Unsupervised classification

How would you describe what you see here?

- coffee
- cafe
- espresso
- latte
- barista
- beans
- brew
- cappuccino
- aroma
- roast

- programming
- code
- software
- development
- algorithm
- python
- function
- variable
- debugging
- Java

- fitness
- exercise
- health
- workout
- gym
- nutrition
- weight
- strength
- cardio
- yoga

Topic 1: Label?	Topic 2: Label?	Topic 3: Label?
• coffee • cafe • espresso • latte • barista • beans • brew • cappuccino • aroma • roast	• programming • code • software • development • algorithm • python • function • variable • debugging • Java	• fitness • exercise • health • workout • gym • nutrition • weight • strength • cardio • yoga

Types of machine learning

1. Supervised
    - An outcome variable is defined
    - Focus is on prediction
2. Unsupervised
    - No outcome variable has been defined
    - Focus is on patterns

What can you do with unsupervised techniques?

- **Discovery**

What can you do with unsupervised techniques?

Scaling and clustering:

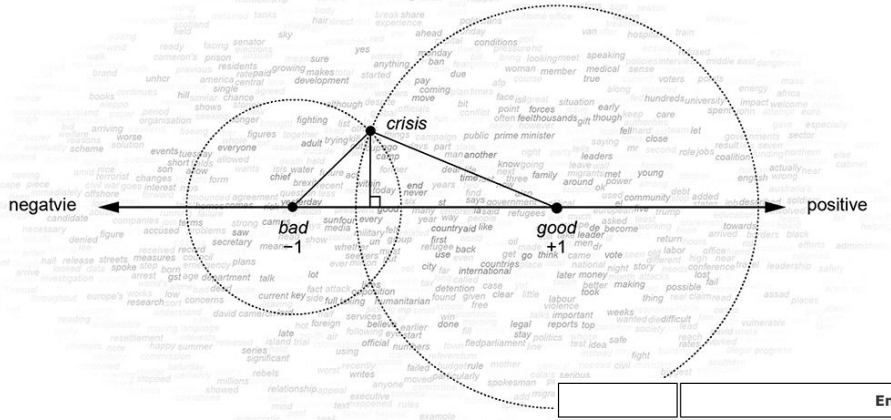
- Can the documents be placed on a **dimension**? What **groups** of documents are in the corpus?
- Example methods: Latent semantic scaling, wordfish, cluster analysis, …

Topic modeling:

- What are the main **topics** in a corpus?
- Example methods: LDA, STM, BERTopic

Grimmer et al. 2022, Barberá, 2018

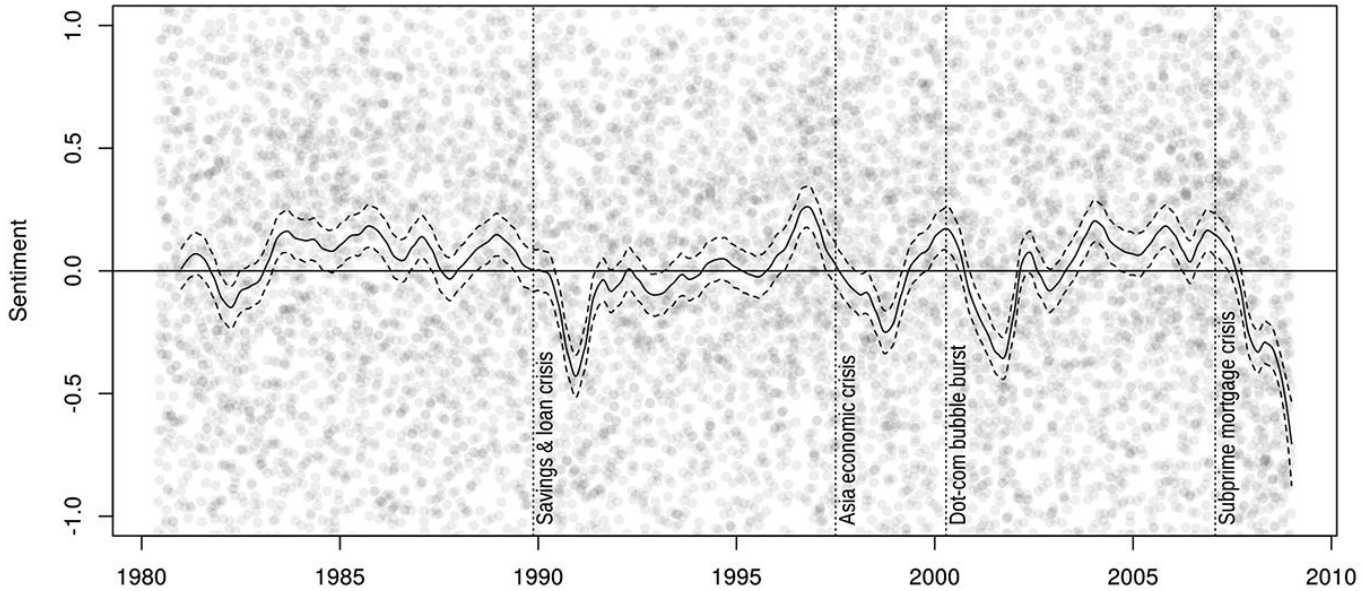
Example: Latent Semantic Scaling



	English (Turney & Littman, 2003)	Japanese (by author)
Positive	good, nice, excellent, positive, fortunate, correct, superior	絶好, 美麗, 秀逸, 卓越, 優雅, 絶賛, 善良
Negative	bad, nasty, poor, negative, unfortunate, wrong, inferior	粗悪, 醜悪, 稚拙, 非礼, 貧相, 酷評, 悪徳

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, *15*(2), 81-102.

Longitudinal analysis of news articles on the economy (English) in the New York Times corpus by LSS.



Curves are LOESS smoothed sentiment scores with 95% confidence intervals.

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, *15*(2), 81-102.

Latent Semantic Scaling (LSS)

Semi-supervised (uses seed words).

Estimates a latent dimension (like sentiment or political ideology) from text by:

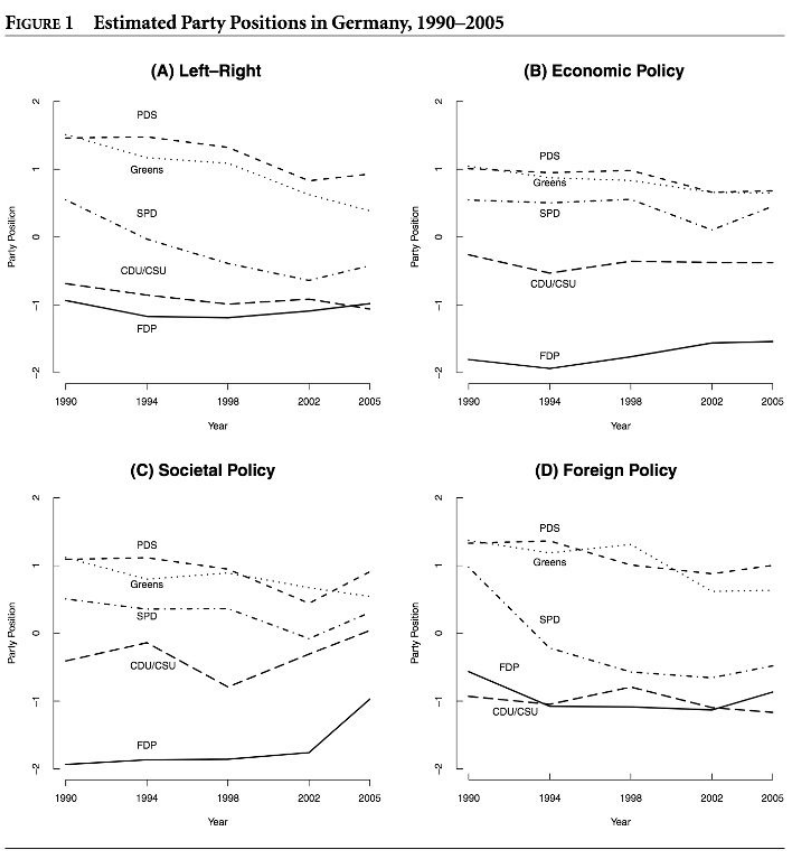
1. Creating a vector space model of word meanings
2. Using dfm + a set of seed words that represent poles of the scale (e.g., *good–bad*, *liberal–conservative*).
3. Computing semantic similarity between each word and the seed words.
4. Aggregating those similarities into document-level scores

It's more nuanced than just counting dictionary hits, it uses the semantic context of words.

Watanabe, K. (2021). Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, *15*(2), 81-102.

Example wordfish

- No seed words
- You don't choose what the scale represents
- finds the dimension that best explains variation in word usage across documents
- Use Case: Political scaling of manifestos or speeches



Slapin and Proksch, 2008: 714.

Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), 705-722.

What can you do with unsupervised techniques?

Scaling and clustering:

- Can the documents be placed on a **dimension**? What **groups** of documents are in the corpus?
- Example methods: Latent semantic scaling, wordfish, cluster analysis, …

Topic modeling:

- What are the main **topics** in a corpus?
- Example methods: LDA, STM, BERTopic

Grimmer et al. 2022, Barberá, 2018

Topic Modelling

- A model to discover **latent topics** in a corpus
- Can be used to organize the collection according to the discovered themes
- **Requires little prior information**, no training set, or human annotation – only a decision on K (number of topics)
- Most common: Latent Dirichlet Allocation (LDA)
  - Bayesian generative hierarchical model
  - First introduced as a way to detect the presence of structured genetic variation (Pritchard, 2000)
  - Adjusted for text analysis ML applications (Blei et al., 2003)

Example 1: Heidenreich et al., 2019

Original:

1

elnök, katonai, orosz, államfő, nemzetközi, külügyminiszter, emberi, török, humanitárius, védelmi, jogi, szervezet, erők, közölte, civil, kormány, amerikai, biztonsági, ország, részt,

---

2

százalékkal, forint, milliárd, millió, százalékos, százalék, száma, százaléka, legnagyobb, összesen, gazdasági, menekültek, ezren, uniós, külföldi, kormány, pénzügyi, csaknem, adatai, adatok,

---

3

kancellár, német, párt, százalék, tartományi, kormányfő, százaléka, szociáldemokrata, pártja, jobboldali, politikai, európai, miniszterelnök, elnöke, politikus, parlamenti, liberális, uniós, közvélemény, százalékos,

---

4

keresztény, kulturális, vallási, európai, társadalmi, világ, politikai, emberi, magyarok, gazdasági, nemzeti, nyugati, emberek, társadalom, menekültek, fontos, terrorizmus, hangsúlyozta, történelmi, keleti,

Example 1: Heidenreich et al., 2019

HU - Google Translate

1
president, military, russian, head of state, international, foreign minister, human, turkish, humanitarian, defense, law, organization, forces, communicated, civil, government,

---

2
percent, percent, number, percent, largest, total, economic, refugees, thousand, EU, foreign, government, financial, almost, data, data,

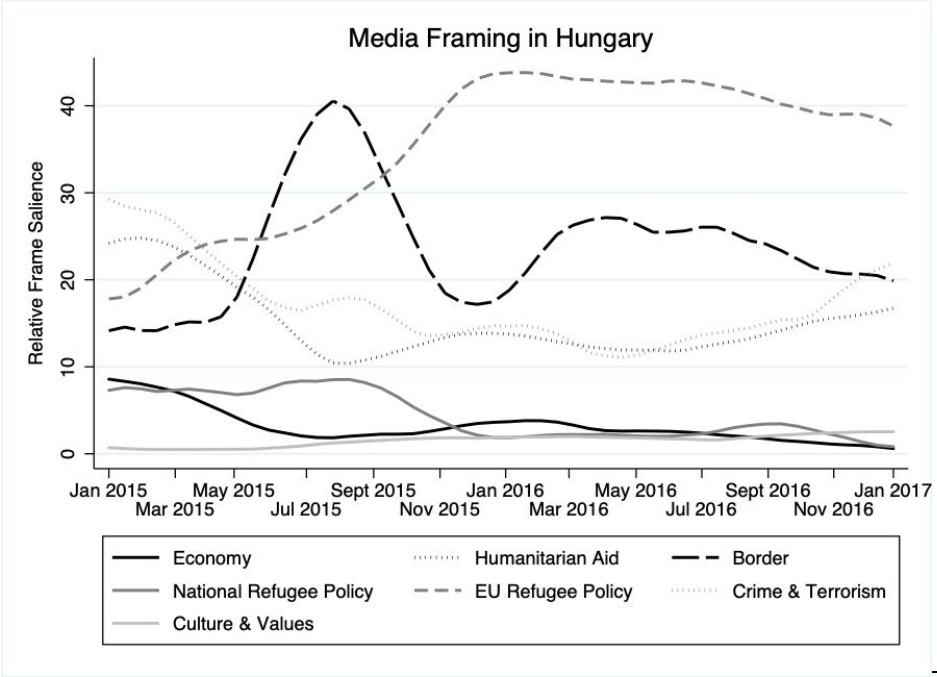
---

3
chancellor, German, party, percentage, provincial, leader, percentage, socialdemocrat, party, right, political, European, prime minister, president, politician, parliamentary, liberal,

---

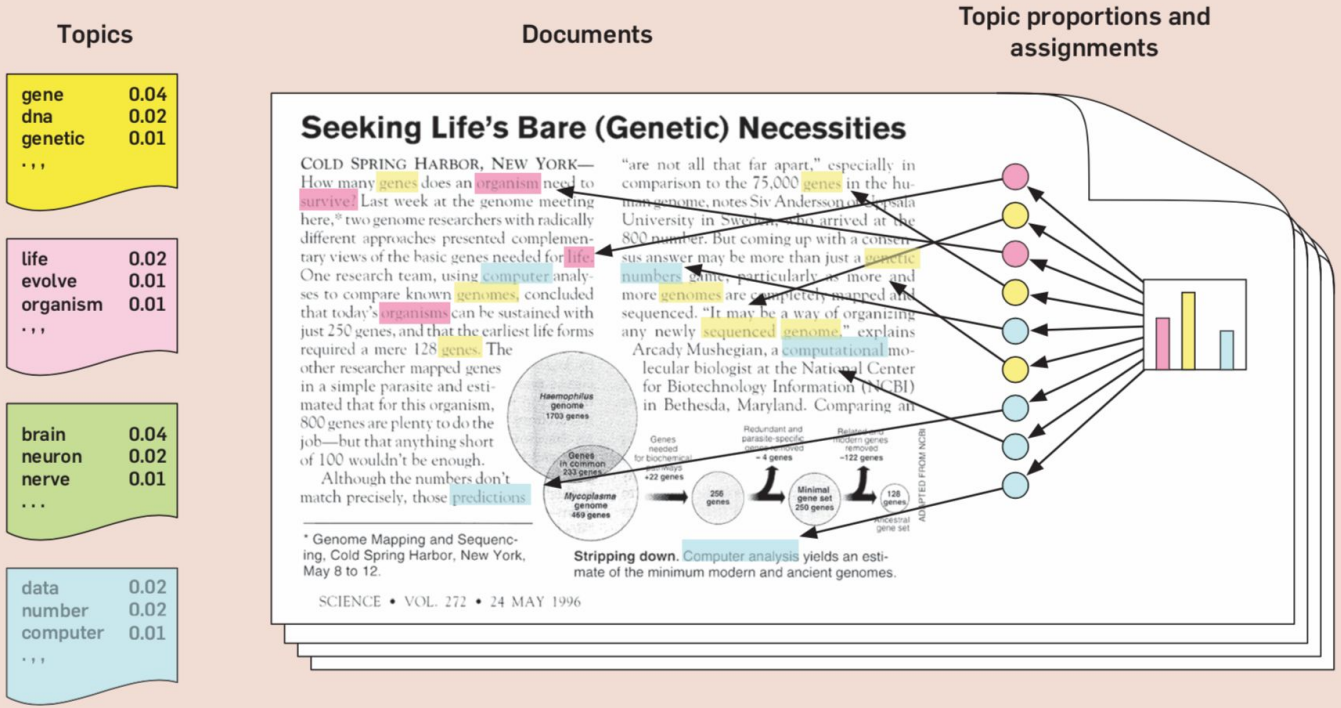
4
cultural, religious, European, social, world, political, human, hungarian, economic, national, western, people, society, refugees, important, terrorism, stressed, historical,

Example 1: Heidenreich et al., 2019



LDA

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



(Blei et al., 2003)

Latent Dirichlet Allocation (LDA)

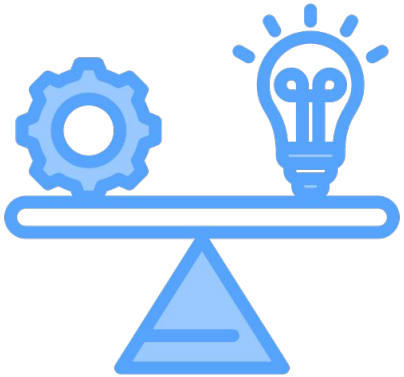
- **Initialization:** LDA starts by randomly assigning each word in each document to one of the topics. These initial assignments serve as a starting point for the algorithm.
- **Iteration:** LDA iteratively updates the topic assignments of words and estimates the topic distributions. In each iteration, it goes through each word in each document and reassigns it to a topic based on the current topic distributions and the observed word frequencies.
- **Inference:** As the iterations progress, LDA gradually infers the most likely topic assignments for each word and the overall topic distributions across the document collection. This inference process continues until the algorithm converges or reaches a predefined stopping criterion.

van Atteveld et al., 2022

Meaning of topics obtained with LDA?

- topics do not necessarily match theoretical concepts of what a topic is defined as

- topic models identify the most discriminatory groups of documents automatically
- topics are created by the LDA algorithm based on patterns of word co-occurrence in documents



Jacobi et al. 2015 DOI: 10.1080/21670811.2015.1093271

How to: Steps

1. Data selection
2. Data preprocessing (what features are good topic representations?)
3. Applying LDA
4. Once the iterations are complete, LDA provides 2 outputs:
   a. the estimated topic distributions for each document
   b. the word distributions for each topic

   These distributions give insights into the main topics present in the document collection
5. Validation: Inspecting top words per topic + the top topics per document; and many more methods

Example: Topic model

- **Data:** 51,528 news stories, NY Times coverage of nuclear technology, published between 1945 and 2013, search terms “nuclear”, “atom” or “atomic” in the headline or lead
- **Preprocessing:** select lemmas and only nouns, verbs, adjectives and proper noun
- **Question:** Is the change in culture surrounding nuclear issues expressed in a change in word use over time as captured by LDA?

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts using Topic Modelling, *Digital Journalism.*

Example: Topic model

Choosing parameters .

Number of topics: defines into how many topics the LDA model should classify the words in the documents

- K = 10 topics (systematic comparison with different numbers of K)

Alpha: controls the distribution of topics in documents. A higher alpha value means that documents are expected to contain a more diverse mixture of topics, while a lower alpha value means that documents are expected to be more focused on a smaller subset of topics.

Common default value for the alpha is 50 divided by the number of topics

- Alpha = 50/K (as the authors want several clearly distinguishable topics)

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts using Topic Modelling, *Digital Journalism.*

Example: Topic model

TABLE 1
LDA results on US nuclear discourse, 10 topics

Topic	Interpretation	Most representative words
Topics with temporal patterns		
1	*Research*	atomic, Energy, WASHINGTON, scientist, energy, bomb, Commission, United, research, weapon
3	*Cold War*	United, States, Union, Soviet, soviet, weapon, arm, missile, President, treaty
7	*Proliferation*	Iran, United, North, Korea, program, weapon, States, official, country, China
8	*Accidents/ Danger*	plant, power, reactor, Island, Nuclear, accident, Commission, official, waste, safety
Continuous topics		
5	*Weapons*	test, submarine, Japan, first, Navy, year, explosion, missile, ship, bomb
9	*Nuclear Power*	power, plant, company, year, energy, percent, utility, cost, Company, reactor
10	*US Politics*	war, President, weapon, Mr., year, military, policy, world, Reagan, House
Irrelevant topics		
2	*Summaries*	New, new, year, government, official, York, people, business, President, state
4	*Book Reviews*	week, life, book, man, woman, John, year, New, family, University
6	*Films & Music*	Street, West, Theater, Mr., Sunday, East, show, New, tomorrow, p.m.

Validation: Inspecting the top words per topic

The labels of the topics “Research”, “Cold War”, etc. as well as their categorization into topic types such as “Topics with temporal patterns” are found via manual inspection

Example: Topic model

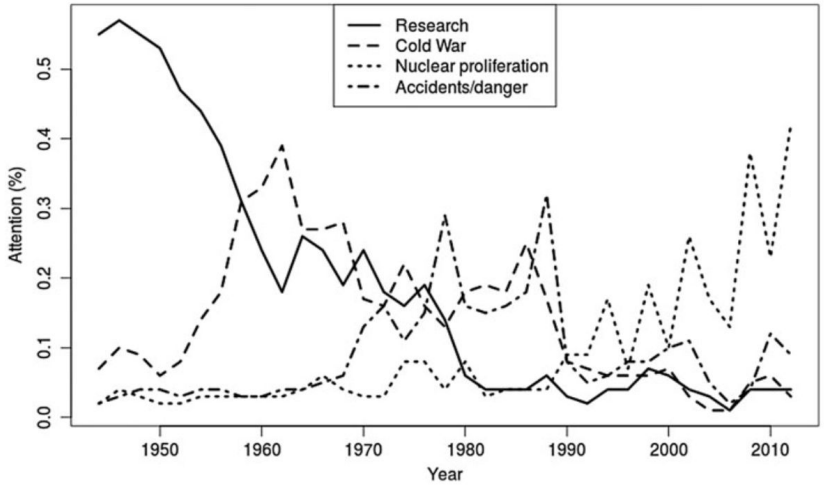
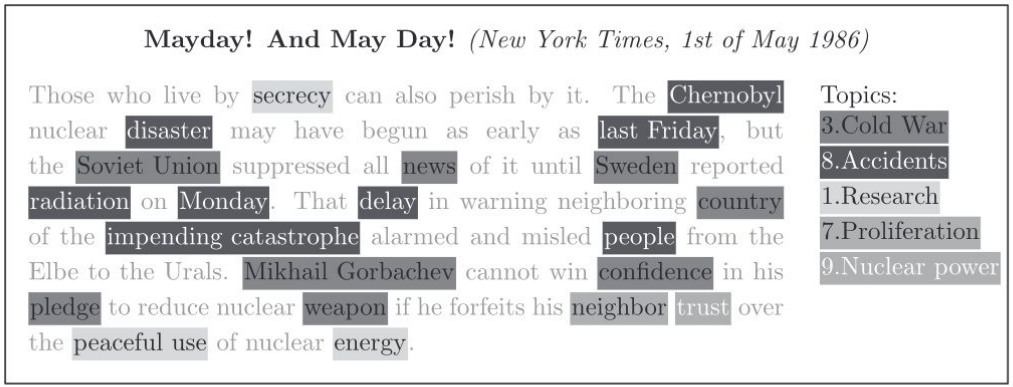


FIGURE 3
Occurrence of topics that have a strong temporal component

Validation: Inspecting the top topics per document and across Comparison with peaks of certain topics with events. E.g. the Chernobyl disaster happened 1986; early 1960s that can be identified as the Cuban missile crisis:
(details see the paper)

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts using Topic Modelling, *Digital Journalism.*

Example: Topic model



Mayday! And May Day! *(New York Times, 1st of May 1986)*

Those who live by secrecy can also perish by it. The Chernobyl nuclear disaster may have begun as early as last Friday, but the Soviet Union suppressed all news of it until Sweden reported radiation on Monday. That delay in warning neighboring country of the impending catastrophe alarmed and misled people from the Elbe to the Urals. Mikhail Gorbachev cannot win confidence in his pledge to reduce nuclear weapon if he forfeits his neighbor trust over the peaceful use of nuclear energy.

Topics:
3.Cold War
8.Accidents
1.Research
7.Proliferation
9.Nuclear power

FIGURE 1
Example article with words from different topics highlighted in the text

Jacobi, C., van Atteveldt, W., and Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts using Topic Modelling, *Digital Journalism.*

Extension of LDA: Structural Topic Models (STM)

We often have information about the document

- Article metadata (year, source)
- Speaker data (year, party, gender)
- Respondent data for open questions

STM allows to use that data

- e.g. topic proportions change over time
- e.g. words differ between speakers, parties

Roberts et al. (2014). Structural Topic Models for Open Ended Survey Responses. AJPS.

Topic modeling with language embeddings

BERTopic (Grootendorst, 2022)

- generates document embedding with pre-trained transformer-based language models, clusters these embeddings
- generates topic representations with the class-based TF-IDF procedure
- transformer-based topic model
- *not* a statistical model (like LDA), but a pipeline of data science techniques

Implementation with Python and Tutorials:

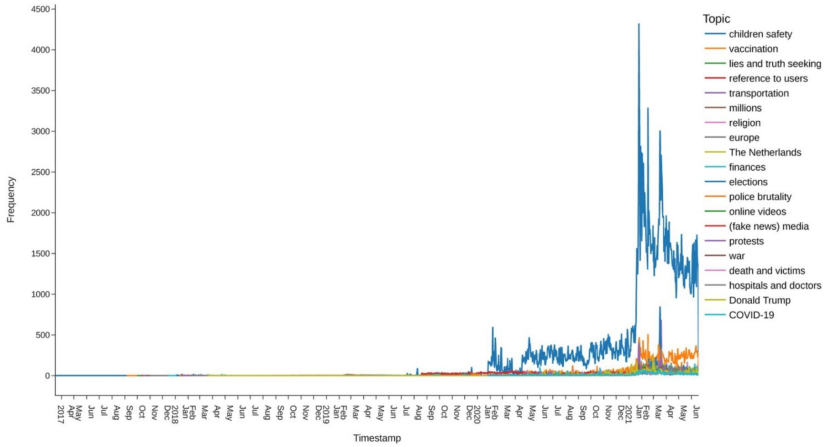
Github: https://github.com/MaartenGr/BERTopic

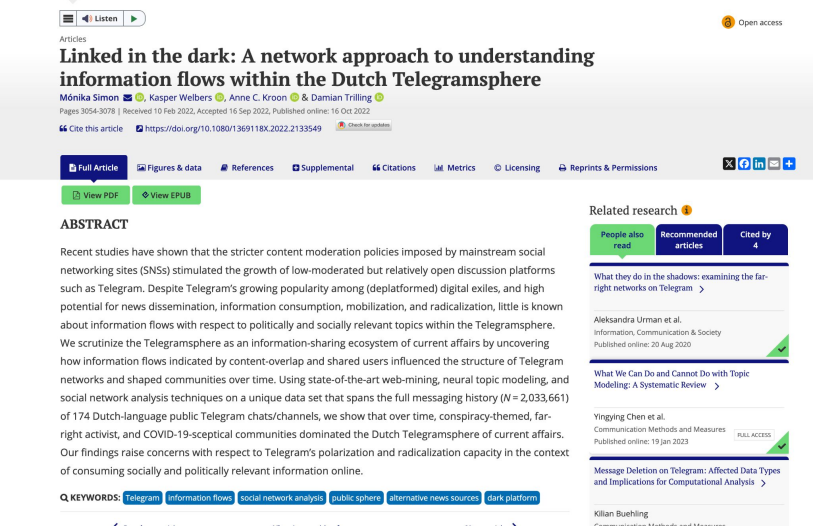
Documentation: https://maartengr.github.io/BERTopic/api/bertopic.html

BERTopic Applications in Social Science

Figure 5 of 6

Figure 5. The top 20 topics overtime (18-03-2017–18-06-2021).





Articles

Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere

Mónika Simon, Kasper Welbers, Anne C. Kroon & Damian Trilling

Pages 3054-3078 | Received 10 Feb 2022, Accepted 16 Sep 2022, Published online: 16 Oct 2022

Cite this article https://doi.org/10.1080/1369118X.2022.2133549

Full Article | Figures & data | References | Supplemental | Citations | Metrics | Licensing | Reprints & Permissions

View PDF | View EPUB

ABSTRACT

Recent studies have shown that the stricter content moderation policies imposed by mainstream social networking sites (SNSs) stimulated the growth of low-moderated but relatively open discussion platforms such as Telegram. Despite Telegram's growing popularity among (deplatformed) digital exiles, and high potential for news dissemination, information consumption, mobilization, and radicalization, little is known about information flows with respect to politically and socially relevant topics within the Telegramsphere. We scrutinize the Telegramsphere as an information-sharing ecosystem of current affairs by uncovering how information flows indicated by content-overlap and shared users influenced the structure of Telegram networks and shaped communities over time. Using state-of-the-art web-mining, neural topic modeling, and social network analysis techniques on a unique data set that spans the full messaging history ($N = 2,033,661$) of 174 Dutch-language public Telegram chats/channels, we show that over time, conspiracy-themed, far-right activist, and COVID-19-sceptical communities dominated the Dutch Telegramsphere of current affairs. Our findings raise concerns with respect to Telegram's polarization and radicalization capacity in the context of consuming socially and politically relevant information online.

KEYWORDS: Telegram | information flows | social network analysis | public sphere | alternative news sources | dark platform

Related research

People also read | Recommended articles | Cited by 4

What they do in the shadows: examining the far-right networks on Telegram

Aleksandra Urman et al.
Information, Communication & Society
Published online: 20 Aug 2020

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

Yingying Chen et al.
Communication Methods and Measures
Published online: 19 Jan 2023

Message Deletion on Telegram: Affected Data Types and Implications for Computational Analysis

Kilian Buehling

Simon, M., Welbers, K. C., Kroon, A., & Trilling, D. (2022). Linked in the dark: A network approach to understanding information flows within the Dutch Telegramsphere. *Information, Communication & Society*, 1–25. https://doi.org/10.1080/1369118X.2022.2133549

Conclusion

- Topic modeling as dimensionality reduction and discovery method
- Topics should make sense (have coherence and meaning)
- Preprocessing, parameters are important!
- LDA interpretable process, (often) plausible results
- Many extensions exist: Of which stm has best R support

Coding Challenge

Have you heard about #oscarssowhite?

What did the discourse look like?

Best movie 2023: Everything Everywhere All at Once

Best movie 2020: Parasite