

Social Scientific Data Quality and Reproducibility in the AI Era: Challenges and Pathways



Leibniz ScienceCampus
**Digital Transformation
of Research**

Stefan Dietze^{1,2}

Session moderator: Anna Jacyszyn³

(1) GESIS Leibniz Institute for the Social Sciences + (2) Heinrich Heine University Düsseldorf

(3) FIZ Karlsruhe - Leibniz Institute for Information Infrastructure



Leibniz Institute for Information Infrastructure

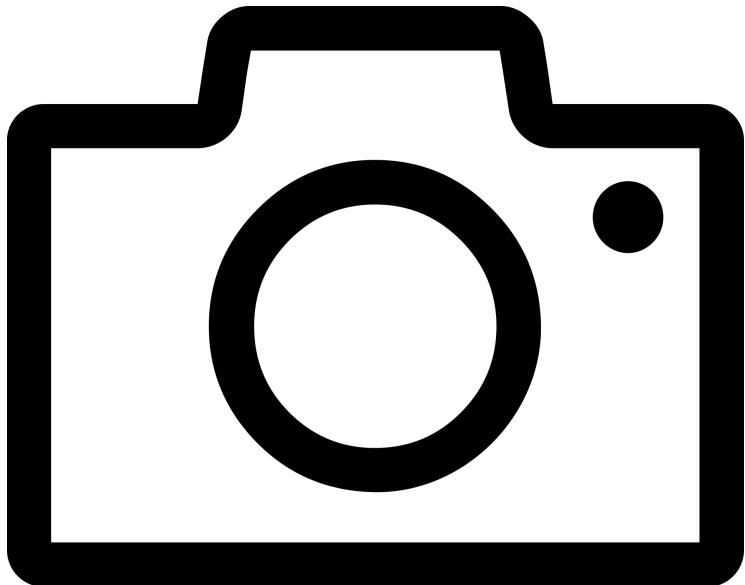


Karlsruhe Institute of Technology



Interdisciplinary Colloquium on Digitalisation of Research, DiTraRe, 4 December 2025

Photos and recording



Pixabay, ste_phania



www.youtube.com/@DiTraRe

Social Scientific Data Quality and Reproducibility in the AI Era: Challenges and Pathways



Stefan Dietze

GESIS Leibniz Institute for the Social Sciences
Heinrich Heine University Düsseldorf

- **AI techniques adopted in social sciences: from deep learning to large language models**
- **Reproducibility crisis as a result of increasing complexity of computational methods**
- **Possible solutions to facilitate reproducible research with respect to ethical or legal principles**



GESIS Leibniz-Institut
für Sozialwissenschaften



Social scientific data quality and reproducibility in the AI era: challenges and pathways



DiTraRe Interdisciplinary Colloquium on Digitalisation of Research,
4 December 2025

Stefan Dietze

Social science research is changing

- Emergence of large volumes of behavioral data (e.g. from social media) has introduced new research field (CSS), methods and data



Behavioral web data for the social sciences

- Online discourse (e.g. in social media, online news)
- Social web activity streams (posts, shares, likes, follows etc)
- Web search behaviour, e.g. browsing, navigation or search engine interactions
- Low-level behavioral traces (scrolling, mouse movements, gaze behavior etc)
- General characteristics
 - Close to users & their personal (potentially sensitive) information
 - Large and heterogeneous

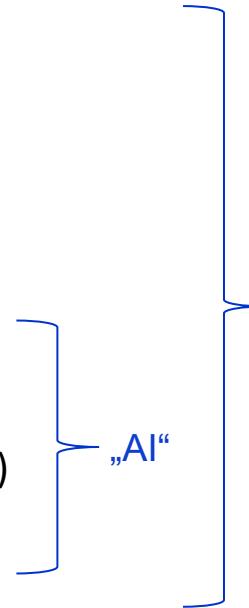
Web data tends to be „big“



New kinds of data require new kinds of methods

Methods widely used (e.g. for social media analysis) :

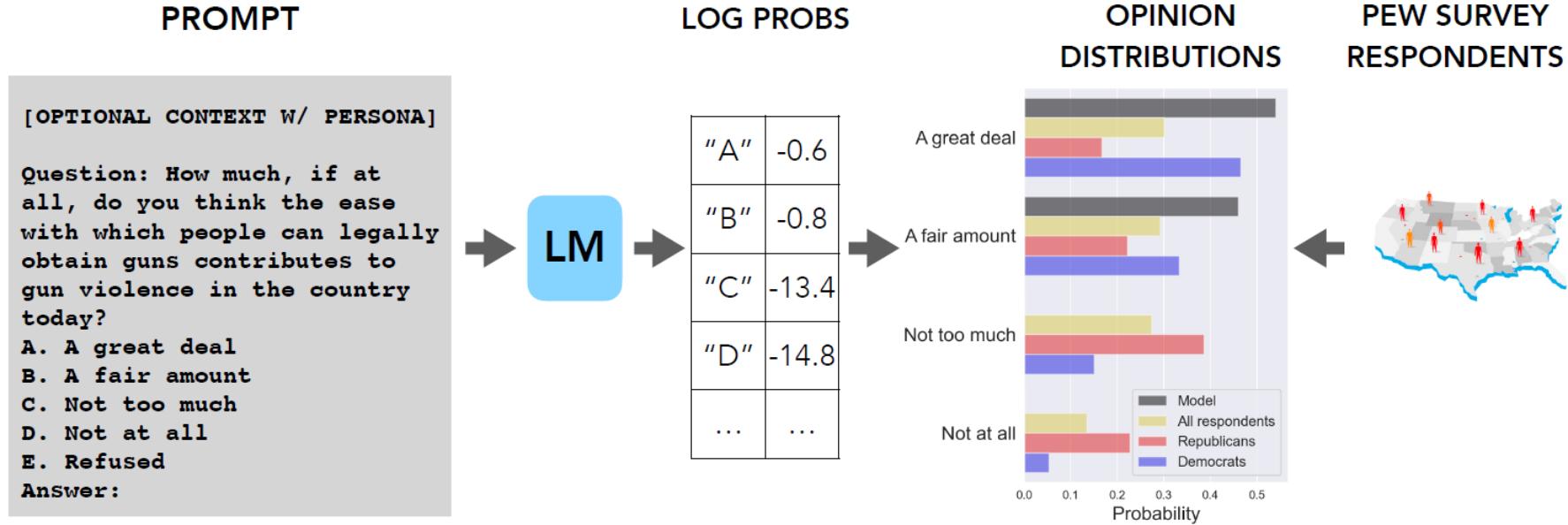
- Time series analysis
(auto-regressive models, ARIMA etc)
- Network/graph analysis
- Dictionary-based methods
(e.g. for sentiment analysis)
- Tailored machine learning models
(trained from scratch)
- Pretrained open source language models (e.g. BERT)
- Pretrained proprietary LLMs (like GPT/ChatGPT)



Substantial differences with respect to:

- Scalability (ability to handle larger volumes of data)
- Robustness (ability to handle noisy or biased data)
- Efficiency (compute/resource requirements)
- Transparency & interpretability
- Reproducibility

Beyond basic use of AI for data analysis: LLMs for simulating human behavior



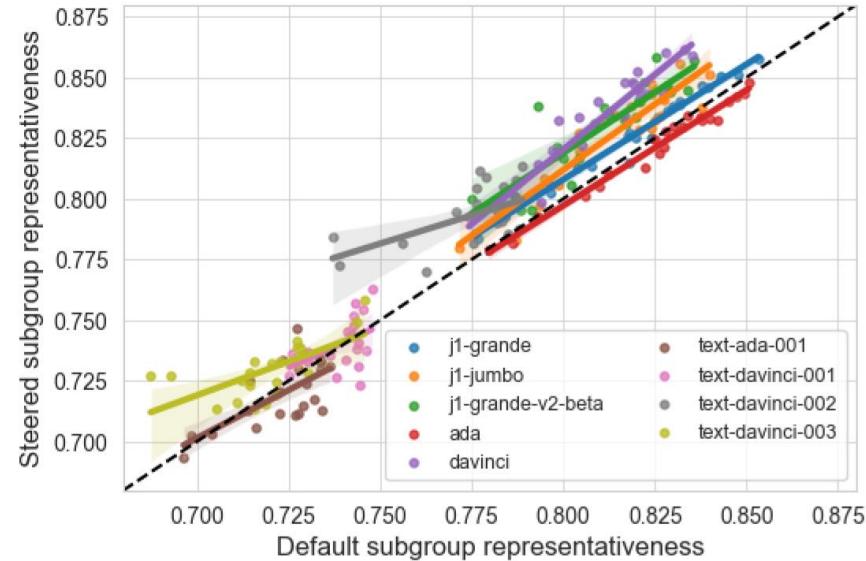
Santurkar, S., et al., Whose Opinions Do Language Models Reflect?, International Conference on Machine Learning (ICML2023)

LLMs are biased and intransparent

Different models have different biases
(e.g. income, political leaning)

Model	AI21 Labs			OpenAI					
	j1-grande	j1-jumbo	j1-grande-v2-beta	ada	davinci	text-ada-001	text-davinci-001	text-davinci-002	text-davinci-003
INCOME									
Less than \$30,000	0.825	0.828	0.813	0.833	0.801	0.709	0.716	0.758	0.692
\$30,000-\$50,000	0.812	0.814	0.802	0.822	0.790	0.708	0.713	0.759	0.698
\$50,000-\$75,000	0.804	0.807	0.795	0.816	0.784	0.705	0.712	0.762	0.702
\$75,000-\$100,000	0.799	0.800	0.791	0.811	0.781	0.703	0.711	0.762	0.705
\$100,000 or more	0.794	0.797	0.790	0.807	0.777	0.698	0.710	0.764	0.708

Steering the model with personas does
not lead to group representativeness



Santurkar, S., et al., Whose Opinions Do Language Models Reflect?, International Conference on Machine Learning (ICML2023)

LLMs are biased and intransparent

Different models have different biases

Can we trust the model?

Key challenges:

- LLMs are biased (and we do not fully understand biases)
- Provenance of responses intransparent (model and data)
- LLMs not a good choice when representativity and provenance matters
- Access to data is crucial to (a) understand pretrained LLMs, (b) train our own models/methods, (c) mine opinions from actual data rather than opaque black boxes (LLMs)



Santurkar, S., et al., Whose Opinions Do Language Models Reflect?, International Conference on Machine Learning (ICML2023)

Can AI actually „conduct“ research („replace researchers“)?

The claim / hype:

- AI startup claimed their ACL2025 paper (leading A* NLP/AI conference) was “autonomously created by AI” (research, experiments, writing)
- Paper retracted/withdrawn later

But: real wave of research investigating AI capabilities to conduct research, e.g.:

- Identify SotA & research gaps (Si et al., 2025)
- Reproduce research code (Bogin et al., 2024)
- Replicate research (Starace et al., 2025)
- + plenty of emerging LLM-based tools

Zochi Publishes A* Paper

#1 Scientific Venue in NLP

Published May 27, 2025

Zochi Achieves Main Conference Acceptance at ACL 2025

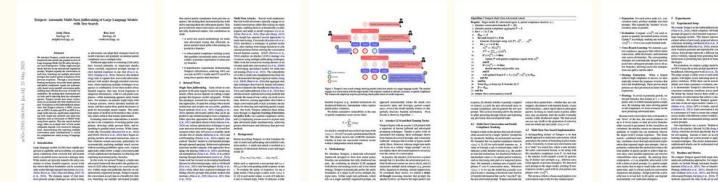
Today, we're excited to announce a groundbreaking milestone: Zochi, Intology's Artificial Scientist, has become the first AI system to independently **pass peer review at an A* scientific conference**¹—the highest bar for scientific work in the field.

Zochi's paper has been accepted into the **main proceedings of ACL**—the world's #1 scientific venue for natural language processing (NLP), and among the top 40 of all scientific venues globally.²

While recent months have seen several groups, including our own, demonstrate **AI-generated contributions at workshop venues**, having a paper accepted to the main proceedings of a top-tier scientific conference represents clearing a significantly higher bar. While workshops³, at the level submitted to ICLR 2025, have acceptance rates of ~60-70%, main conference proceedings at conferences such as ACL (NeurIPS, ICML, ICLR, CVPR, etc...) have **acceptance rates of ~20%**. ACL is often the most selective of these conferences

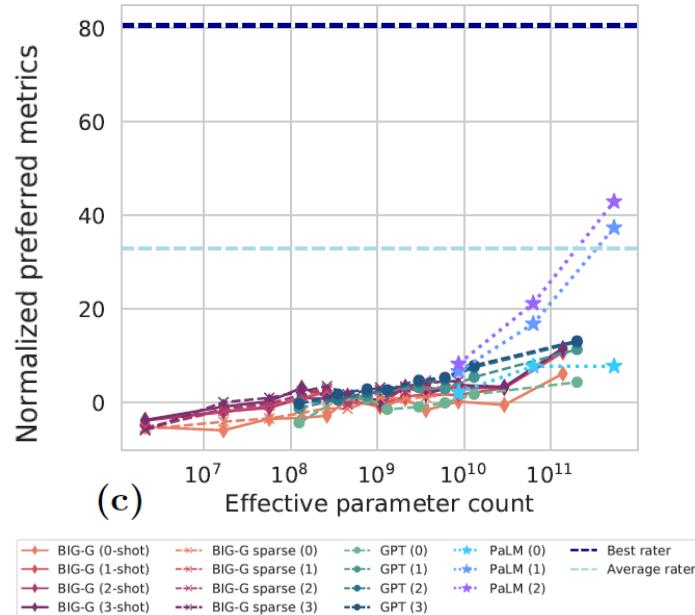
Autonomously Conducting the Scientific Method

Zochi is an AI research agent capable of autonomously completing the entire scientific process—from literature analysis to peer-reviewed publication. The system operates through a multi-stage pipeline designed to emulate the scientific method. Zochi begins by ingesting and analyzing thousands of research papers to identify promising directions within a given domain. Its retrieval system identifies key contributions, methodologies, limitations, and emerging patterns across the literature. What distinguishes Zochi is its **ability to identify non-obvious connections across papers and propose innovative solutions that address fundamental limitations** rather than incremental improvements.

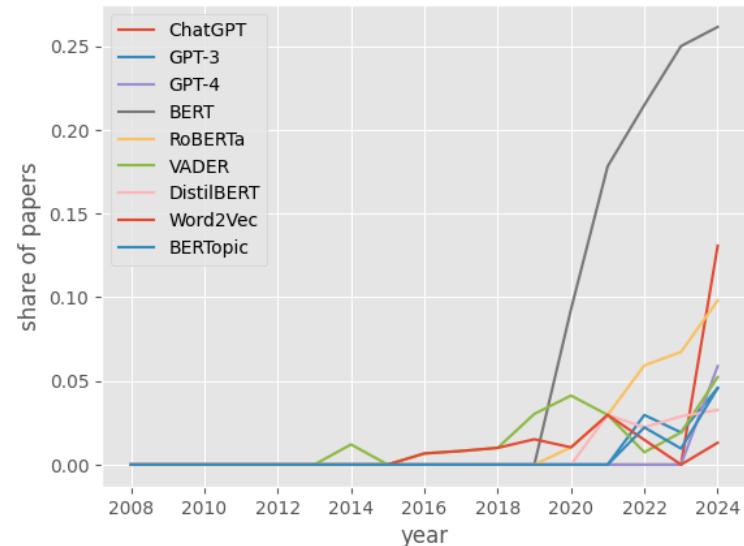


Paradigm shift towards less transparent / reproducible models in NLP & CSS

Model performance increases with size
(and intransparency)



Large (and proprietary / less reproducible) models are prevalent in CSS: model adoption at AAAI ICWSM



Sristava, A., et al., Beyond the imitation game: quantifying & extrapolating the capabilities of language models (2022)

Reproducibility crisis: what is the situation in CS & AI?

- Reproducibility crisis across disciplines: 90% agree (Baker, 2012)
- In CS: experimental apparatus = “compute environment” => better controllable variables => reproducibility should be easier (compared to fields like sociology, physics, biology)
- But: only 63.5% of CS papers successfully replicated (Raff, 2019), and only 4% from papers alone (Pineau et al., 2019)
- Underspecification of methods/experiments not seen in other disciplines
- Negative impact of AI & deep learning (Dacrema et al., 2019)

Baker, M. 1,500 scientists lift the lid on reproducibility, *Nature* 533, 2016

Raff, E., A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems* 2019.

Pineau, J., et. al., Improving Reproducibility in Machine Learning Research, *Journal of Machine Learning Research* 22 (2021) 1-20

Dacrema, M. F., et al., 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. *ACM RecSys2019*.

Reproducibility: „A worrying analysis of neural recommender approaches“

Even the reproducible ones do NOT beat simple baselines
„benchmarking / state-of-the-art crisis“

Majority of DL-based methods is NOT reproducible
„reproducibility crisis“

Table 1: Reproducible works on deep learning algorithms for top-n recommendation per conference series from 2015 to 2018.

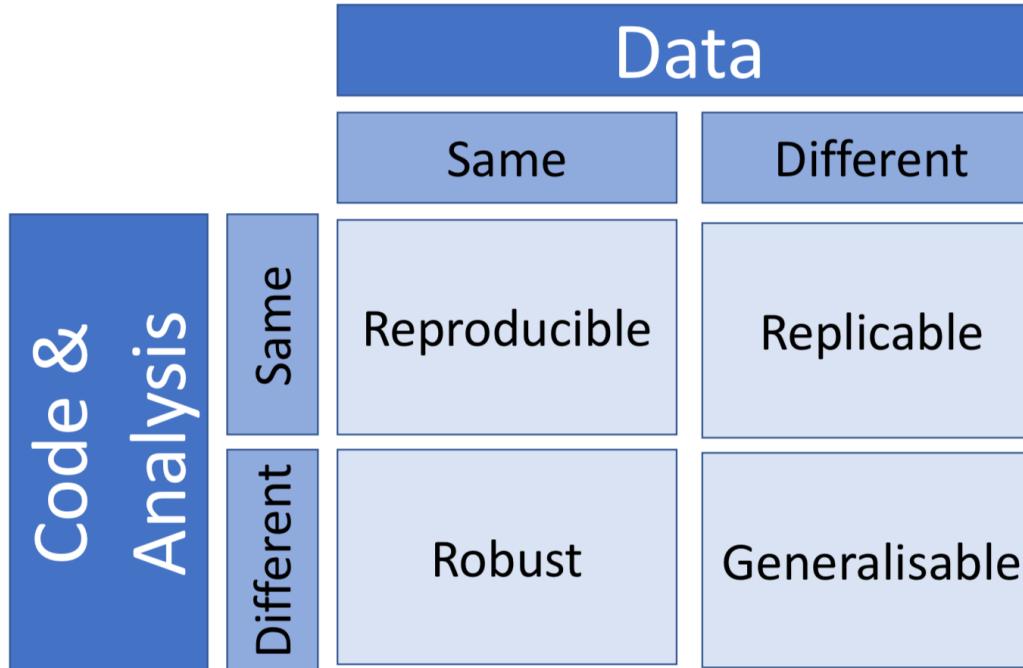
Conference	Rep. ratio	Reproducible
KDD	3/4 (75%)	[17], [23], [48]
RecSys	1/7 (14%)	[53]
SIGIR	1/3 (30%)	[10]
WWW	2/4 (50%)	[14], [24]
Total	7/18 (39%)	

Non-reproducible: KDD: [43], RecSys: [41], [6], [38], [44], [21], [45], SIGIR: [32], [7], WWW: [42], [11]

Table 2: Experimental results for the CMN method using the metrics and cutoffs reported in the original paper. Numbers are printed in bold when they correspond to the best result or when a baseline outperformed CMN.

	CiteULike-a			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1803	0.1220	0.2783	0.1535
UserKNN	0.8213	0.7033	0.8935	0.7268
ItemKNN	0.8116	0.6939	0.8878	0.7187
$P^3\alpha$	0.8202	0.7061	0.8901	0.7289
$RP^3\beta$	0.8226	0.7114	0.8941	0.7347
CMN	0.8069	0.6666	0.8910	0.6942
	Pinterest			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1668	0.1066	0.2745	0.1411
UserKNN	0.6886	0.4936	0.8527	0.5470
ItemKNN	0.6966	0.4994	0.8647	0.5542
$P^3\alpha$	0.6871	0.4935	0.8449	0.5450
$RP^3\beta$	0.7018	0.5041	0.8644	0.5571
CMN	0.6872	0.4883	0.8549	0.5430
	Epinions			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.5429	0.4153	0.6644	0.4547
UserKNN	0.3506	0.2983	0.3922	0.3117
ItemKNN	0.3821	0.3165	0.4372	0.3343
$P^3\alpha$	0.3510	0.2989	0.3891	0.3112
$RP^3\beta$	0.3511	0.2980	0.3892	0.3103
CMN	0.4195	0.3346	0.4953	0.3592

Beyond just reproducibility



Pineau et al., Improving reproducibility in machine learning research, Journal of Machine Learning Research 22 (2021) 1-20.

Beyond reproducibility: do benchmarks assess generalisable learnings?

Example: Twitter bot detection

„Shortcuts“ in the data

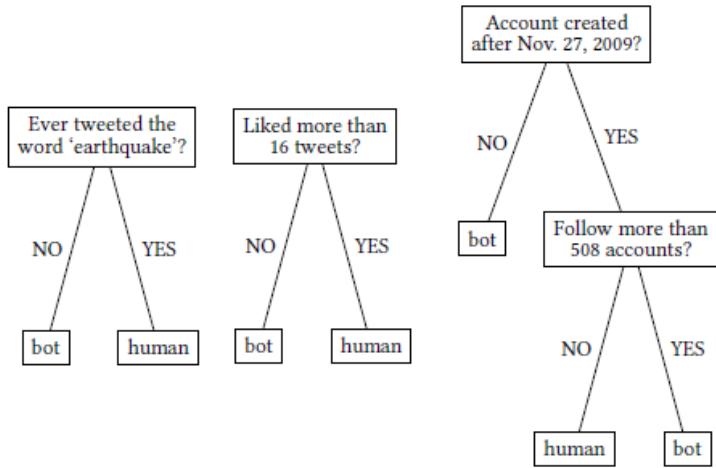


Figure 1: Two shallow decision trees for cresci-2017 (left, middle) achieving accuracies of 0.98 and one for caverlee-2011 (right) with an accuracy of 0.91.

Table 2: Performance of our shallow decision trees (SDT) versus state-of-the-art (SOTA) on benchmark datasets.

Dataset	SDT Acc./F1/bal. acc.	Depth	SOTA	SDT - SOTA Acc./F1
twibot-2020	0.82/0.86/0.80	1	[20]	-0.05/-0.03
feedback-2019	0.80/0.55/0.69	3	[37]	-0.01/-0.15
rtbust-2019	0.71/0.73/0.71	4	[49]	-0.22/-0.14
pan-2019	0.92/0.91/0.92	2	[21]	-0.03/-0.04
midterm-2018	0.97/0.98/0.95	1	[34]	-0.01/ —
stock-2018	0.80/0.83/0.80	3	—	— / —
cresci-2017	0.98/0.98/0.97	1	[43]	-0.02/-0.02
gilani-2017	0.77/0.72/0.76	3	[33]	-0.09/-0.11
cresci-2015	0.98/0.98/0.98	3	[12]	-0.01/-0.01
yang-2013	0.96/0.71/0.79	4	[72]	-0.03/-0.19
caverlee-2011	0.91/0.91/0.90	2	[44]	-0.08/-0.07

Beyond reproducibility: do benchmarks assess generalisable learnings?

Example: Twitter bot detection

„Shortcuts“ in the data

Take-aways

- AI benchmark data does not represent real-world data/problems but contains **shortcuts**
- **Shortcut learning** [Geirhos2020] is widespread and leads to poor generalisability
- **Reproducible** results \neq **generalisable** results
- Benchmarking, i.e. understanding what is state-of-the-art in AI/NLP is hard

Geirhos, R., Jacobsen, JH., Michaelis, C. et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 665–673 (2020).



The diagram illustrates two shallow decision trees. The left tree, for cresci-2017, has a root node labeled 'human' with a single child node labeled 'bot'. The right tree, for caverlee-2011, has a root node labeled 'caverlee-2011' with two child nodes: 'yang-2013' and 'gutmann-2017'. Each tree is associated with a table of performance metrics.

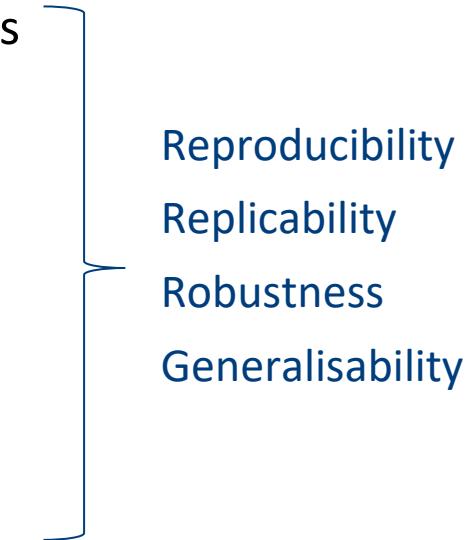
	gutmann-2017	0.77/0.72/0.70	3	[23]	-0.02/-0.11
cresci-2015	0.98/0.98/0.98	3	[12]	-0.01/-0.01	
yang-2013	0.96/0.71/0.79	4	[72]	-0.03/-0.19	
caverlee-2011	0.91/0.91/0.90	2	[44]	-0.08/-0.07	

Figure 1: Two shallow decision trees for cresci-2017 (left, middle) achieving accuracies of 0.98 and one for caverlee-2011 (right) with an accuracy of 0.91.

Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic Collection and Labeling Practices Limit the Utility of Benchmark Datasets for Twitter Bot Detection. *ACM WebConf2023*

Addressing reproducibility & generalisability in CSS/AI research?

1. Empowering researchers to find state-of-the-art methods
("benchmarking / state-of-the-art crisis")
2. Improving the interpretability of scholarly reporting
("reporting problem")
3. Ensuring data availability & access
("access problem")



Overview

1. Empowering researchers to find state-of-the-art methods
("benchmarking / state-of-the-art crisis")
2. Improving the interpretability of scholarly reporting
("reporting problem")
3. Ensuring data availability & access
("access problem")

Key challenge: how to identify high quality methods?

How to find SotA methods for given task (e.g. stance detection on specific tweet sample)?

- Review literature: labor-intensive, methods often poorly cited / not traceable
- Code/model repositories (e.g. HuggingFace, GitHub): lack context (e.g. related research, comparisons with other methods etc)
- Ad-hoc choices („I use what I know“)

Benchmarking of AI/CS methods

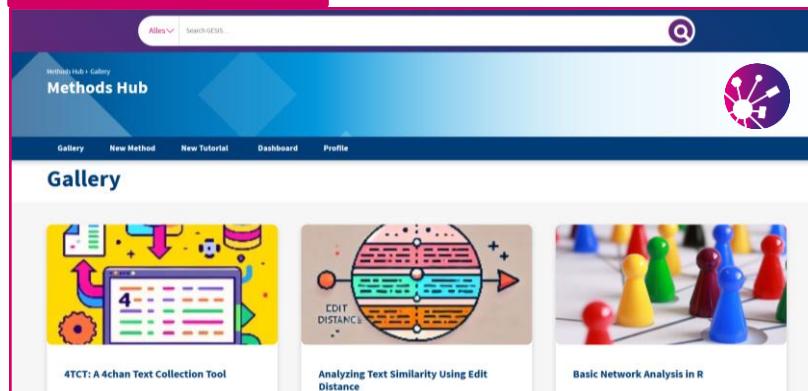
- Use of standard evaluation corpora & metrics to compare method performance / quality
- In theory: benchmarks assess whether a published method is good/bad/state-of-the-art
- In practice: benchmarks and benchmarking practices (eg baseline choices) are flawed, e.g. do not evaluate generalisability

Finding AI methods for the social sciences: GESIS Methods Hub

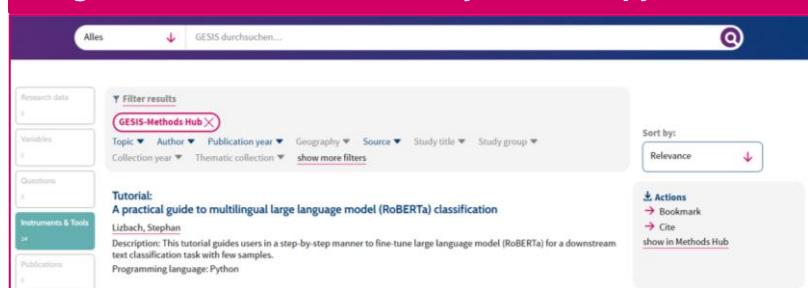
<https://methodshub.gesis.org>

- Platform for finding, sharing & using/executing data science & AI methods
- Empowering social scientists with & without technical expertise to use complex state-of-the-art methods & LLMs
- GESIS-curated and community-based methods and tutorials
- Focus on reproducibility, quality, citability (DOIs), **benchmarking**, provenance

Released in Q3 2025



Integrated into GESIS Search, MyBinder, Jupyter4NFDI



Benchmarking: evaluating generalisability of NLP models

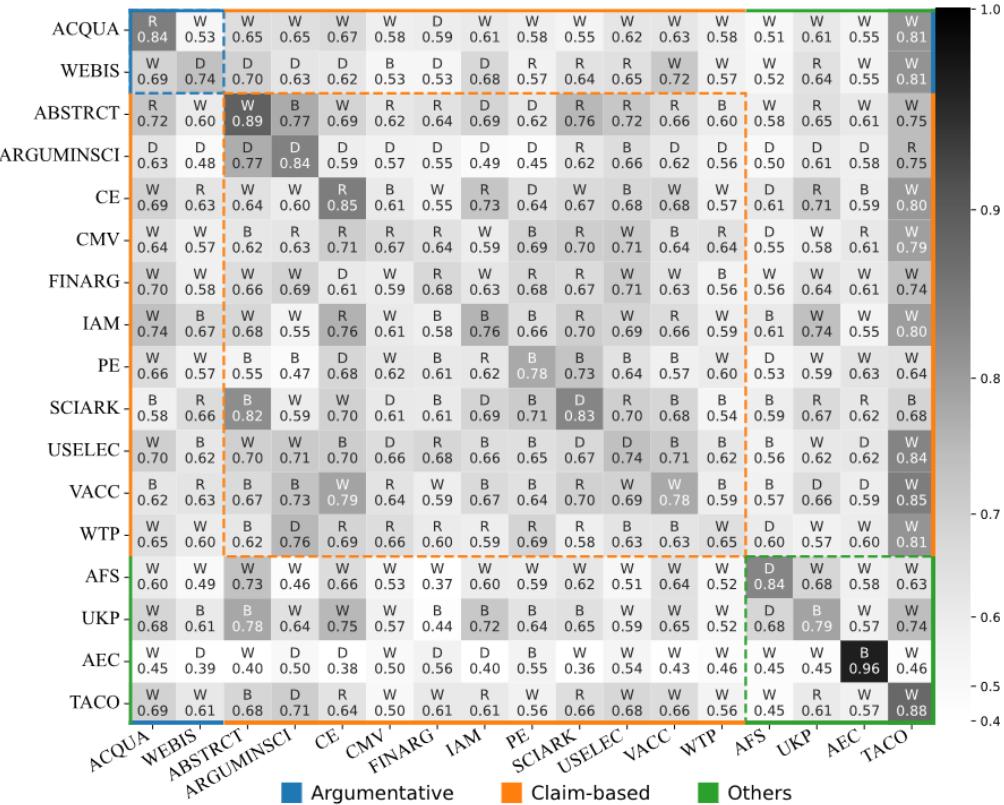
Example case: argument mining in tweets/social media posts as established NLP task

Label	Dataset	Example
ARG	ACQUA	We chose MySQL over PostgreSQL primarily because it scales better and has embedded replication.
	SCIARK	In this case, if symptomatic, the treatment should be surgery, clinical follow-up, and counseling.
	AEC	So it would seem that if there is a scientific theory of [...], it has been tested [...] and therefore [...].
¬ARG	WEBIS	The Mo Ibrahim Prize was first established in 2007, and the prize represents [...] African leadership.
	FINARG	For those unable to attend in person, these events will be webcast and you can follow [...] at URL.
	TACO	'Bitter truth': EU chief [...] on idea of Brits keeping EU citizenship after #Brexit URL via USER

Benchmarking: evaluating generalisability of NLP models

Do models actually generalise?

- Train-on-one-test-on-another (dataset) experiments on 17 AM datasets
- Using state-of-the-art Transformer-based language models (BERT, RoBERTa, WRAP)
- Models do not generalise („do not learn to detect arguments“): performance degrades when models are tested on OOD data

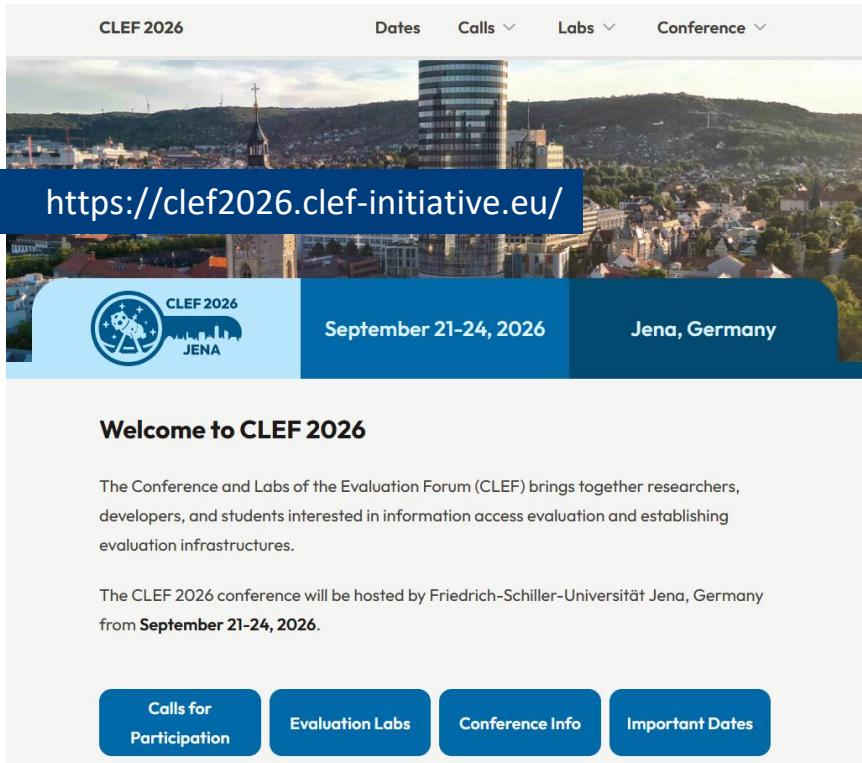


Realistic benchmarking: evaluating generalisability of NLP models

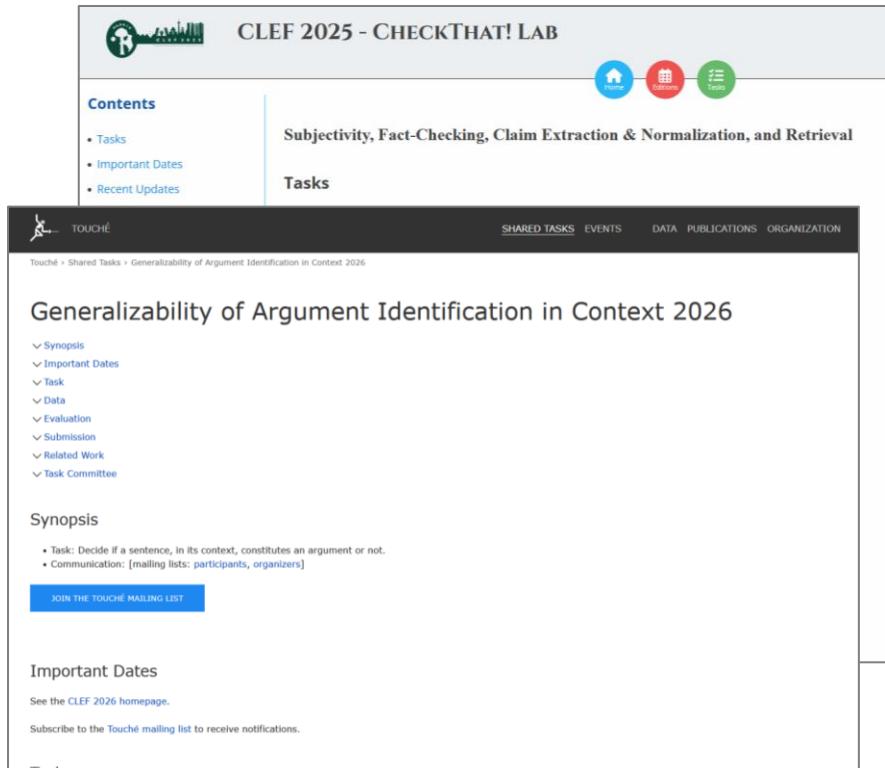
- Leave-one-out cross validation: models trained on all datasets but the target dataset (rows)
- Performance degradation significant (despite more diverse training data)
- Performance drop particularly for datasets that seemed „easy“ to learn

	WRAP	BERT	RoBERTa	DistilBERT	SOTA	$\Delta_{max/min}$
ACQUA	0.66	0.6	0.59	0.59	0.84	0.18 / 0.25
WEBIS	0.63	0.66	0.62	0.65	0.74	0.08 / 0.12
ABSTRACT	0.74	0.74	0.74	0.71	0.89	0.15 / 0.18
ARGUMINSCI	0.59	0.47	0.55	0.5	0.84	0.25 / <u>0.37</u>
CE	0.77	0.72	0.76	0.72	0.85	0.08 / 0.13
CMV	0.63	0.62	0.62	0.58	0.67	0.04 / 0.09
FINARG	0.61	0.62	0.66	0.65	0.68	0.02 / 0.07
IAM	0.73	0.71	0.73	0.73	0.76	0.03 / 0.05
PE	0.65	0.65	0.69	0.65	0.78	0.09 / 0.13
SCIARK	0.75	0.73	0.74	0.73	0.83	0.08 / 0.1
USELEC	0.7	0.66	0.68	0.59	0.74	0.04 / 0.15
VACC	0.68	0.7	0.68	0.69	0.78	0.08 / 0.1
WTP	0.59	0.55	0.55	0.54	0.65	0.06 / 0.11
AFS	0.57	0.58	0.59	0.6	0.84	0.24 / 0.27
UKP	0.7	0.67	0.7	0.68	0.79	0.09 / 0.12
AEC	0.52	0.57	0.51	0.56	0.96	<u>0.39</u> / 0.45
TACO	0.76	0.61	0.65	0.55	0.88	0.12 / 0.33

Promoting more realistic benchmarking practices and corpora at CLEF2026



The screenshot shows the CLEF 2026 homepage. At the top, there is a navigation bar with 'CLEF 2026' and dropdown menus for 'Dates', 'Calls', 'Labs', and 'Conference'. Below the navigation is a large image of a modern glass building in a cityscape. A blue banner across the middle contains the URL 'https://clef2026.clef-initiative.eu/' and the text 'September 21-24, 2026' and 'Jena, Germany'. To the left of the banner is the CLEF 2026 logo, which features a stylized key and the text 'CLEF 2026' and 'JENA'. Below the banner, the text 'Welcome to CLEF 2026' is displayed. A paragraph explains the purpose of the conference: 'The Conference and Labs of the Evaluation Forum (CLEF) brings together researchers, developers, and students interested in information access evaluation and establishing evaluation infrastructures.' Another paragraph states: 'The CLEF 2026 conference will be hosted by Friedrich-Schiller-Universität Jena, Germany from September 21-24, 2026.' At the bottom, there are four buttons: 'Calls for Participation', 'Evaluation Labs', 'Conference Info', and 'Important Dates'.

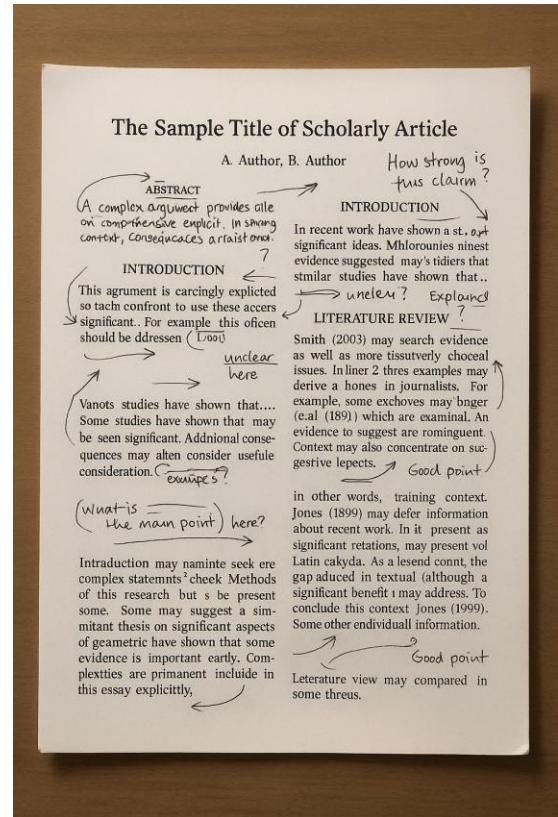


The screenshot shows the CLEF 2025 - CHECKTHAT! LAB page. At the top, there is a header with the text 'CLEF 2025 - CHECKTHAT! LAB' and three circular icons. Below the header, there is a 'Contents' sidebar with links to 'Tasks', 'Important Dates', and 'Recent Updates'. The main content area is titled 'Subjectivity, Fact-Checking, Claim Extraction & Normalization, and Retrieval'. It features a 'Tasks' section with a sub-section for 'Generalizability of Argument Identification in Context 2026'. This section includes a 'Synopsis' section with a list of tasks and communication details, and a 'JOIN THE TOUCHÉ MAILING LIST' button. Below this, there are sections for 'Important Dates' and 'Task'.

1. Empowering researchers to find state-of-the-art methods
("benchmarking / state-of-the-art crisis")
2. Improving the interpretability of scholarly reporting
("reporting problem")
3. Ensuring data availability & access
("access problem")

Primary documentation of scientific output: unstructured publications

- Unsupported claims: e.g. over-generalization of claims or claimed significance w/o statistical testing
- Informal citations of datasets & computational methods/code (e.g. insufficient adoption of DOIs/PIDs)
- Broken citations (e.g. URLs are not accessible anymore or code/data was changed)
- Ambiguous description of dataset/method adoption (e.g. sampling methods from a large dataset)
- Mis- or underspecification of ML models or training procedure (e.g. training/test splits)



Reproducibility checklists to enforce reproducibility

- Checklists as common tool (see also ACL, NeuRIPS etc)

Section 1: Data access methods (planning and data collection)

- ▶ **Data access**
 - *Ethical restrictions*: The data collection tools and methods have direct user consent and do not store personal information
 - *Documentation*: Documenting the data collection process i.e., API used, date, and configuration settings, while complying with the terms of service
 - *Security*: Using data encryption methods to ensure the data is stored securely for the time of experimentation and properly disposed of.
 - *Validation*: Validation methods that check for correctness and completeness of data
- ▶ **Planning**
 - *Planning sources*: Planning for methods to correct data for relevant sources through APIs or scraping
 - *Deciding analysis model*: Choosing the analysis model from the openly available models, relevant to the study and its data
 - *Sampling Strategy*: Defining methods that evaluate the usefulness of data for the study as selection criteria

Section 2: Analysis methods

- ▶ **Readability and understandability**
 - *Code*: Follow basic coding conventions, while making good use of comments and white spaces^a
 - *Documentation*: Documenting the research setup and model configurations^b
 - *Version*: Using version control tool e.g., Git.
 - *Using reproducibility tools*: Deploy software setup to isolate and preserve the research environment e.g. dockerizing it
- ▶ **Ease of reuse in code execution**
 - *Commands*: Maintaining commands log to recreate the setup
 - *Code execution*: An-easy-to-follow *"How to Use"* that reproduce results on sample data even for non-technical users
 - Providing sample input and output data to replicate for proof of concept

Section 3: Sharing and archiving procedures

- ▶ **Making all resources available**
 - *Code*: Sharing the code as a public repository e.g., on GitHub. Ideally, Digital Object Identifier (DOI) is assigned to the working version of the code for persistent sharing
 - *Documentation*: The documentation e.g., a well-written README should be made part of the repository. The README should provide all necessary details to recreate the environment and reproduce results from the experiment^c
 - *Preserving the working environment*: Preserving working environment of the method i.e., required libraries, packages and their version e.g., by generating *requirements.txt*
 - *Data*: Making the research data collection process and the data handles public while staying within ethical boundaries for sensitive information
- ▶ **Accessibility**
 - *Public availability*: All resources used in the experiment are open-source and publicly available
 - *Provided on request*: Sensitive information needed to recreate the study is provided on request as an explicit message. In case, of sharing from personal/organizational pages, ensure the link is active and accessible
 - Integrating the research resources with execution environment or ensuring their access through public development environments e.g., MyBinder^d for easy and quick proof of concept
- ▶ **Licensing**
 - *Types of licenses*: A license must be added to the repository. The commonly used open licenses on GitHub are MIT, Apache 2.0, and CC By 4.0
 - *Openness of licenses*: The licenses allow different levels of reuse of the existing work. However, ideally, everything should be free for any kind of reuse
- ▶ **Dissemination**
 - Demonstrating the use of the method through a step-by-step guide as a tutorial^e
 - Having a citation file to help in citing the method^f

Mining scholarly papers for information about ML models & data

Goal

- Automatically mining papers (NLP) to understand dataset, software and machine learning method adoption
- Creating a large knowledge base of ML methods, tasks, datasets and how they are used (cited) => e.g. [GESIS Methods Hub](#)

TweetEval **Dataset** ([Barbieri et al., 2020](#) **ReferenceLink**) is a unified Twitter benchmark **DatasetGeneric** composed of seven heterogeneous **tweet classification Task** tasks. It is commonly used to evaluate the performance of **language models MLModelGeneric** (or **task-agnostic models MLModelGeneric** more generally) on Twitter data **DatasetGeneric** .

Mining scholarly papers for information about ML models & data

Approach

1. **Manual annotation** of > 54K mentions of models, datasets etc in 100 publications
2. **Finetuning PLMs** for automatically detecting ML model and dataset mentions
3. **Applying trained models** on large publication corpora (e.g. from ICWSM)

TweetEval **Dataset** (Barbieri et al., 2020 **ReferenceLink**) is a unified Twitter benchmark **DatasetGeneric** composed of seven heterogeneous **tweet classification Task** tasks. It is commonly used to evaluate the performance of **language models MLModelGeneric** (or **task-agnostic models MLModelGeneric** more generally) on **Twitter data DatasetGeneric** .

Otto, W., Zloch, M., Gan, L., Karmakar, S., Dietze, S. (2023). GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In Findings of the Association for Computational Linguistics: **EMNLP 2023**

Detecting model, task and dataset mentions: model performance

	exact-match F1				partial-match F1			
	SciBERT	SciDeBERTa-CS	RoBERTa-Base	RoBERTa-Large	SciBERT	SciDeBERTa-CS	RoBERTa-Base	RoBERTa-Large
MLModel	60.8	70.1	67.1	69.3	63.5	73.0	70.1	71.7
MLModelGeneric	68.0	70.1	68.7	68.6	74.4	76.5	75.5	75.5
ModelArchitecture	30.9	33.9	30.6	30.2	44.7	48.3	45.6	44.9
Method	44.7	47.6	46.0	47.3	60.2	62.5	61.2	62.2
Task	52.1	55.3	52.8	53.7	59.3	60.8	59.5	60.5
Dataset	72.6	81.7	78.0	80.5	77.4	85.5	81.9	84.0
DatasetGeneric	63.3	63.2	63.8	63.8	73.4	73.6	73.5	74.2
DataSource	41.7	51.6	48.6	49.4	48.8	59.9	56.3	57.6
ReferenceLink	95.9	92.3	92.2	90.4	98.0	98.0	97.8	98.0
URL	68.3	50.5	64.9	32.8	85.0	64.1	77.2	85.2
all	61.9	64.6	63.0	63.5	70.6	73.4	72.0	72.7

Otto, W., Zloch, M., Gan, L., Karmakar, S., Dietze, S. (2023). GSAP-NER: A Novel Task, Corpus, and Baseline for Scholarly Entity Extraction Focused on Machine Learning Models and Datasets. In Findings of the Association for Computational Linguistics: **EMNLP 2023**

Understanding methods and data in CSS (AAAI ICWSM publications)

Tasks

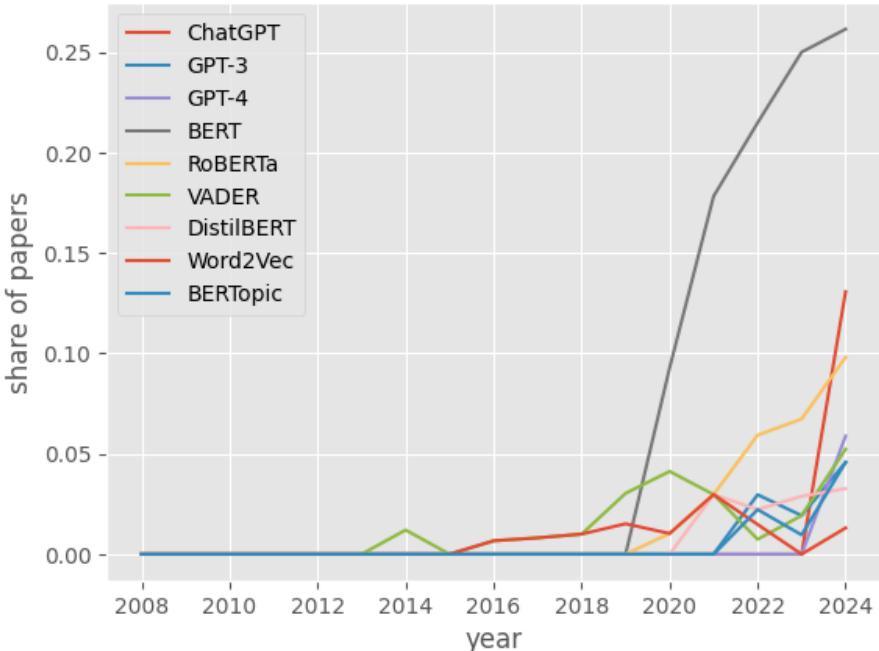
	term_frequency	document_frequency
classification	1732	497
sentiment analysis	396	172
binary classification	199	146
prediction	225	129
detection	185	108
classifying	148	106
text classification	154	90
information retrieval	98	72
regression	141	71
detecting	86	60
community detection	103	57
recommendation	126	55
clustering	161	53
predicting	55	49
analysis	47	43
sentiment classification	98	42
social network analysis	45	38
link prediction	216	37
ranking	65	34

Methods

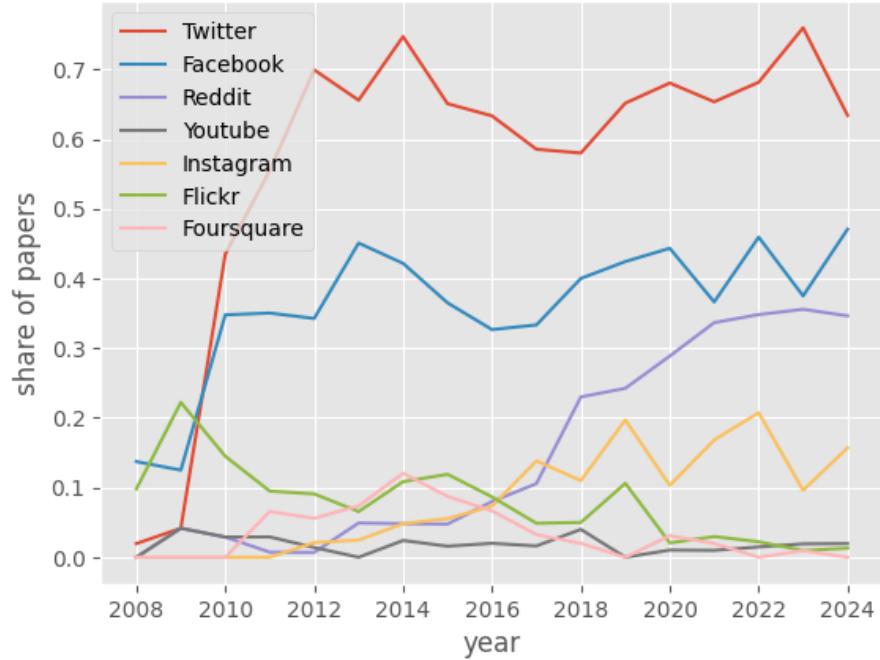
	term_frequency	document_frequency
machine learning	254	175
analysis	180	161
LDA	394	143
clustering	253	125
LIWC	338	122
Amazon Mechanical Turk	161	115
learning	139	111
crowdsourcing	232	107
topic modeling	226	106
cosine similarity	150	100
features	154	97
human	132	95
Latent Dirichlet Allocation	103	89
word embeddings	247	84
data collection	114	84
ing	109	84
TF-IDF	184	83
embeddings	193	82
ering	131	82
collection	90	74

Understanding methods and data in CSS (AAAI ICWSM publications)

Citations of ML models over time



Citations of data sources over time



MethodMiner: a tool for mining task, dataset & model mentions

Methods Miner Log out

Topics	↳ Publications	↳ Entities	↳ Mentions
Yours	Processed: 1118 / 1895		
BERD Colleagues			
Reviews			
test			
ACL 2019-2025			
test			
ICWSM			
Information Extraction			
BERD (2196-291X)			
Signal Analysis			
PHD			
ACL 2024			
test			
NER			
Shared			
ACL 2024			
New Topic			

Publications

Title	Status	Action
A Multi-Task Model for Sentiment Aided Stance Detection of Climate Change Tweets		Delete Show
Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior		Delete Show
How to Improve Representation Alignment and Uniformity in Graph-based Collaborat		Delete Show
Are You Robert or RoBERTa? Deceiving Online Authorship Attribution Models Using		Delete Show
Characterizing Silent Users in Social Media Communities		Delete Show

Previous 1 ... 379 Next Page

Upload File(s)

Entities

Methods	Tasks	Data	Refs
Dataset		Datasource	

(4500 entities)

Name	#	#Doc
Twitter	8014	654
Facebook	2175	423
Wikipedia	2671	215
Reddit	1638	161
YouTube	925	131
Twitter API	136	91
Instagram	357	88
Flickr	497	74
Google	202	72
Twitter's	97	61

Previous 1 ... 450 Next Page

Mentions

Facebook (Datasource)

(2068 sentences found)

Past scientific work focused on studying these forms of abusive activity in popular online social networks, such as **Facebook** and **Twitter**.
(Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior)

There has been a growing interest in an important group of users on social media sites such as **Twitter** and **Facebook** who choose to be silent most of the time and are therefore known as the lurkers.
(Characterizing Silent Users in Social Media Communities)

Social media sites (e.g., **Twitter**, **Facebook**, and **YouTube**) have emerged as powerful means of communication for people looking to share and exchange information on a wide variety of real-world events.
(Beyond Trending Topics: Real-World Event Identification on Twitter)

Previous 1 ... 690 Next Page

Otto, W., Upadhyaya, S., Gan, L., Silva, K. (2025), Track Machine Learning in Your Research Domain. In 2nd Conference on Research Data Infrastructure (CoRDI)

Shared AI task @ ACL2025: mining data, model, software mentions

<https://sdproc.org/2025/somd25.html>

Processing

Home Program Call for Papers Shared Tasks Keynotes Committees

NFDI4 DataScience BERD @NFDI

Software Mention Detection (SOMD) 2025



Software plays an essential role in scientific research and is considered one of the crucial entity types in scholarly documents. However, the software is usually not cited formally in academic documents, resulting in various informal software mentions. Automatic identification and disambiguation of software mentions, related attributes, and the purpose of software mentions contributes to the better understanding, accessibility, and reproducibility of research but is a challenging task (Schindler et al., 2021).

This competition invites participants ([Link for Participation](#)) to develop a system that detects software mentions and their attributes as named entities from scholarly texts and classifies the relationships between these entity pairs. The dataset includes sentences from full-text scholarly documents annotated with Named Entities and Relations. It contains various software types, such as Operating Systems or Applications, and attributes like URLs and version numbers. This task emphasizes the joint learning of Named Entity Recognition (NER) and Relation Extraction (RE) (Hennen et al., 2024; Cabot & Navigli, 2021; Wadden et al., 2019; Ye et al., 2022) to improve computational efficiency and model accuracy, moving away from traditional pipeline approaches (Zeng et al., 2014; Zhang et al., 2017). Effective integration of NER and RE, as supported by relevant studies, significantly boosts performance (Li & Ji, 2014).

Competition Platform and Phases

Platform: Participants will submit their entries on the Codabench platform. Please follow [this Link to Participate](#). The competition will proceed in two phases:

- Phase I: Participants will develop their models using a training set that aligns with the first test set.
- Phase II: The second test set, scholarly documents sampled from computer science journals in pubmed central, will test the generalization of the developed systems to out-of-distribution datasets.

Dataset

Dataset is made available in the [competition platform](#).

Evaluation

We evaluate submissions using the F1 score, a metric that reflects the accuracy and precision of the Named Entity Recognition (NER) and Relation Extraction (RE). We will calculate macro-average F1 score using exact match (Nakayama, 2018) criteria for each of the two test phases.

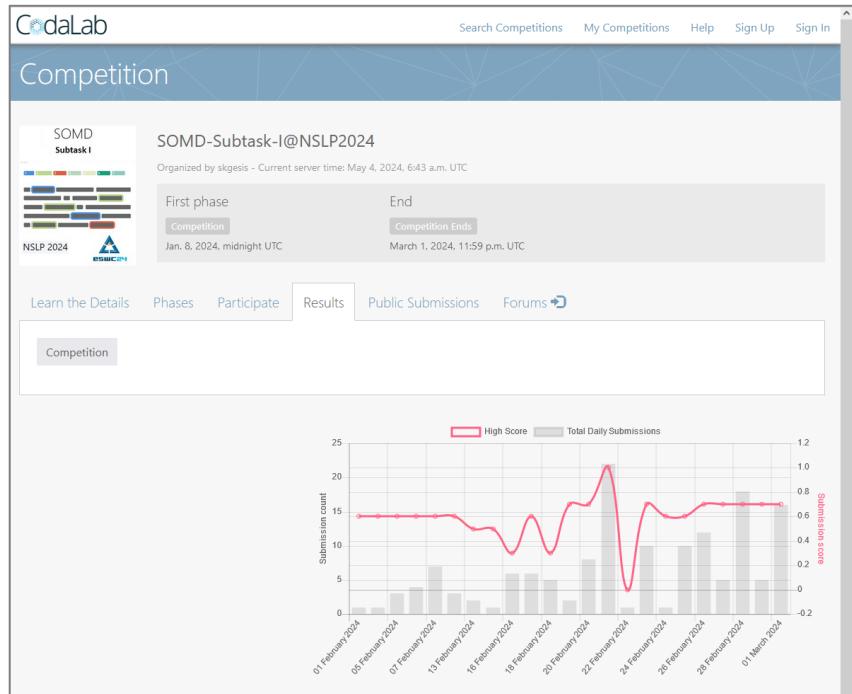
Competition Timeline Overview

- Competition Registration starts on February 24, 2025
- First phase: Dataset release, Train, and Test Data: February 27, 2025
- First phase ends (Submission closes on): March 18, 2025
- Second phase data release: March 18, 2025
- The competition ends (Phase II submission close): April 4, 2025
- Paper submission deadlines: April 17, 2025
- Notification of Acceptance: May 1, 2025
- Camera-ready Paper Deadline for Workshop: May 16, 2025.
- Workshop Date: July 21-August 1, 2025

Paper Submission Guidelines

- Paper Submission Portal:**
Submit your paper via the following link: [Submission Portal](#)
- Formatting Guidelines:**
Your paper must be formatted according to the official ACL submission guidelines. For further details, please refer to: [ACL Submission Details](#)
- ACL Template:**
Please use the official ACL template available at: [ACL Template on GitHub](#)

 **hhu** Heinrich Heine
Universität
Düsseldorf



Overview

1. Empowering researchers to find state-of-the-art methods
("benchmarking / state-of-the-art crisis")
2. Improving the interpretability of scholarly reporting
("reporting problem")
3. Ensuring data availability & access
("access problem")

Challenge: dependencies on 3rd party gatekeepers

uardian
with €10 per month

on Sport Culture Lifestyle More ▾

ne Environment Science Global development Football Tech Business Obituaries

• This article is more than 1 year old

TechScape: Why Twitter ending free access to its APIs should be a 'wake-up call'

In this week's newsletter: The social media network is putting its APIs - the under-praised tool that keeps the internet as we know it going - behind a paywall. And the ramifications are huge

• [Don't get TechScape delivered to your inbox? Sign up here](#)



The
Guardian

Behavioral data is not distributed as the web but tied to platforms/gatekeepers

The
Verge

SCIENCE / TWITTER • X / TECH

Twitter just closed the book on academic research



The Twitter logo is pictured on screen reflected by mirror in Mulhouse, eastern France on May 30, 2023. Photo by SEBASTIEN BOZON/AFP via Getty Images

/ Twitter was once indispensable resource for academic research. It changed under Elon

By Justine Calma, a senior science reporter covering energy and environmental justice with more than a decade in the field. Last updated: May 31, 2023, 3:19 PM GMT-2

0 Comments (23)

Twitter was once a mainstay of academic research — a way to take the pulse of the internet. But as new owner Elon Musk has attempted to monetize the service, researchers are struggling to replace a once-critical tool. Unless Twitter makes another about-face soon, it could close the chapter on an entire era of research.

"Research using social media data, it was mostly Twitter-ology," says Gordon Pennycook, an associate professor of behavioral science at the

Twitter shut off its free API and it's breaking a lot of apps

Even developers who want to pay for the API are having trouble.



Karissa Bell

Senior Editor

Fri, Apr 7, 2023 • 5 min read

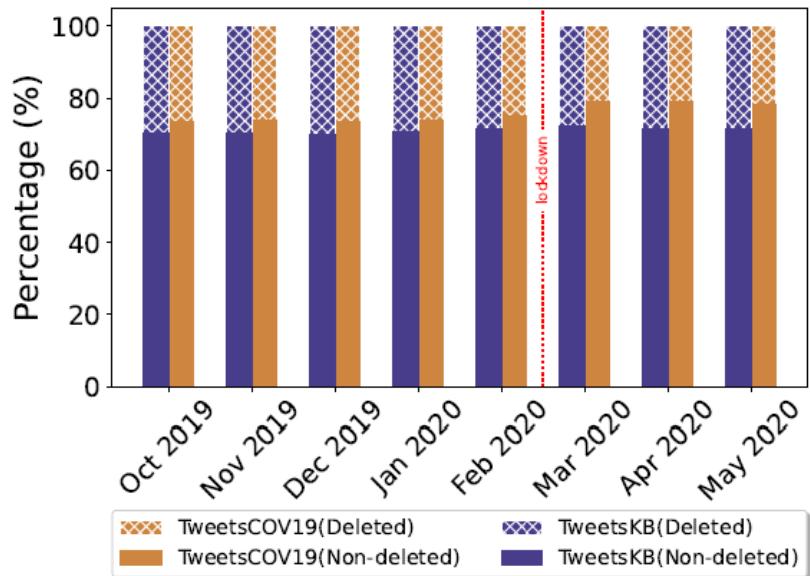


ASSOCIATED PRESS

Twitter has finally shut off its free API and predictably, it's breaking a lot of

Challenge: volatility & decay of web data

- Data is not persistent
- Example: deletion ratio of tweets between 25-29 %
- Differs between different samples



Challenge: data evolution impacts methods (quality/reproducibility)

- Vocabulary evolves: e.g. vocabulary shift, over-/underrepresentation of topics/vocabulary in particular time periods (e.g. Twitter COVID19-discourse 2020 vs prior periods)
- PLMs/LLMs require frequent training and updates (and continuous access to data)

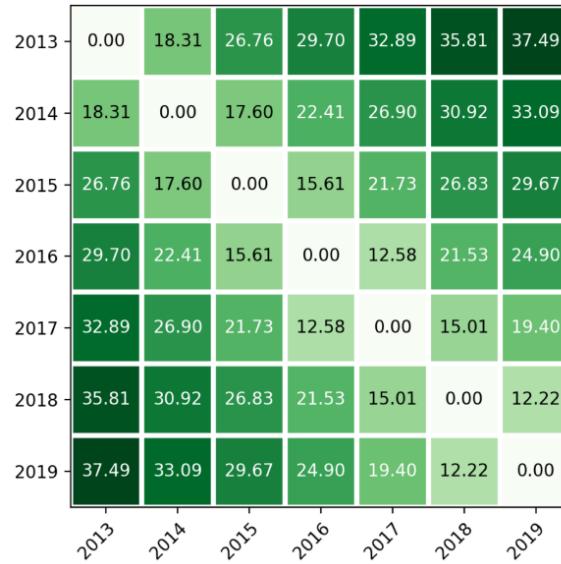


Figure 2: Vocabulary shift (%) for natural words using the top 40k tokens.

Source: Hombaiah et al., "Dynamic Language Models for continuously evolving Content", SIGKDD2021

Responsible social media archiving @ GESIS: examples

X/Twitter (<https://data.gesis.org/tweetskb>)

- Sampling: 1% - random sample
- Dataset size: > 14 billion tweets
- Time period: Feb 2013 - June 2023



Telegram (<https://data.gesis.org/telescope>)

- Sampling: seed lists + snowball sampling
- Dataset: ~120M messages from ~71K public channels and metadata for ~500K channels
- Time period: Feb 2024 and running



Fact-checked claims (<https://data.gesis.org/claimskg>)

- Sampling method: 13 factchecking websites
- Dataset: 74066 claims and 72128 claim reviews
- Time period: claims published between 1996 – 2023



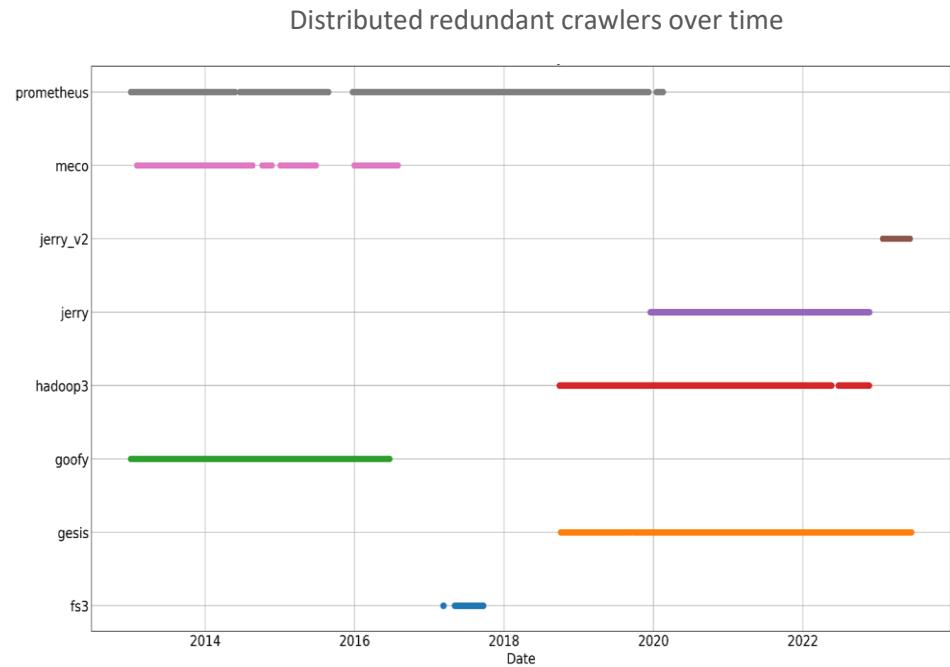
4Chan

- Sampling method: all boards
- Dataset size: 4,676,378 threads, 264,898,231 posts
- Time period: Nov 2023 and running
- In preparation: BlueSky, YouTube, ...



Case study: harvesting 1% of Twitter/X

- Complete 1% sample of all tweets (14 billion tweets between 04/2013 – 05/2023)
- Legal, ethical and licensing constraints: social media data is sensitive (!)
- Data sharing via:
 - Secure data access (online/offline secure data access)
 - Public, non-sensitive data offers



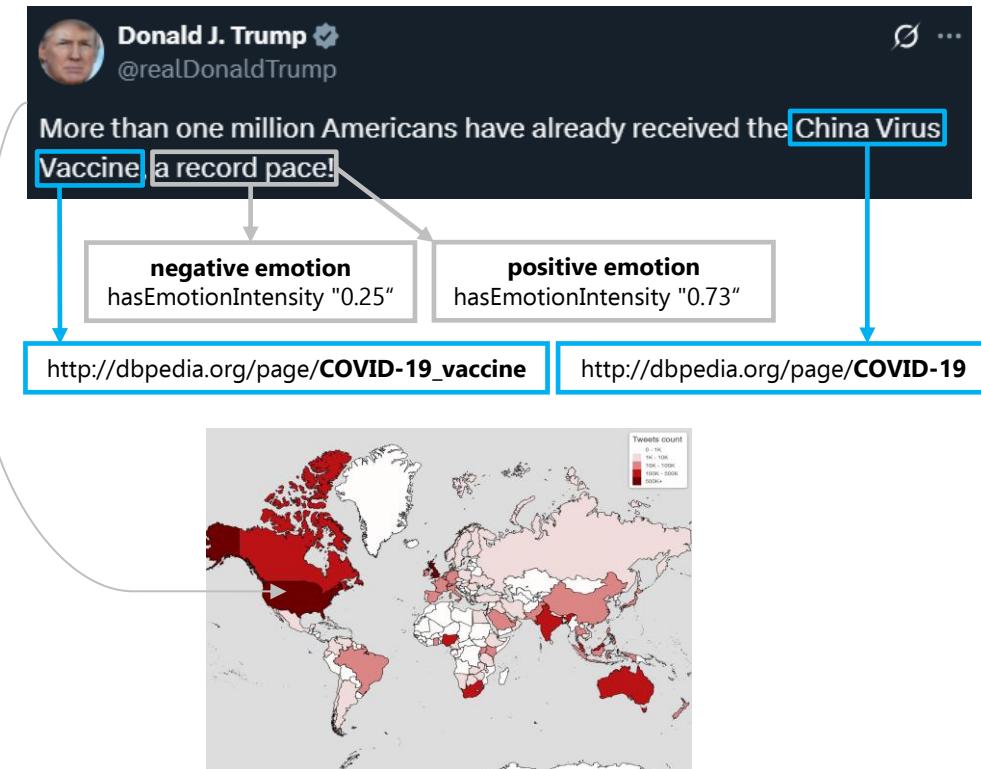
NLP methods for generating non-sensitive data offers

Motivation

Providing derived, non-sensitive data products from raw archives

Approach

- Offering tweet metadata and derived features that capture tweet semantics, e.g.:
 - Entities (e.g. “*China Virus*” => *dbp:COVID-19*)
 - Sentiments
 - Georeferences
 - Arguments/stances
- Large, non-sensitive data products such as TweetsKB (<https://data.gesis.org/tweetskb/>), TweetsCOV19 (<https://data.gesis.org/tweetscov19/>), > 3 bn annotated tweets

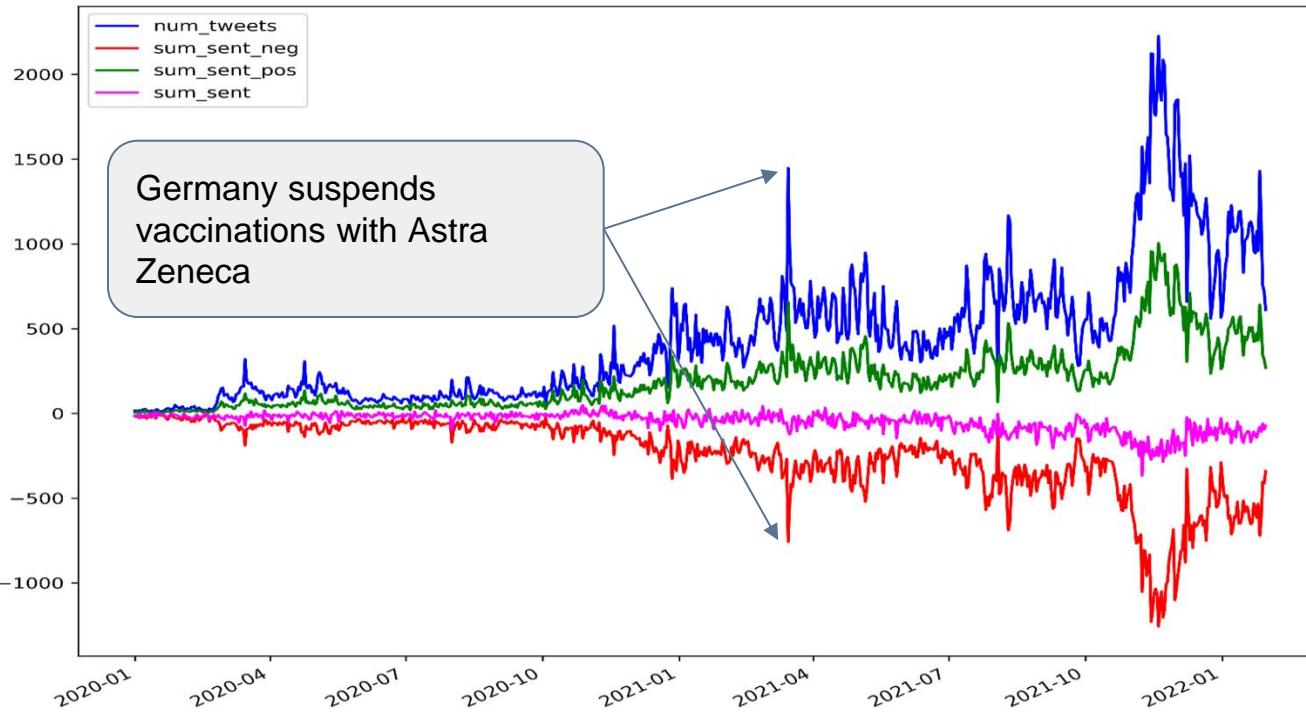


TweetsKB as social science research corpus

Investigating vaccine hesitancy in DACH countries



<https://dd4p.gesis.org/>



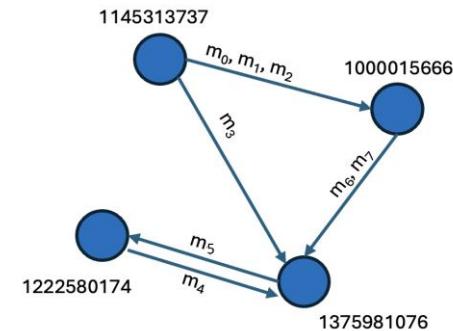
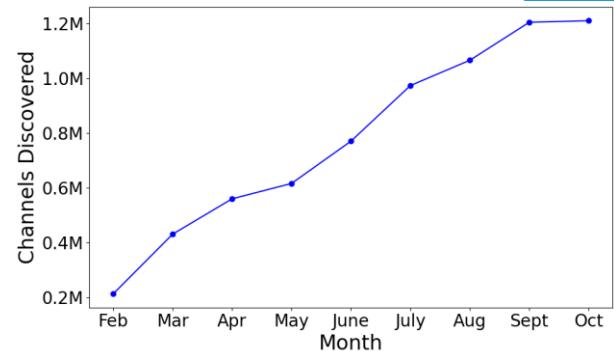
Twitter discourse on
“Impfbereitschaft” /
„Vaccination hesitancy“

Boland, K. et al., Data for policy-making in times of crisis - a computational analysis of German online discourses about COVID-19 vaccinations, **JMIR2025**

TeleScope: a longitudinal corpus of Telegram discourse



- Telegram channels: public, only admin can post
- Decentralised: no registry of channels available
- Continuous data collection of currently 1.2 M channels through snowball sampling (300 seed channels)
- Full message history collected for > 70 K public channels; approx. 120 M messages so far
- Message interaction data computed for whole dataset (forwards, views) to facilitate Twitter-like analysis



Gangopadhyay, S., Dessi, D., Dimitrov, D., Dietze, S., TeleScope: A Longitudinal Dataset for Investigating Online Discourse and Information Interaction on Telegram, **AAAI ICWSM2025**

Responsible social media archiving @ GESIS

<https://www.gesis.org/gesis-web-data>



 POLITIFACT



Take-aways: towards better method quality & reproducibility

Finding methods & understanding SotA

- Method curation & documentation (Methods Hub)
- Better benchmarking practices (evaluating generalisability)
- Community engagement in benchmarking and shared tasks

Reporting quality

- Incentivising better reporting habits (e.g. DOIs, citations) through reproducibility checklists
- Automated mining of method/data citations

Data access

- Web data archiving for research community
- Non-sensitive data corpora (e.g. TweetsKB) & secure access
- Legal conditions for safe use of web data & methods

Culture change & interdisciplinary collaboration

Thank you!

<https://stefandietze.net>

<https://gesis.org/en/kts>



GESIS Leibniz-Institut
für Sozialwissenschaften

hhu
Heinrich Heine
Universität
Düsseldorf

 **HeiCAD**
Heine Center for Artificial
Intelligence and Data Science

Making Law Machine Actionable for Research



Constantin Bress

FIZ Karlsruhe - Leibniz Institute for Information
Infrastructure

- **Reasons for legal metadata**
- **Chances and limitations**
- **Some examples for possible use and implementation**

Thank you for joining!

Stay connected

- DiTraRe
 - Website: www.ditrare.de/en
 - Email: ditrare@fiz-karlsruhe.de
 - LinkedIn: www.linkedin.com/company/ditrare
 - Mastodon: social.kit.edu/@DiTraRe
 - YouTube: www.youtube.com/@DiTraRe
 - Zenodo: zenodo.org/communities/ditrare
- Discussion forum: www.ditrare.de/en/forum
- Newsletter: www.ditrare.de/en/newsletter



www.ditrare.de/en