maier_2018_applying_lda_topic_modeling_in_communication_research_toward_a_valid_and_reliable_methodology

Year

2018

Author(s)

Daniel Maier and A. Waldherr and P. Miltner and G. Wiedemann and A. Niekler and A. Keinert and B. Pfetsch and G. Heyer and U. Reber and T. H{\"a}ussler and H. Schmid-Petri and S. Adam

Title

Applying LDA topic modeling in communication research: Toward a valid and reliable methodology

Venue

Communication Methods and Measures

Topic labeling

Manual

Focus

Secondary

Type of contribution

Established approach

Underlying technique

Manual labeling assisted by associated documents and generated summaries

Topic labeling parameters

\

Label generation

The authors decided that 13 topics should be removed because they showed no indication of being either meaningful or coherent. The remaining 37 topics were subject to the final validation and labeling step.

we reviewed the document-topic distributions from θ for the remaining topics. Ten randomly sampled documents were read, all containing relatively large proportions of the respective topic ($\theta_{d,k} > .5$).

For the sampled topics, brief summary descriptions of their content were written, and suggestions about the topic labels were proposed. Subsequently, the researchers deliberately decided in a discussion
(a) whether a topic was semantically coherent and, thus, a valid topic in theoretical terms and
(b) what label should be given to the topic.

The label is the product of determining what catches the notion of the underlying concept, in our case the “issues,” most concisely.
In this phase of in-depth investigation, eight of the remaining 37 topics were further discarded because they either did not reveal a coherent semantic meaning or solely represented contents from a single website unconnected to aspects of the debate about food safety.
Thus, 29 validated, manually labeled “issues” in the food-safety debate remained.

Table 1. Validated topic model for the online text corpus about food safety in the U.S.

K	Label	Share % *M* (SD)	HHI *M* (SD)	Top-5 Words
Agriculture				
25	GM Food	3.94 (0.90)	0.04 (0.01)	food, label, genetically, monsanto, gmo
9	Organic Farming	2.58 (0.37)	0.02 (0.00)	organic, food, farm, farmer, agriculture
20	Livestock	2.55 (0.18)	0.03 (0.00)	meat, food, animal, beef, milk
10	Antibiotics	2.21 (0.46)	0.10 (0.02)	antibiotic, animal, health, drug, human
Consumption and Protection				
22	Foodborne Diseases	4.06 (1.34)	0.06 (0.02)	food, outbreak, salmonella, illness, report
8	FS Regulation	3.48 (0.40)	0.04 (0.01)	food, fda, safety, product, consumer
7	Contaminated Food	2.77 (0.63)	0.04 (0.01)	safety, recall, produce, fda, outbreak
29	Food Consumption	2.26 (0.14)	0.03 (0.01)	product, company, consumer, store, sell
27	Restaurant Inspection	2.14 (0.98)	0.09 (0.04)	food, restaurant, safety, health, inspection
16	Tap Water	1.53 (1.03)	0.22 (0.23)	water, food, public, protect, watch
39	BPA-packaging	1.50 (0.83)	0.15 (0.11)	chemical, bpa, safe, toxic, health
Science and Technology				
6	Health Reports	3.48 (0.25)	0.02 (0.00)	health, report, public, risk, datum
19	Chemicals	2.28 (0.28)	0.02 (0.00)	study, chemical, level, health, human
37	GM Technology	1.84 (0.12)	0.02 (0.00)	research, test, science, article, study
Environment				
44	Bees and Pesticides	3.14 (1.90)	0.41 (0.28)	bee, pesticide, epa, food, center
43	Environment	1.41 (0.28)	0.05 (0.02)	read, fish, salmon, environment, specie
50	Fracking	1.37 (0.30)	0.04 (0.02)	energy, gas, oil, water, environmental
31	Climate Change	1.34 (0.22)	0.03 (0.01)	climate, change, report, world, warm
Personal Health and Wellbeing				
21	(Un)healthy Diet	2.32 (0.44)	0.04 (0.01)	food, fat, sugar, diet, health
35	Health and Nutrition	2.31 (0.24)	0.01 (0.01)	program, community, work, education, child
38	Recipes	2.26 (0.41)	0.03 (0.01)	cook, eat, meat, make, recipe
1	School Food	2.00 (0.52)	0.17 (0.08)	food, school, pew, safety, project
12	Dietary Therapy/Prevention	1.42 (0.18)	0.03 (0.01)	cancer, disease, woman, blood, child
42	Medical Information	1.29 (0.39)	0.07 (0.08)	doctor, medicine, take, day, skin
Background Topics				
14	Politics	2.65 (0.28)	0.03 (0.01)	bill, state, obama, law, house
11	Economy	2.50 (0.29)	0.02 (0.01)	company, market, country, million, u.s.
24	Law and Order	2.20 (0.34)	0.02 (0.00)	report, year, police, official, court
2	Infectious Diseases	2.03 (0.62)	0.06 (0.02)	health, coli, pet, animal, case
48	Health Care	1.07 (0.46)	0.13 (0.11)	drug, health, care, medical, patient

Note. HHI = Hirschman-Herfindahl-Index; GM = genetically modified; BPA = Bisphenol A; FS = food safety; K = index of the topic.

Motivation

describing the substantive content of the topic

Topic modeling

LDA

Topic modeling parameters

Nr of topics: {30, 50, 70}

α: 0.5

β: 0.02

Nr. of topics

30 (UK), 50 (Germany), 50 (US)

Label

Representing an “issue” in the food-safety debate

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic modeling
Dataset: Food safety

Problem statement

In applying LDA to textual data, researchers need to tackle at least four major challenges that affect these criteria: (a) appropriate pre-processing of the text collection; (b) adequate selection of model parameters, including the number of topics to be generated; (c) evaluation of the model’s reliability; and (d) the process of validly interpreting the resulting topics. We review the research literature dealing with these questions and propose a methodology that approaches these challenges. Our overall goal is to make LDA topic modeling more accessible to communication researchers and to ensure compliance with disciplinary standards.

Corpus

Origin: Various websites concerned with the issue of food safety
Nr. of documents: 344,456 (87,692 after processing)
Details:

- from June 2012 to November 2014 (30 months)
- food-safety-related content from Germany, the U.K., and the U.S.,

Document

Webpage document content

Pre-processing

- Removal of boilerplate content, such as navigation bars, page markups, ads, teasers, and other items regarded as irrelevant.
- filtered for relevant content by only including those documents containing the (combination of) issue-specific key terms
- document duplicate detection
- tokenisation
- all capital letters are converted to lowercase
- removal of punctuation and special character
- remove stop-words
- lemmatization
- strip very rare and extremely frequent word occurrences

```
@article{maier_2018_applying_lda_topic_modeling_in_communication_research_toward_a_valid_and_reliable_methodology,
  author = {Daniel Maier and A. Waldherr and P. Miltner and G. Wiedemann and A. Niekler and A. Keinert and B. Pfetsch and G. Heyer and U. Reber and T. H{\"a}ussler and H. Schmid-Petri and S. Adam},
  date-added = {2023-04-08 20:54:13 +0200},
  date-modified = {2023-04-08 20:54:13 +0200},
  doi = {10.1080/19312458.2018.1430754},
  journal = {Communication Methods and Measures},
  month = {feb},
  number = {2-3},
  pages = {93--118},
  publisher = {Informa {UK} Limited},
  title = {Applying {LDA} Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology},
```

```
  url = {https://doi.org/10.1080%2F19312458.2018.1430754},
  volume = {12},
  year = 2018}
```

#Thesis/Papers/BS