

Topics (Automated Content Analysis)

AUTHOR

Valerie Hase

KEYWORDS

topic, issue, agenda setting, news value, supervised machine learning, LDA topic modeling, structural topic modeling

BRIEF DESCRIPTION

Topics describe the main issue discussed in an article, for example: Does an article deal with politics, economics or sports?

FIELD OF APPLICATION/THEORETICAL FOUNDATION

In the context of “*Agenda Setting*”, studies analyze which issues are on the public agenda. In the context of “*News Values*”, studies may analyze why some topics are more prominently covered than others.

REFERENCES/COMBINATION WITH OTHER METHODS OF DATA COLLECTION

Many studies combine manual inspection of topics with their automated detection. Quinn et al. (2010) demonstrate for their analyses of legislative speeches how manual inspection may increase the validity of results. Similarly, Hase et al. (2020) use automated content analysis to find and map similar topics for which manual coding is then conducted. Such combinations may contribute to a better and more detailed understanding of topics than automated analyses by themselves.

The datasets referred to in the table are described in the following paragraph:

Puschmann (2019a) uses New York Times articles (1996-2006, N = 30,862) as well as articles from Die Zeit (2011-2016, N = 377) to identify topics using supervised machine learning. In an

other tutorial, Puschmann (2019b) uses Sherlock Holmes stories (18th century, N = 12), articles from Die Zeit (2011-2016, N = 377) and debate transcripts (1970-2017, N = 7,897) to apply LDA and structural topic modeling. In her tutorials, Silge (2018a, 2018b) also uses Sherlock Holmes stories (18th century, N = 12) and a news corpus also containing comments (2006-ongoing, N = 100,000). Silge and Robinson (2020) apply LDA topic modeling on news stories by the Associated Press (1992, N = 2,246) as well as books by Dickens, Wells, Verne and Austen (18th century, N = 4). Roberts et al. (2019) use blogposts (2008, N = 13,248) for structural topic modeling. Watanabe and Müller (2019) apply LDA topic modeling on newspaper articles from The Guardian (2016, N = 6,000). Van Atteveldt and Welbers (2019, 2020) use State of the Union speeches (1981-2017, N = 10 and 1789-2017, N = 58) for their analyses. Lastly, Wiedemann and Niekler (2017) use the same data containing State of the Union speeches (1790-2017, N = 223).

REFERENCES:

- Hase, V., Engelke, K., Kieslich, K. (2020). The things we fear. Combining automated and manual content analysis to uncover themes, topics and threats in fear-related news. *Journalism Studies*, 21(10), 1384-1402.
- Puschmann, C. (2019). Automatisierte Inhaltsanalyse mit R. Retrieved from <http://inhaltsanalyse-mit-r.de/index.html>
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1), 209-228.



<https://doi.org/10.34778/1e>

© 2021, the authors. This work is licensed under the “Creative Commons Attribution – NonCommercial – NoDerivatives 4.0 International” license (CC BY-NC-ND 4.0)

- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Model. *Journal of Statistical Software*, 91(2), 1–40.
- Silge, J. (2018a). The game is afoot! Topic modeling of Sherlock Holmes stories. Retrieved from <https://juliasilge.com/blog/sherlock-holmes-stm/>
- Silge, J. (2018b). Training, evaluating, and interpreting topic models. Retrieved from <https://juliasilge.com/blog/evaluating-stm/>
- Silge, J., & Robinson, D. (2020). Text Mining with R. A tidy approach. Retrieved from <https://www.tidytextmining.com/>
- van Atteveldt, W., & Welbers, K. (2019). Structural Topic Modeling. Retrieved from https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/r_text_stm.md
- van Atteveldt, W., & Welbers, K. (2020). Fitting LDA models in R. Retrieved from https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/r_text_lda.md
- Watanabe, K., & Müller, S. (2019). Quanteda tutorials. Retrieved from <https://tutorials.quanteda.io/>
- Wiedemann, G., Niekler, A. (2017). Hands-on: a five day text mining course for humanists and social scientists in R. Proceedings of the 1st Workshop Teaching NLP for Digital Humanities (Teach4DH@GSCL 2017), Berlin. Retrieved from <https://tm4ss.github.io/docs/index.html>

Table 1. Measurement of „Topics“ using automated content analysis.

Author(s)	Sample	Procedure	Formal validity check with manual coding as benchmark*	Code
Puschmann (2019a)	(a) Newspaper articles (b) Newspaper articles	Supervised machine learning	Reported	http://inhaltsanalyse-mit-r.de/maschinelles_lernen.html
Puschmann (2019b)	(a) Sherlock Holmes stories (b) Newspaper articles (c) United Nations General Debate Transcripts	LDA topic modeling; structural topic modeling	Not reported	http://inhaltsanalyse-mit-r.de/themenmodelle.html
Silge (2018a) & Silge (2018b)	(a) Sherlock Holmes stories (b) News stories and comments	Structural topic modeling	Not reported	https://juliasilge.com/blog/sherlock-holmes-stm/ https://juliasilge.com/blog/evaluating-stm/
Silge & Robinson (2020)	(a) News articles (b) Books	LDA topic modeling	Not reported	https://www.tidytextmining.com/topicmodeling.html

Author(s)	Sample	Procedure	Formal validity check with manual coding as benchmark*	Code
Roberts et al. (2019)	Blogposts	Structural topic modeling	Not reported	https://www.jstatsoft.org/article/view/v091i02
Watanabe & Müller (2019)	Newspaper articles	LDA topic modeling	Not reported	https://tutorials.quanteda.io/machine-learning/topicmodel/
van Atteveldt & Welbers (2019)	State of the Union speeches	Structural topic modeling	Not reported	https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/r_text_stm.md
van Atteveldt & Welbers (2020)	State of the Union speeches	LDA topic modeling	Not reported	https://github.com/ccs-amsterdam/r-course-material/blob/master/tutorials/r_text_lda.md
Wiedemann & Niekler (2017)	State of the Union speeches	LDA topic modeling	Not reported	https://tm4ss.github.io/docs/Tutorial_6_Topic_Models.html
Wiedemann & Niekler (2017)	State of the Union speeches	Supervised machine learning	Reported	https://tm4ss.github.io/docs/Tutorial_7_Klassifikation.html

* Please note that many of the sources listed here are tutorials on how to conduct automated analyses – and therefore not focused on the validation of results. Readers should simply read this column as an indication in terms of which sources they can refer to if they are interested in the validation of results.