

Building Large-Scale Instagram Datasets for Political Communication Research: Lessons from the 2025 German Federal Election Campaign

Jakob Berg, M.A.
Professorship for Political Science Research Methods
Institute of Political Science
Faculty of Philosophy, Art, History and Social Sciences



Universität Regensburg

Content

1. Scraping
2. Add-On: Optical Character Recognition & Captioning
3. Legal Challenges: Data Protection & Copyright
4. Storing
5. Dataset Overview & Output
6. Problems
7. Evaluation
8. Study Results
9. Outlook

Universität Regensburg

Scraping: Identification



Die Bundeswahlleiterin

- 29 Parties
- 4.506 Candidates
- 24 persons called „Müller“

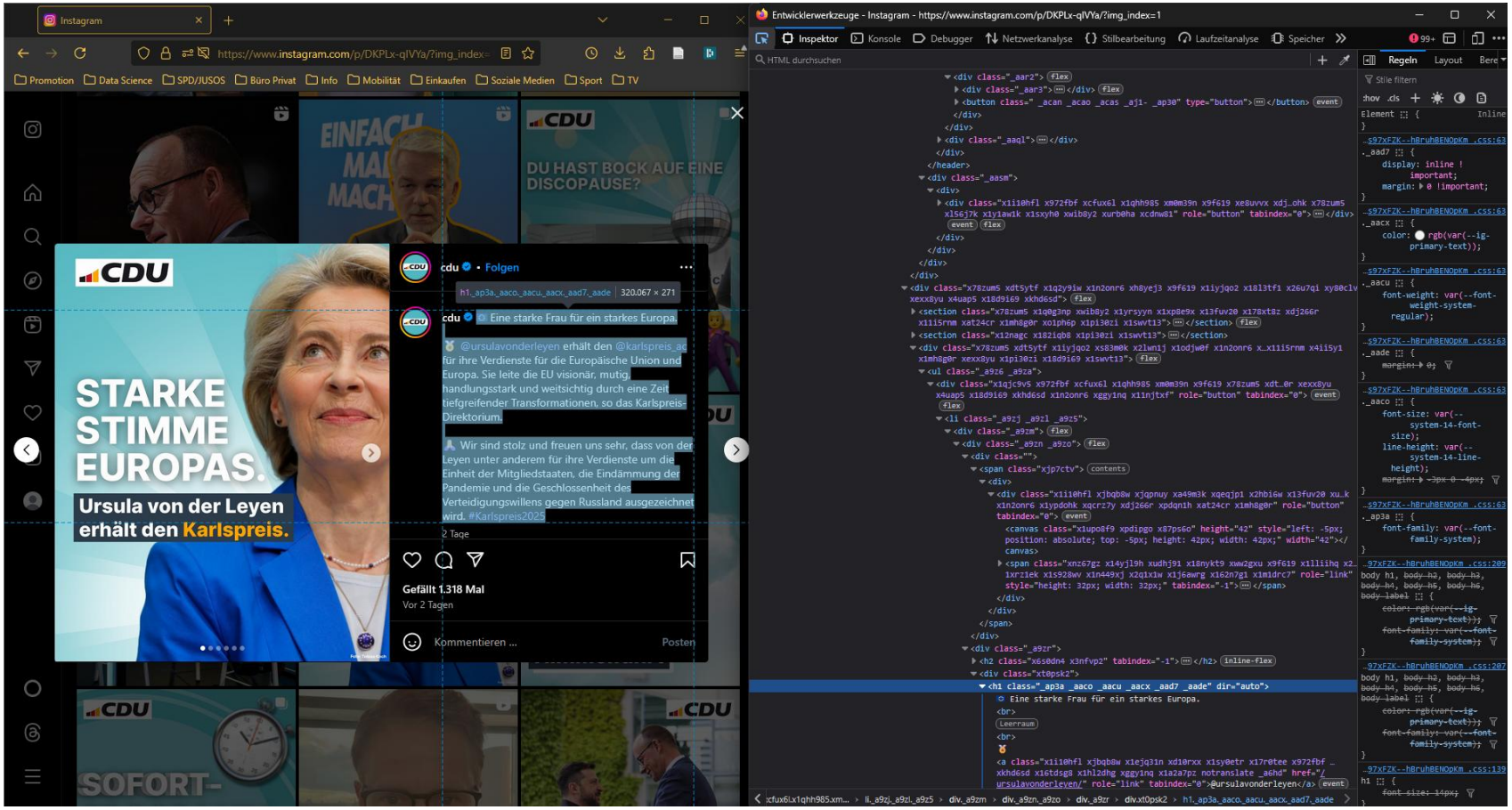
→ 58,1% could be linked with an Instagram account

Scraping: Identification

1. Identification via user name search
2. Identification via possible following or followed accounts, e.g. party accounts or other politicians from the local area
3. Identification via party und candidate websites

Universität Regensburg

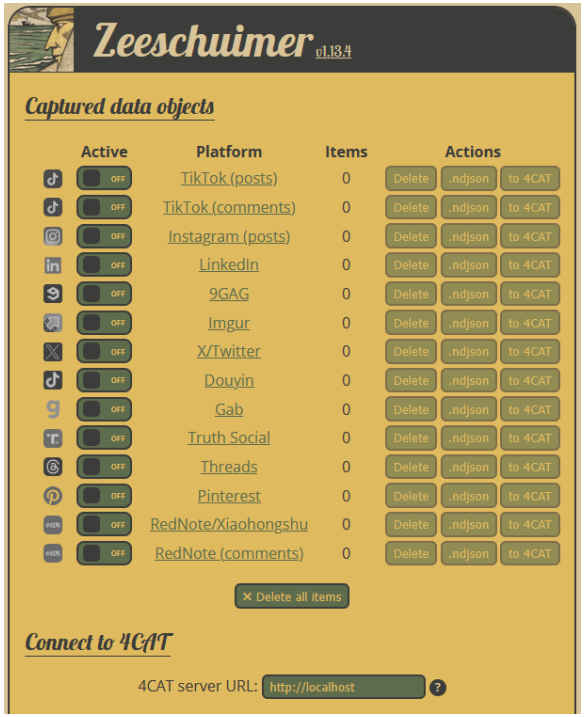
Scraping: Zeeschuimer



Universität Regensburg

Scraping: Zeeschuimer

- HTML elements of the Instagram page are loaded into the browser through data requests during scrolling
- FoxScroller enables automated and high-speed scrolling through Instagram profiles, facilitating efficient data collection
- During this process, the relevant information is extracted and consolidated into an NDJSON file

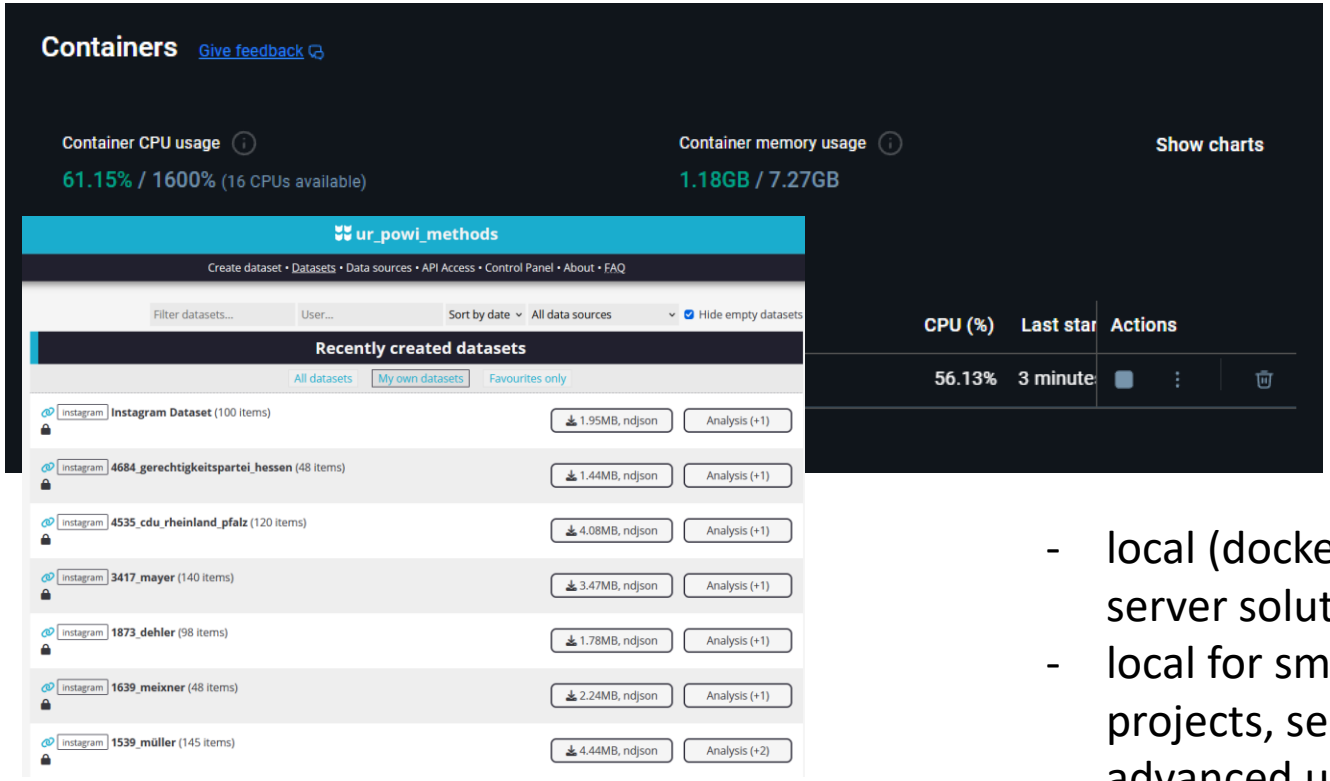



```
{"name": "Anna", "alter": 28}
{"name": "Ben", "alter": 34}
{"name": "Clara", "alter": 22}
```

{"code": "DGya9hrNFAt", "pk": "3581043230968664109", "id": "3581043230968664109_1590419560", "ad_id": null, "boosted_status": null, "boost_unavailable_identifier": null, "boost_unavailable_reason": null, "caption": {"has_translation": null, "created_at": 1742154157, "pk": "18021918956663910", "text": "#s\u00f6derdreht Niemand kann so schnell den Kopf drehen wie Markus S\u00f6der. Vielleicht noch Friedrich Merz. \nNat\u00fcrlich ist es richtig, die Schuldenbremse zu reformieren! Das fordern wir seit Jahren, haben dazu der Opposition immer wieder Vorschl\u00e4ge gemacht und waren im Wahlkampf ehrlich. \nW\u00e4hrend die Union hier unehrlich war. @florian_ossner rechnete gar beim bdslandshut vor, beim B\u00fcrgergeld 18 Milliarden sparen zu wollen. Die dann ausreichen sollten f\u00fcr die notwendigen Investitionen in die Infrastruktur. Beides falsch: so viel kann beim B\u00fcrgergeld nicht eingespart werden und der Investitionsbedarf ist weit h\u00f6her.\n\nDie Situation ist seit dem Wahltag kein bisschen neu. Trump stellte zuvor bereits die NATO in Frage und wie auch vor der Wahl brauchen wir massive Investitionen in die Verteidigung, die Infrastruktur, in Stromleitungen und Speicherkapazit\u00e4t, in unsere Schulen und die Bildung, in die Schiene und vieles mehr. \n\nWas wir nicht brauchen sind teure Sonderschulden f\u00fcr Wahlgeschenke. \n\nDer erneute Sieg von Trump machte bereits deutlich, dass die L\u00fcge sich lohnt. Genau das schadet der Demokratie, weil die st\u00e4ndige L\u00fcge dazu f\u00fchrt, dass die Menschen gar nichts mehr glauben.\n\nOb Merz morgen wohl vorschl\u00e4gt Kapitalertr\u00e4ge st\u00e4rker bei der Stabilisierung der Sozialversicherungen einzubeziehen? Vielleicht mit einer progressiven Kapitalertragssteuer?"}, "caption_is_edited": true, "feed_demotion_control": null, "feed_recs_demotion_control": null, "taken_at": 1741113641, "inventory_source": null, "video_versions": null, "is_dash_eligible": null, "number_of_qualities": null, "video_dash_manifest": null, "image_versions2": {"candidates": [{"url": "https://scontent-fra3-1.cdninstagram.com/v/t51.2885-15/482406256_18490970113019561_4576936578759673738_n.jpg?stp=dst-jpg_e35_tt6&efg=eyJ2ZW5jb2R1X3RhZyI6IkZFRUQuaW1hZ2VfdXJsZ2VuLjEzNTl4MTU3Ny5zZHIuZjc1NzYxLmRlZmF1bHRfaW1hZ2UifQ&_nc_ht=scontent-fra3-1.cdninstagram.com&_nc_cat=105&_nc_oc=Q6cZ2QEoFZsGGXUUh9fRZ8AHiOKGs5iBv5XxfdtsL67nomqNv8R5Jy3VZFBAwZWyG9g-TXY&_nc_ohc=tVqjjC03AhkQ7kNvwHUJbqc&_nc_gid=xH7Nebhg0J1lljMyAWgbVA&edm=AP4sbd4BAAAA&ccb=7-5&ig_cache_key=MzU4MTA0MzIzMDk2ODY2NDEwOQ%3D%3D.3-ccb7-5&oh=00_AfExST7LfmQGutxoHabZdpqYDN_92YgUQJ2x7v-58Njvpw&oe=67FD3872&_nc_sid=7a9f4b", "height": 1577, "width": 1359}, {"url": "https://scontent-fra3-1.cdninstagram.com/v/t51.2885-15/482406256_18490970113019561_4576936578759673738_n.jpg?stp=dst-jpg_e35_p750x750_sh0.08_tt6&efg=eyJ2ZW5jb2R1X3RhZyI6IkZFRUQuaW1hZ2VfdXJsZ2VuLjEzNTl4MTU3Ny5zZHIuZjc1NzYxLmRlZmF1bHRfaW1hZ2UifQ&_nc_ht=scontent-fra3-1.cdninstagram.com&_nc_cat=105&_nc_oc=Q6cZ2QEoFZsGGXUUh9fRZ8AHiOKGs5iBv5XxfdtsL67nomqNv8R5Jy3VZFBAwZWyG9g-

Universität Regensburg

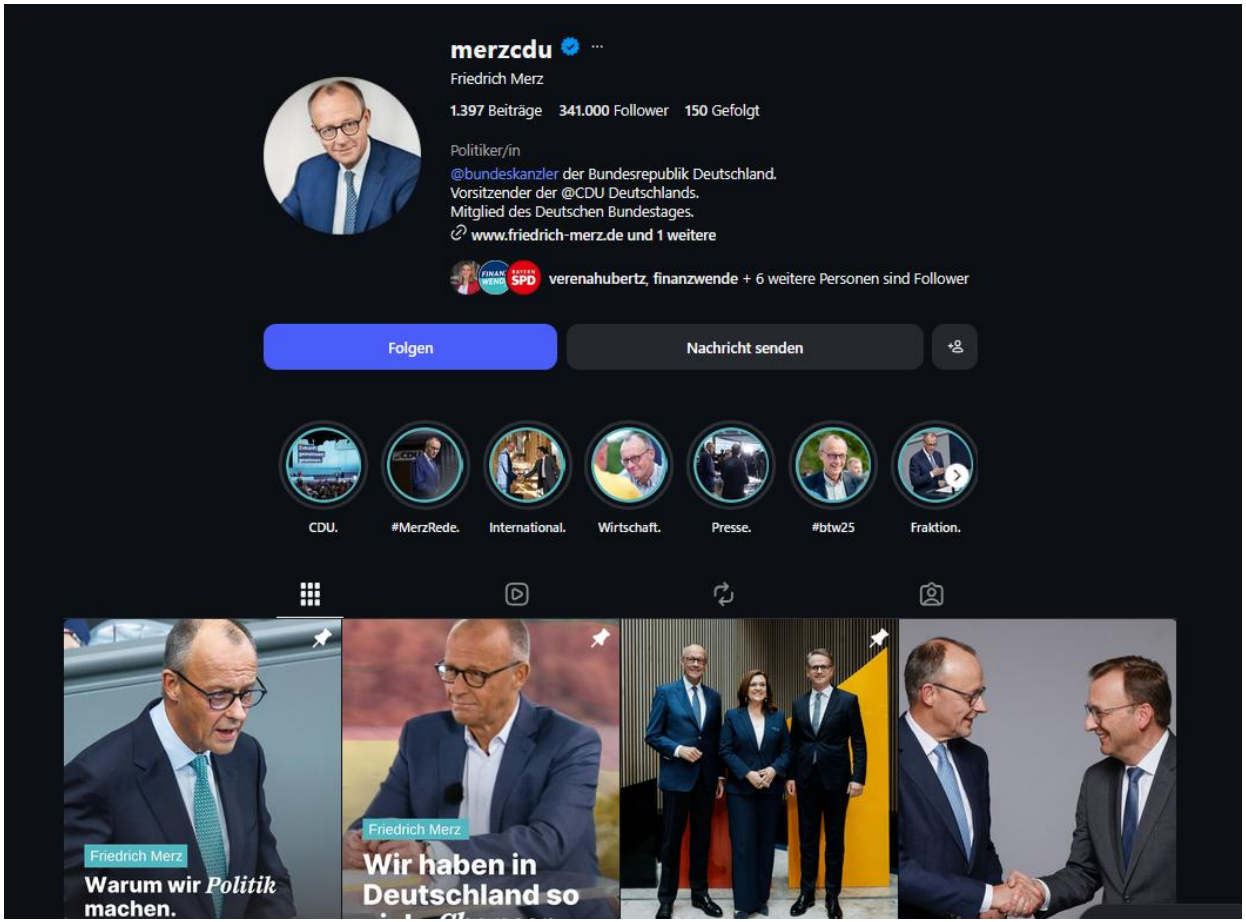
Scraping: 4CAT



- local (docker.desktop) or server solution
- local for smaller data projects, server solution for advanced users and research groups

Universität Regensburg

LIVE DEMO



Universität Regensburg

Optical Character Recognition

A process used to automatically recognize text from images or scanned documents and convert it into machine-readable text.



MORGEN IST WEIHNACHTEN!



T FOR vielfalt FüRUNSERE ZUKUNFT MIt ROBERT BLEIBTS COOA

Sidequest: Captioning

- Image captioning automatically generates textual descriptions of visual content, enables the analysis of visual content using text-based methods.
- Resource-intensive and time-consuming (in my use case!)
- Limited output, no classification benefit or performance improvement (in my use case!)

Universität Regensburg

Legal Challenges: Data Protection

Disclaimer: German/European Law

- Subject of research: Personal data that can be used to draw conclusions about an identifiable individual.
- Legal basis: For scientific research purposes, Article 89 DSGVO provides certain exemptions, subject to specific conditions (e.g., purpose limitation and anonymization), commonly referred to as the research exemption.
- Informing data subjects is often not practicable in large-scale research projects. In such cases, Article 14(5) DSGVO may apply if providing the information would involve a disproportionate effort.
- Legitimate interest (Article 6(1)(f) DSGVO) may serve as a legal basis for scientific research, provided that a careful balancing of interests is conducted.
- Under German and European legal practice, politicians and electoral candidates are generally considered public figures or persons of public interest during election campaigns; therefore, anonymization is typically not required.

Universität Regensburg

Legal Challenges: Copyright

Disclaimer: German/European Law

- Content such as images, videos, and texts is protected by copyright.
- Quotation is permitted, but only within the scope of scientific use and with proper source attribution.
- Text and data mining is permitted in the European Union for scientific research purposes (Article 3 of the DSM Directive; research exception under Section 60d of the German Copyright Act, *UrhG*).
- Data must be processed, analyzed, and stored in a transparent, secure, and purpose-specific manner.
- No publication of copyrighted content beyond what is necessary for scientific reporting and citation.

Universität Regensburg

Storing



Universität Regensburg

Storing

- Data set is archived in the University of Regensburg DataHub/Publication Repository.
- Repository provides permanent access to the dataset documentation and codebook.
- Researchers can request access to the dataset and associated image material.
- Access requests are reviewed to ensure that the intended use serves legitimate scientific research purposes.
- Upon approval, data may be shared in accordance with the applicable provisions of the GDPR research exemption framework.
- Persistent DOI is assigned, enabling reliable citation and long-term accessibility.
- Transparency, reproducibility, and controlled data access.

Universität Regensburg

FrontEnd-Scraping vs. API-Solutions

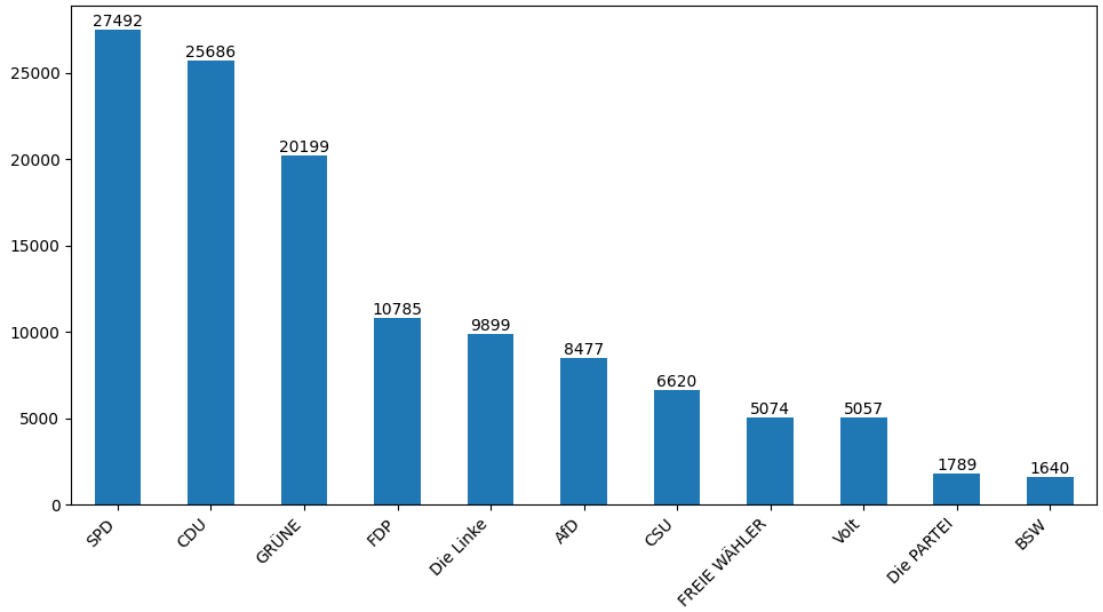
- The Zeeschuimer/4CAT workflow is legally compliant and aligned with Meta's Terms of Use.
- Finding relevant profiles remains the main bottleneck, regardless of the data source.
- Beyond profile identification, the practical differences between API-based and scraping-based approaches are often limited.
- No dependence on Meta approvals, applications, or access permissions.
- Supports transparent, independent, and reproducible Open Science research practices.
- Researchers retain full control over data collection and research design.

Universität Regensburg

Output

- 2,888 campaign profiles, comprising 270 party accounts and 2,618 candidate profiles (out of a total of 4,506 candidates).
- A total of 149,365 posts published during the observation period from 6 November 2024 to 23 February 2025 were collected.

Top 11 most frequent parties



Universität Regensburg

Output - Variables

- IDs (Unique, URL, Scraping)
- Timestamps
- Media Type (video, mixed, picture)
- Body/OCR-Output
- Performance Data (likes, comments, followers, interaction_rate)
- Location
- Personal Data (birth year, job, sex)
- Political Data (party, list position, electoral district)

Problems & Missing Data

- Follower?
- Colab?
- Ads?
- Comments?
- Whole Reels/Videos?

Evaluation

- A stratified sample evaluation was conducted to ensure the quality of the collected dataset.
- 6,000 posts were manually reviewed for structural errors.
- Posts were sampled across different strata to ensure representative quality assessment.
- The evaluation focused on identifying and correcting data extraction and formatting issues.
- Results were used to validate the reliability and integrity of the dataset.

But: What can be done with the dataset?

Universität Regensburg

But: What can be done with the dataset?

Project 1: Political Advertising and Engagement

Universität Regensburg

But: What can be done with the dataset?

Project 1: Political Advertising and Engagement

Project 2: Celebrity Endorsements and Political Communication

Universität Regensburg

Research Project I – Political Advertising on Instagram

- Combination of Instagram data and Meta Ads Library data
- Research question: Do sponsored posts achieve higher engagement?
- Key findings: Sponsored posts show slightly higher interaction rates after controlling for account effects
- Higher advertising spending is associated with higher engagement, although effects remain limited

Universität Regensburg

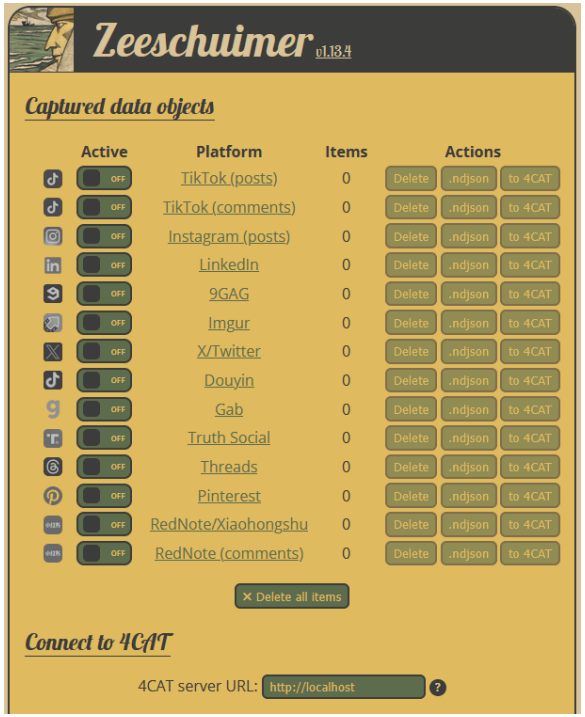
Research Project II – Celebrities and Political Communication

- Additional Dataset: 330,562 Instagram posts from 200 German and U.S. celebrities
- Research Question: How prevalent is political communication by celebrities on Instagram, and how does it affect engagement with political content?
- Automated classification of political celebrity communication and endorsements using LLMs
- Key findings: Political communication by celebrities is rare but measurable, Celebrity support increases interaction rates of political posts by approximately **44%**

Universität Regensburg

Outlook

- Ephemeral Content (Stories)
- Other Platforms (TikTok, LinkedIn)
- Video Scraping



Universität Regensburg

Thanks to:

Prof. Dr. Melanie Walter Rogg, Sonja Bachleitner, Lisa Brosch, Dominik Dächert, Dominik Ganslmeier, Christoph Götz, Amelie Holder, Lucas Hundhammer, Luca Klein, Adina Kükelhahn, Marisa Ludt, Anne Nothtroff, Leo Mühlbauer, Lea Meridies, Anett Pietsch, Timo Ritter, Nikolas Gregor Sadovsky, Nick Soestwöhner, Bianca Schwarz, Franziska Schierlinger, Phillip Sommer, Vincent Welsch…

and **Michael Achmann** for introducing me to the Zeeschuimer-4CAT-Workflow!

Universität Regensburg

Sources:

- Die Bundeswahlleiterin (2025): Wahlbewerberinnen und Wahlbewerber für die Wahl zum 21. Deutschen Bundestag (23. Februar 2025). Stand: 07.02.2025, 15:25 Uhr. Verfügbar unter: https://bundeswahlleiterin.de/bundestagswahlen/2025/ergebnisse/opendata.html Lizenz: Datenlizenz Deutschland – Namensnennung – Version 2.0 (https://www.govdata.de/dl-de/by-2-0)
- Peeters, S. (2023). Zeeschuimer (v1.8.0). Zenodo. https://doi.org/10.5281/zenodo.8399900
- Peeters, S., Hagen, S., & Wahl D. (2023). 4CAT Capture and Analysis Toolkit (v1.36). Zeno-do. https://doi.org/10.5281/zenodo.8139174
- Berg, J. (2026). Instagram-Beiträge aus dem Bundestagswahlkampf 2025 [Datensatz]. Universität Regensburg. https://doi.org/10.5283/epub.81002
- Berg, J. (2026). Instagram-Beiträge von Prominenten aus den USA und Deutsch-land [Datensatz]. Universität Regensburg. https://doi.org/10.5283/epub.78663

Universität Regensburg

Thank you for your attention!

Scan for Contact, LinkedIn, GitHub and Datasets: