Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Interactive Visualization of Telegram

Jeong-Eun Choi, Karla Schäfer, Martin Steinebach

Fraunhofer Institute for Secure Information Technology — ATHENE, Rheinstraße 75, 64295 Darmstadt, Germany
jeong-eun.choi@sit.fraunhofer.de

Abstract

We present an interactive visualization for exploring a large-scale Telegram dataset. The demo primarily features (1) the interactive exploration of information flows between channels/groups and (2) a content analysis based on established NLP techniques. It demonstrates how diverse information extracted from complex, heterogeneous social media data can be integrated into a cohesive interface. The demo highlights the potential of technical tools to support interdisciplinary research, such as studies on disinformation, and provides users with interpretable insights for further analysis.

Dataset — https://zenodo.org/records/16994657
Video — https://youtu.be/HSJwSDrIdro
Github — https://github.com/jechoi2021/TelegramDemo

Introduction

Social media data offers rich insights into diverse aspects of human activity, including entertainment, education, economics, public discourse, and private communication, making it a valuable resource across disciplines. However, drawing meaningful insights from raw datasets is challenging without a clear understanding of platform structures, user behavior, and information flow. Even with such knowledge, integrating heterogeneous information into coherent analysis remains complex.

Currently, journalists, social scientists, and communication scholars rely heavily on subjective observation and qualitative methods. This reliance can create dilemmas for journalists when deciding what to publish (Ferreira and Daoust 2025). Although quantitative approaches are increasingly used due to the growing availability of data and analytical techniques, commonly referred to as computational social science, recent studies still report a gradual and steady increase in qualitative research (Zakopoulos and Xanthopoulou 2026). Concepts such as collectivism and perspectivism (Lapinski and Rimal 2006), for example, play important roles in selecting meaningful patterns or subsets relevant to specific research questions. While these approaches are valuable and often necessary, they face limitations in scalability, replicability, and objectivity. Although scalability and replicability are being increasingly addressed through the use of appropriate big data technologies and the availability of datasets, subjectivity remains less explored. Such subjectivity can lead, for example, to an overemphasis on studying negative phenomena in social media, while opportunities to steer research toward positive outcomes remain underexplored (Antonini 2023).

From a technical perspective, data-driven metrics and insights can complement subjective approaches by providing additional analytical perspectives. However, many non-technical users struggle to understand the capabilities and limitations of AI-based tools, such as deep learning models, whose lack of explainability may lead to either uncritical adoption or avoidance. This motivates the need for tools that provide transparent, interpretable cues, enabling practitioners to explore their data proactively and align insights with their goals.

Developed within the interdisciplinary Dynamo project[1][2], our demo is designed for researchers, practitioners such as journalists, and even the general public with non-technical backgrounds. Rather than providing definitive answers, it fosters exploratory engagement with complex datasets, helping balance subjective interpretation with data-driven exploration. By embracing the non-linear and dynamic nature of social media data (Anderson and Millard 2023), the tool encourages flexible and reflective analysis. While the current implementation focuses on Telegram, the underlying design is adaptable to other platforms with similar data structures, supporting analytical, research-driven exploration of content and interactions between entities.

Motivation & Target Users

The target users of our demonstrator are researchers and practitioners from non-technical backgrounds who aim to understand and analyze social media data. Social media platforms are complex and dynamic, and interpreting them often relies on subjective observation, which can introduce biases and limit systematic understanding. While the increasing availability of AI and computational methods offers opportunities for more structured analysis, current tools still em-

[1]Project: https://www.dynamo.sit.fraunhofer.de/

[2]Online event in German: https://www.youtube.com/watch?v=vtErfvgyLT8

phasize either management or narrowly defined tasks, such as monitoring influence or detecting misinformation, rather than supporting exploratory, research-driven inquiry.

Many online tools, often categorized as social media listening or monitoring platforms, offer insights into social media activity but typically rely on basic metadata to measure aspects such as influence or virality. Recently, there has been a growing trend toward integrating AI for enhanced content analysis or even content creation, as seen in tools such as Swat.IO[3], Meltwater[4] and Hootsuite[5]. However, they are primarily designed for social media management, with clear goals such as optimizing communication strategies or tracking marketing performance. Consequently, platforms such as Telegram, which are less suitable for marketing purposes, are often not covered by these tools. Moreover, these platforms rarely aim to provide a deeper understanding of the broader ecosystem. Instead, their strengths lie in real-time monitoring geared toward financial or branding outcomes.

Others focus specifically on Telegram data. For example, TeleCatch (Ruscica, Tucci, and Carneiro 2025) is designed to collect Telegram data and prepare it for analysis. The MST platform (Claudino et al. 2023) describes itself as a data management system for Telegram data and evaluates how a pre-trained classifier for misinformation can be used together with the data processed by the platform. Fact Flow, developed by Newtral [6], is designed to detect disinformation messages and monitor the narratives of such messages as well as suspicious channels using a AI-based pipeline.

In contrast, our demonstrator is designed to support exploratory and analytical engagement with social media data, particularly for social and research-driven objectives. When data is provided, our demonstrator can be hosted locally. Unlike systems such as the MST platform or Fact Flow, which focus on specific domains such as misinformation detection, our approach does not require prior domain specification. Furthermore, rather than primarily focusing on collecting and managing data, as tools like TeleCatch do, our system emphasizes understanding the complex interaction dynamics between entities and posts and provides richer content analysis.

System Description

Our dataset consists of 995 public Telegram entities (channels/groups) and was collected via Telegram API from 25-03-2022 to 30-06-2023. These entities were selected through interviews with journalism and fact-checking experts, conducted by our partners at Hochschule der Medien as part of the project Dynamo to identify influential Telegram actors in the spread of disinformation in Germany. As such, interpretations of the dataset should be made with care and contextual awareness. Further details are available in our analysis papers (Choi, Schäfer, and Steinebach 2024; Choi, Schäfer, and Yannikos 2024; Schäfer and Choi 2023).

About our Demo Through metadata analysis, we were able to identify and measure the flow of information between entities. In our visualization, users can select one entity (main node) and will then be provided with entities in connection (neighbouring nodes). In the graph, arrows are used to indicate the direction of content sharing. An inward arrow (green) indicates that the selected entity has reposted content from another entity, i.e. it was influenced by that entity. Conversely, an outward arrow (orange) indicates that other entities have reposted content from the selected entity, i.e. it exerted influence on others. The thickness of the edges represents the strength of this information exchange, measured by the number of posts forwarded. Users can apply a threshold filter to adjust the visibility of edges, showing only those with higher volumes of information exchange or use filters to view only incoming or outgoing edges.

We defined an influential metric for each channel and group, combining features such as the number of participants, average posts per day, the ratio and mean of forwarded original posts, and the percentage of forwards originating from channels within the dataset. This metric provides a compact estimation of a channel's or group's influence, based on metadata tailored to the ecosystem, as defined in the following formula:

$\psi :=$ Number of participants,
$\delta :=$ Average posts per day,
$\omega :=$ Number of original posts with forwards,
$\phi :=$ Mean forwarded counts of original posts,
$\mu :=$ % of Forwarded Posts coming from within the dataset

$$\text{Influential\_Metric} = \psi + 0.8\,\delta + 0.3\,\frac{\omega^2 \cdot \phi}{\text{Original\_Posts}} + 0.1\,\mu$$

Yellow nodes indicate entities identified as influential within the dataset, with node size corresponding to the number of participants. Selecting a neighboring node displays its metadata analysis on the right in the InfoCard. Bolded numbers in the InfoCard represent values in the 90th percentile, highlighting extreme or notable feature metrics. See the demo video for further visualization details.

The "More Info" button on the InfoCard displays the results of the content analysis by showing the top five topics identified through topic modeling. we used BERTopic (Grootendorst 2022) as a topic modeling approach, and SentenceTransformer all-MiniLM-L6-v2[7] as embedding. For channels or groups with a large number of posts, we randomly selected 10,000 posts for topic modeling. Additionally, only texts containing more than six words were considered. For efficiency and runtime considerations, we applied UMAP for dimensionality reduction and HDBSCAN for clustering. To generate topic names, we selected a representative post from each cluster and used microsoft/Phi-3.5-mini-instruct[8] to produce an appropriate topic label. For each topic, we performed sentiment analysis to classify posts as positive, neutral, or negative and present

[3]https://swat.io/de/

[4]https://www.meltwater.com/

[5]https://www.hootsuite.com/

[6]https://www.journalismai.info/programmes/innovation/innovation-challenge-2024/newtral

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[8]https://huggingface.co/microsoft/Phi-3.5-mini-instruct

the results as percentages. Since the dataset primarily consists of German posts, we used the GermanSentiment (Guhr et al. 2020) model for sentiment classification.

Our demo provides a visual workspace for exploring Telegram using a crawled dataset of interest. Topic Modeling and sentiment analysis are central components that go beyond mere description. Topic-level information allows users to identify dominant themes across channels and posts, supporting tasks such as profiling channels by content focus, summarizing trends, and clustering posts with similar content. Sentiment-level information provides insight into the emotional valence of discussions, enabling researchers to examine how positive, neutral, or negative content propagates across the network. For more, see our Github or video.

Future Work and Limitations

In future work, the demo can be extended with additional functionalities. These include filtering interactions for posts of interest, such as on topics or sentiments relevant to specific research questions; downloading findings and results as files for further analysis; visualizing topical relatedness and reconstructing the dissemination behavior of individual posts or aggregated posts with similar content; profiling channels and groups, building on our existing design for channel-level visualization; evaluating applicability and scalability across multiple publicly available datasets; and providing influential metrics to provide multiple ways of measuring influence within the observed dataset.

The implementation of these functionalities are supported by a clear conceptual mapping of channels and posts, enabling the development of rich entity-level profiles and a better understanding of interactions between these elements. For the current implementation, individual users are not tracked, both because they are often anonymized and to ensure that the analysis respects user privacy.

The main limitation of the demo is the lack of real-time analysis. Preparing the full 37.22 GB dataset required approximately 15 hours, with topic modeling being the most time-consuming step. Data preparation for React visualization also takes several hours, though parallelization could reduce processing time.

While user testing has been limited, planned extensions will enable broader evaluation with additional researchers and datasets, helping validate the applicability and scalability of the demo in diverse research scenarios.

Conclusion

Our demo offers an interactive tool for non-technical users to explore Telegram. It facilitates the exploration of social media behaviour and content dissemination, fostering data-driven insights. It helps users to balance subjective interpretation with objective analysis, making it a platform for interdisciplinary research.

Acknowledgments

This research work was supported by the National Research Center for Applied Cybersecurity ATHENE [9] as well as within the German Federal Ministry of Education and Research project DYNAMO [10].

References

Anderson, M. W. R.; and Millard, D. E. 2023. Seven Hypertexts. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702327.

Antonini, A. 2023. Positive by Design: The Next Big Challenge in Rethinking Media as Agents? In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702327.

Choi, J.-E.; Schäfer, K.; and Steinebach, M. 2024. Creating Visual Persona Profiles in Telegram using NLP. *Electronic Imaging*, 36(1): 359–1–359–6.

Choi, J.-E.; Schäfer, K.; and Yannikos, Y. 2024. Scientific Appearance in Telegram. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 2091–2096.

Claudino, I.; Gadelha, T.; Vinuto, T.; Franco, J. W.; Monteiro, J. M.; and Machado, J. 2023. A Real-Time Platform to Monitoring Misinformation on Telegram. In *Proceedings of the 25th International Conference on Enterprise Information Systems - Volume 1: ICEIS,*, 271–278. INSTICC, SciTePress. ISBN 978-989-758-648-4.

Ferreira, R. R.; and Daoust, J.-F. 2025. To Report or Not to Report? A Qualitative Analysis of Journalists' Perspectives on Harm to Public Opinion. *Public Opinion Quarterly*, 89(SI): 683–715.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Guhr, O.; Schumann, A.-K.; Bahrmann, F.; and Böhme, H. J. 2020. Training a Broad-Coverage German Sentiment Classification Model for Dialog Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1620–1625. Marseille, France: European Language Resources Association.

Lapinski, M. K.; and Rimal, R. N. 2006. An Explication of Social Norms. *Communication Theory*, 15(2): 127–147.

Ruscica, G.; Tucci, G.; and Carneiro, B. 2025. TeleCatch: An open-access software for visualizing, filtering and extracting Telegram messages data. *Software Impacts*, 23: 100736.

Schäfer, K.; and Choi, J.-E. 2023. Transparency in Messengers: A Metadata Analysis Based on the Example of Telegram. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702327.

Zakopoulos, V.; and Xanthopoulou, P. 2026. Mapping Qualitative Research in Social Sciences and Humanities: A Bibliometric Review. *Encyclopedia*, 6(3).

[9] https://www.athene-center.de/forschung/revise

[10] https://www.sit.fraunhofer.de/de/dynamo/

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? NA
   (g) Did you discuss any potential misuse of your work? NA
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? NA
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? - YES (the published data is pseodonymized)
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA