

Semantic distances between comments in online newspaper forums increase over time

Max Pellert (<https://mpellert.at>)

Barcelona Supercomputing Center



[https://mpellert.at/
ccs_talk_09_25/
ccs_talk_09_25.pdf](https://mpellert.at/ccs_talk_09_25/ccs_talk_09_25.pdf)

DerStandard

DerStandard is a major Austrian newspaper, online user can post below news articles

We created a dataset from the platform that spans ten years (2013–2022)

Includes over 75 million user comments, more than 400 million votes and detailed metadata, also on articles and user interactions

All in a nicely organised data repository:

[https://dataverse.bsc.es/dataset.xhtml?
persistentId=perma:BSC/P32CXW](https://dataverse.bsc.es/dataset.xhtml?persistentId=perma:BSC/P32CXW)

The detailed data set publication: Fraxanet, E., Gómez, V., Kaltenbrunner, A., & Pellert, M. (2025). A Decade of News Forum Interactions: Threaded Conversations, Signed Votes, and Topical Tags (arXiv:2506.22224). arXiv. <https://doi.org/10.48550/arXiv.2506.22224>

Christiane 1:4 vor einer Stunde

38  3

Schon ein Wahnsinn

Jetzt ist der Klimawandel schuld, dass es Sommer

[antworten](#)

► Puderquaste vor 5 Minuten

Echt? Noch immer nicht kapiert????

[antworten](#)

► Mr. Pudel 1:4 vor 28 Minuten

Die Ironie ist, ...

... dass genau die Jahreszeiten eine bestimmte F
den im jährlichen Rhythmus darstellen.

[antworten](#)

- It's like a Wilson +
- subside, permanent, station +
- Sir Pudel +
- Quilombo +
- The Return +
- VELKED +
- + Make Make By +
- Enfield Tennis Academy +
- + Strandford +
- not THE real sciencigay. Not ... +

mehr anzeigen



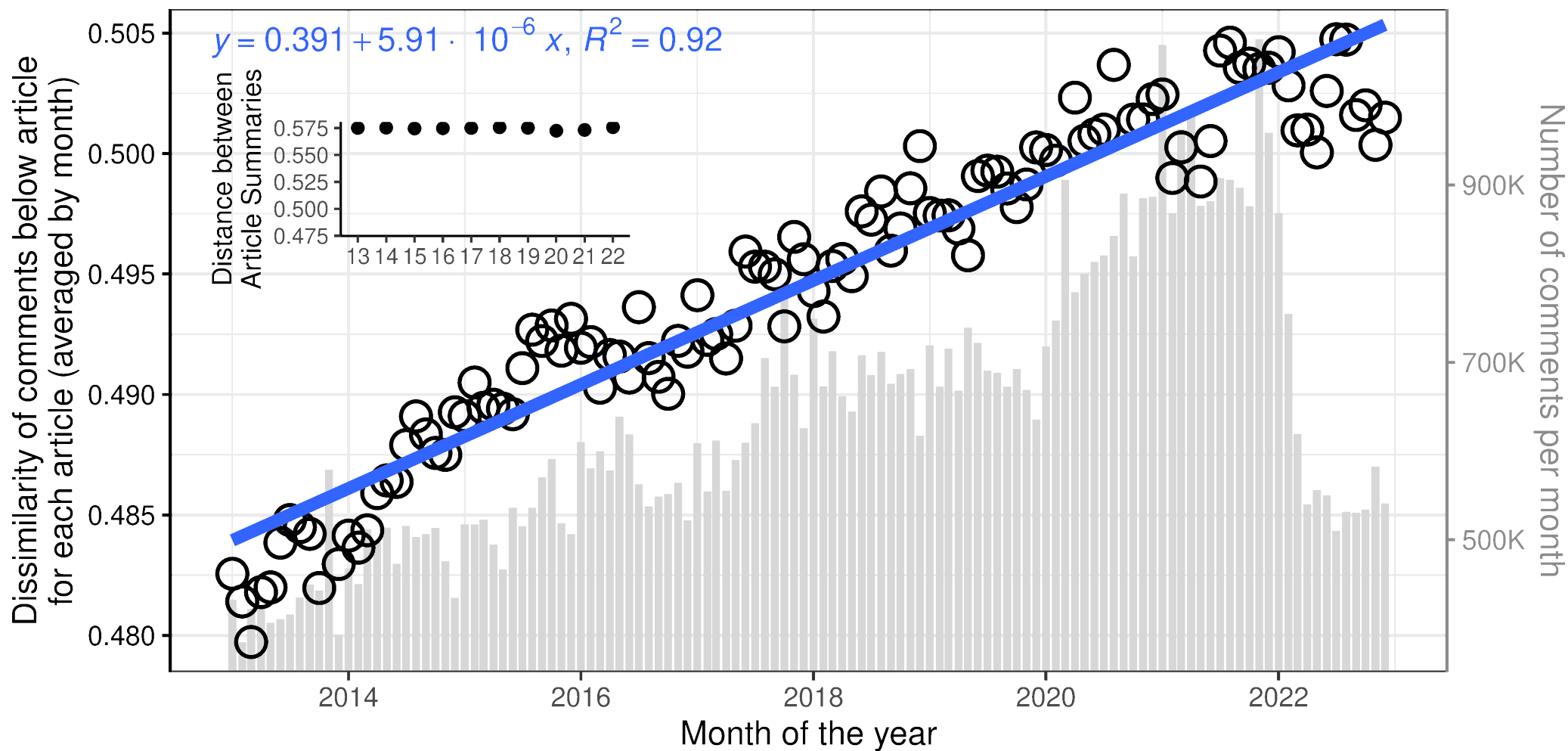
(Dis-)Similarity of postings over time

We embed the text of each posting with “KaLM-embedding-multilingual-mini-v1”

Next, we compute cosine similarity of all pairs of postings below each newspaper article

We average by month of each year

To check for divergence of *contents*, we compute similarity between short text summaries of each pair of article in each year



A proxy for polarization?

Interesting phenomenon that could be an artifact

Like singular topics that overlay everything for some time at least
(COVID-19 or Russo-Ukrainian War)

But dominating news topics hardly a new phenomenon and that should also affect article summary similarity?

Otherwise, it could be a revealing trend

Some sign of polarization?

If yes, the trend should differ between article contents → we can make use of the topical tags on articles provided by the platform

Topic Tag (G)	Topic Tag (E)	Average increase of distance (in 10^{-4})
diestandard	genderissues-political	2.934
diskurs	discourse	1.716
etat	media	1.134
gesundheit	health	3.093
immobilien	real estate	2.496
inland	national (domestic)	1.822
international	international	1.342
karriere	career	2.424
kultur	culture	1.980
lifestyle	lifestyle	0.955
panorama	reports	1.157
sport	sport	1.982
web	web	1.632
wirtschaft	economy	2.072
wissenschaft	science	1.804

Further analyses

I speculate that within-group similarity may have become higher over time, leading to higher dissimilarity between groups and an increasing overall mean

To compare groups, we have (ideological) user factions based on signed networks

Also checking for Simpson's paradox (unlikely)

Definitely, better metrics needed

What are the mechanisms behind that?

(Data Records) A Decade of News Forum Interactions: Threaded Conversations, Signed Votes, and Topical Tags

Version 1.0



Fraxanet Morales, Emma; Pellert, Max; Kaltenbrunner, Andreas; Gómez, Vicenç, 2025, "(Data Records) A Decade of News Forum Interactions: Threaded Conversations, Signed Votes, and Topical Tags", <https://dataverse.bsc.es/BS/SC/P32CXW>, BSC Dataverse, V1, UNF:6:MmzkAl6KMTPYXLdJYALuKw== [fileUNF]

[Cite Dataset](#)

Learn about [Data Citation Standards](#).

[Access Dataset](#)

[Contact Owner](#)

[Share](#)

Dataset Metrics

10 Downloads

Description

This dataset contains the full set of data records described in the "A Decade of News Forum Interactions: Threaded Conversations, Signed Votes, and Topical Tags" publication. It includes the data records for User-level metadata, Comment-level data, Voting behavior, Article metadata, and Pre-computed text embeddings. Additionally, it includes: Annotated subsets related to previous work, Derived Summary metadata, Supplementary files addressing data inconsistencies, Usage notes for reading and decompressing the files, and Basic example code for accessing and processing the data. All data files are provided in compressed TSV format, organized by type and temporal resolution (monthly or yearly). (2025-05-29)

Subject

Computer and Information Science; Social Sciences

Keyword

Discussion Forums, Networks, Text Embeddings, Social Media Platforms, Signed Networks, Discussion Threads, Longitudinal Data

License/Data Use Agreement



[CC BY-NC-SA 4.0](#)

[Files](#)

[Metadata](#)

[Terms](#)

[Versions](#)

Change View

Table

Tree

[Codebook.md](#) (3.1 KB)

[README.md](#) (4.8 KB)

▸ [articles_files](#)

▸ [comments_files](#)

▸ [embeddings](#)

▸ [scripts](#)

▸ [summary_files](#)

▸ [users_files](#)

▸ [votes_files](#)