** Mit Multiverse-Analysen testen, wie robust Ergebnisse bei verschiedenen Preprocessing-Schritten & Hyperparametern bleiben und Analysen validieren (s. **Sitzung 5!**)

Wie geht es weiter?

ZEITPLAN

 Mi, 24. Juli

- 09:00 - 12:00: **1** *Einführung & Preprocessing*
- 12:00 - 13:00:  *Mittagspause*
- 13:00 - 15:00: **2** *Co-Occurrence-Analysen*
- 15:00 - 17:00: **3** *Diktionäre*

 Do, 25. Juli

- 09:00 - 12:00: **4** *Topic Modeling*
- 12:00 - 13:00:  *Mittagspause*
- 13:00 - 15:00: **5** *Qualitätskriterien*
- 15:00 - 16:00: **6** *Ausblick*

Danke! Fragen?



Dr. Valerie Hase
IfKW, LMU Munich



valeriehase



valerie-hase.com



Luisa Kutlar
IfKW, LMU Munich



luisakutlar