Media and Communication
2024 • Volume 12 • Article 7789
https://doi.org/10.17645/mac.7789

cogitatio

ARTICLE

Open Access Journal

The Challenges of Replicating Volatile Platform-Data Studies: Replicating Schatto-Eckrodt et al. (2020)

Philipp Knöpfle [1] and **Tim Schatto-Eckrodt** [2]

[1] Department of Media and Communication, Ludwig Maximilian University of Munich, Germany
[2] Department of Communication and Journalism Studies, Hamburg University, Germany

Correspondence: Philipp Knöpfle (philipp.knoepfle@ifkw.lmu.de)

Submitted: 31 October 2023 **Accepted:** 26 February 2024 **Published:** 15 April 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Replication studies in computational communication science (CCS) play a vital role in upholding research validity, ensuring reliability, and promoting transparency. However, conducting such studies in CCS often proves challenging due to the data environments' dynamic nature and the complexities surrounding data and software sharing. To shed light on these challenges, we examine the replication process with CCS studies by computationally reproducing and replicating Schatto-Eckrodt et al.'s (2020) computational analysis of the X (formerly Twitter) debate about the term "gaming disorder" being added to the International Classification of Diseases 11. Our results indicate a reproduction success rate of 88.46% of the original findings. Replicating the analysis presents several obstacles, particularly in data access and availability. Five years after the original data collection, we were able to recollect only 55.08% of the initial sample, primarily due to user and platform activities, including account deletions, user suspensions, and privacy settings. Our reproduction and replication efforts revealed intricate challenges in conducting CCS research, particularly concerning data access and third-party platforms. To enhance replication in CCS, we emphasize the crucial role of data sharing, increased transparency, extensive documentation, and regulatory processes. Thus, our analysis underscores replications' critical role in enhancing CCS research validity and reliability.

Keywords

computational communication science; replicability; replication; reproducibility; Twitter

1. Introduction

Replication studies play a critical role in scientific research, functioning as a litmus test of research findings' validity and reliability. Their significance cannot be underestimated in their contribution to promoting transparency (Munafò et al., 2017) and fostering trust in empirical results (Rosenthal, 1991). Replications should be viewed as a social science principle because they strengthen the foundation of rigorous and verifiable research, fortifying empirical findings' validity and reliability (Benoit & Holbert, 2008). In meta-science, the concepts of reproduction and replication play pivotal roles in assessing scientific research reliability and robustness. Conducting a reproduction involves faithfully recreating a study's original conditions, methodologies, and analyses, including the acquisition of original data, to verify whether the same or quantitatively similar results can be obtained. Thus, reproduction concerns verifying the original study's validity and reliability (National Academies of Sciences, Engineering, and Medicine, 2019). However, replication entails conducting a novel study that makes variations to the original study's parameters (e.g., methodology, context, sample populations, etc.) to investigate the initial results' generalizability and robustness (National Academies of Sciences, Engineering, and Medicine, 2019). Replications can span a continuum of methodological resemblance to an initial study (LeBel et al., 2017; Machery, 2020). For example, a direct replication, characterized by high methodological similarity to the original study, involves reiterating the latter using methods as closely aligned with the original as reasonably possible. The goal is to anticipate consistent results based on current understanding of the phenomenon (Nosek et al., 2012). Conversely, conceptual replications can be used as experimental procedures designed to assess generalizability and veracity by varying a study's selected operational characteristics, such as omitting or including certain variables (Hendrick, 1990). In communication science, conceptual replications are more common than direct replications (Keating & Totzkay, 2019).

Nevertheless, the pursuit of replication and reproduction studies entails overcoming multiple barriers, such as (technical) resource constraints and complex research designs (Peng, 2011). This is particularly true in the field of computational communication science (CCS), in which researchers regularly face delicate and complex data environments and experimental contexts (van Atteveldt & Peng, 2018). CCS researchers also face multiple obstacles when it comes to conducting replications, e.g., working with personal or sensitive data, dealing with copyright restrictions, and relying on third-party social media platform data. Researchers in the CCS field commonly lack control over the data they work with, as they lack authority over generation of analyzed data, e.g., setting up or documenting experimental data collection protocols. In the replicability context, this means that while reacquiring experimental data for replication studies involves reproducing original experimental conditions to collect new empirical observations, reacquisition of content data is a very different endeavor. Content data, particularly in the case of social media data, is primarily not designed for research purposes and is inherently subject to algorithmic confounding (e.g., algorithmic opacity, data sampling biases, etc.), i.e., beyond researchers' control (Haim, 2023). Such data are made accessible to both developers and researchers as a byproduct of the data's inherent availability within the platform's framework (Bruns, 2019). Consequently, most content data are not intentionally structured for seamless replication (Davidson et al., 2023). Thus, CCS researchers are subject to stringent constraints imposed by technical restrictions, as well as terms of service imposed by platform providers to control access to and distribution of their data (Puschmann, 2019). These technical restrictions create significant barriers to sharing platform data with potential replicators, presenting a notable challenge because data access frequently functions as an essential prerequisite for successful reproduction and replication of a scientific analysis (Peng & Hicks, 2021).

Large social media platforms, such as X and Reddit, have become staple suppliers of data for CCS researchers (van Atteveldt & Peng, 2018). However, this dependence on third-party platform providers introduces significant challenges to achieving replication success (Davidson et al., 2023), particularly in the contemporary "post-API" (application programming interfaces) era, in which researchers face heightened difficulties in accessing data from social media platforms (Tromble, 2021). Re-collecting a study's original data is a necessary and pivotal step in the reproduction and replication process, thereby eliciting the question of to what extent reproductions and replications of social media analyses remain feasible in the current platform ecosystem. Our analysis investigates this by conducting a reproduction and replication of an analysis by Schatto-Eckrodt et al. (2020), who investigated discourse on the X platform concerning the introduction of the term "gaming disorder" into the International Classification of Diseases 11 (ICD-11). We reproduce the original analysis by reconducting the original analysis on Schatto-Eckrodt et al.'s data. A reproduction to us denotes researchers' capacity to re-execute or re-implement the same computational analyses and yield results that are quantitatively identical (Christensen & Miguel, 2018; Cohen-Boulakia et al., 2017; Peng, 2011). In the second step, we conduct a direct replication with rehydrated data from the original analysis. Our replication approach is to re-collect Schatto-Eckrodt et al.'s data on the X platform and conduct an identical analysis to verify to what extent a replication of original results is possible using today's data. The concepts of reproducibility and replicability are intertwined in our study. While we have access to the initial data and code necessary for our reproduction, for our replication, we must re-collect the data from the X platform. If the original authors had not provided us with their data, our reproduction approach would have necessitated a similar data re-collection procedure. Consequently, our replication also can be understood as a computational reproduction scenario comprising a situation in which the data are not initially accessible to the reproduction team. Thus, our insights into the re-collection of social media platform data are applicable to both replication and reproduction scenarios, providing additional valuable considerations for situations in which the reproduction team must recollect the original study's data in today's context. In our study, our objective is to shed light on the intricacies and challenges associated with conducting replication and reproduction studies in the CCS field, particularly in the contexts of data access and third-party platform providers, as Freiling et al. (2021) and Davidson et al. (2023) highlighted. We showcase unique obstacles that CCS researchers face and underscore recent API restrictions and cost increases' implications for replication analyses' feasibility.

2. Methodology

2.1. Reproduction: Data Collection and Analytical Strategy

In many cases, replicating scientific studies proves challenging due to research designs' complexities. As a result, reproduction has become the prevalent standard for evaluating scientific claims' credibility and significance (Peng, 2011; Peng & Hicks, 2021). Nevertheless, reproduction often suffers from constraints analogous to conducting replication analyses. One crucial step behind every reproduction is the reacquisition of the data from the original study. Overstating data sharing's importance is difficult because replicators benefit significantly when they can access data. Unfortunately, in most cases, the original study's researchers often have little incentive to share their data, e.g., sharing their materials can incur technical expenses (Longo & Drazen, 2016), and they may be under legal restrictions that prohibit them from disclosing their data (Rosenberg et al., 2020). If a study's original data are not available to reproducers, initial sample data from the original study must be re-collected, which frequently involves substantial initial expenditures, offering limited prospects for the reproducer, as exact recollection of primary data is not guaranteed. The latter is particularly problematic when reproducing content data analyses on social media platforms, where users and providers constantly generate, edit, and/or remove content. Moreover, most social media platforms, such as X, generally prohibit public sharing of user-level data obtained from the platform in their developer agreements. For our reproduction, we acquired the primary data set directly from the original authors, allowing us to compare the re-queried data with the original data. Considering that our study utilizes components from prior research—including code, data, and documentation—to reproduce initial findings, we classified our investigation as a computational reproduction (Ziemann et al., 2023), which differs from other types of reproducibility, such as analytical reproducibility (Hardwicke et al., 2018), also known as recreated reproducibility (Dreber & Johannesson, 2023), in which the goal is to reproduce an original study's findings based on information and documentation provided in the original article without access to the raw or processed data.

Notably, even though no shortage of resources exists for promoting reproducible research practices (Alston & Rick, 2021; Munafò et al., 2017; Stodden et al., 2014), no standardized templates or structured guidelines to date exist for documenting systematic reproduction of a computational social science article. Our approach aimed to address this gap using a simple methodology: We started by cataloging all critical empirical claims, including text passages and visual elements, from Schatto-Eckrodt et al., and setting out to reproduce these claims using the data and code that the original authors provided. This comparative analysis allowed us to verify each claim's accuracy on a granular level systematically, which facilitated the assessment of reproduction success. We viewed a reproduction as "successful" when we could reproduce the original result quantitatively, ensuring that it aligns with the claim presented in the original research. Notably, different standards are used to assess a reproduction's success rate, e.g., the exact quantitative agreement of results, margin of error (i.e., whether the reproduction results in a predefined range or margin of deviation), or by comparing statistical measures' (e.g., coefficient signs, confidence intervals, and/or hypothesis decisions) congruence. We chose the exact quantitative agreement of results, as it is the strictest criterion for assessing reproduction success. We allowed for minor code modifications to accommodate changes in software and dependencies over time, ensuring that the research remains adaptable and relevant.

2.2. Data Collection Replication Strategy

The Schatto-Eckrodt et al. data set was obtained originally by querying a database generated from the Decahose X API for gaming-disorder-related search terms. The Decahose API provides access to a 10% random sample of all public posts. Our study goal was to analyze to what extent we could replicate the data set with the data available today, so we aimed to re-collect the data under today's X access opportunities. Unfortunately, the Decahose API currently is limited to enterprise-level users, rendering it unavailable for our data collection purposes.

This left us with two viable data collection options: The first approach entailed reinitiating the identical search query applied to the Decahose data set on the X API. Subsequently, all posts and their associated information were collected accordingly. However, current X API access solutions have certain limitations, particularly when conducting queries encompassing wide-ranging terms, such as "gaming/games" and "disorder," as in our case. Open-ended, extensive queries swiftly deplete the allocated monthly post query limit, rendering the process time-consuming and costly. For this project, we adopted an alternative strategy. A "loophole" for researchers is to share each post IDs in their initial sample such that replicators selectively can re-query these exact posts at the API, a process also known as *rehydration*. This gives us precise knowledge about what specific information to request from the X API. Under current API circumstances, this allows for the most resource-efficient data collection. Moreover, the rehydration approach is well-suited for projects with a clearly defined event timeline because the chance of "missing out" on additional data outside of the primary sample is minimal when an analysis has a fixed scope and time frame. To sum up, rehydration emerged as the most cost-effective and pragmatic choice for our replication under current X access model restrictions. Notably, none of the approaches guaranteed that the initial data set could be fully re-collected in its original form.

3. Reproduction of Original Findings

3.1. Reproduction Setup and Computational Environment

The reproduction process was conducted in both the original computational environment (as documented in Schatto-Eckrodt et al.; Supplementary File A) and a contemporary updated environment (see our software bibliography in the Supplementary Material or our Open Science Framework [OSF] repository), with all *R* programming language dependencies updated to their latest versions. Schatto-Eckrodt et al. provided the computational environment and analysis code in an OSF project. The shared materials comprised individual files for each analysis, a README file with usage instructions, an *R* session info file, and all shareable data compliant with X's terms of service. For more detailed recommendations on how to make computational communication research more reproducible and accessible, see van Atteveldt et al. (2019). By reproducing the study in the original computational environment, we ensured that the results were consistent with the conditions under which the original findings were generated. Simultaneously, the examination in the updated environment ensured that the research would remain adaptable and relevant in the face of evolving open-source software landscapes. The reproduction was executed using the statistical programming language *R* (Version 4.0.0 in the original environment and Version 4.3.1 in the updated environment) in July and September 2023. Tim Schatto-Eckrodt was part of the original study, and his participation in the reproduction presented a conflict of interest, so Philipp Knöpfle conducted all reproduction activities and evaluations.

3.2. Reproduction Evaluation

In our computational reproduction, we identified 26 empirical claims in the original paper overall, including text passages, figures, and tables. We successfully reproduced 23, an 88.46% reproduction success rate. All tables from the original study were reproduced. Notably, minor challenges emerged during the reproduction of Figures 2 and 3 from the original article due to the deterministic reproducibility requirement for network illustration methods, which requires setting a seed, i.e., an aspect overlooked in the original script. To address this issue in future analyses, it is recommended that a seed be determined for illustration methods, incorporating a randomness component to ensure consistent and reproducible visualizations. Furthermore, one *t*-test statistic, as documented on page 211 of Schatto-Eckrodt et al., could not be recalculated in our reproduction. However, this non-execution did not alter the test outcome's results substantially and was used in a comparison argument, thereby not affecting the core conclusions of their research. Overall, the replication phase proceeded without larger issues. All code was executed without critical errors in its initial environment. Only minor code adjustments due to depreciated package dependencies for the exporting of results had to be made in the updated environment. Overall, this finding was positive for stability across computing environments. Our reproduction protocol can be found in the article's supplementary files, as well as the article's OSF repository (https://osf.io/2jb9m).

4. Replication of Findings

4.1. Rehydrated Sample

The "basic" subscription model—which replaced the previous, original, free access model in March 2023—permits a rate limit of 10,000 posts per month at a cost of $100 per month. The data collection process spanned two months—August and September 2023—yielding 9,270 successfully queried posts in the rehydrated sample. Table 1 illustrates the rehydrated sample's status, revealing that only 55.08% of the initial sample remained accessible on the X platform in September 2023. The data set experienced considerable changes: Users removed 28.32% of posts (deleted); 5.58% were set to private (protected); 10.62% were removed from the X platform because they violated the X terms of service (suspended); and 0.39% were generated by users who chose to deactivate their accounts temporarily (deactivated). This data landscape mirrored findings from a similar analysis in the original study, in which the original authors investigated how much of their sample was still available in 2020 by querying the X Compliance Firehose API. More than a quarter of the data already were unavailable, just two years after the data were collected initially in 2018. While existing literature on data loss rates on social media platforms supports these findings, other scholars have reported a diverse range of deletion rates. Depending on the time frame, deletion rates varied widely, with examples including 2% within one day after posting (Almuhimedi et al., 2013), 35.14% after one week (Zhou et al., 2016), 11.11% after five weeks (Bhattacharya & Ganguly, 2021), and 3.2% after two months (Petrovic et al., 2013). Specifically analyzing suspended X users' characteristics, Wei et al. (2016) found that X suspended 7.19% of the users after three years. While the rates found in the literature depended greatly on the studied content and user group (e.g., Zhou et al., 2016, focused on individual users' "regrettable" posts by individual users), studies conducted on random samples generated similar deletion rates: 7.27% after one week and 21.65% after six months (Schatto-Eckrodt, 2022).

Our analysis substantiated this, finding that the X trend of users opting to delete or protect their posts has increased in the past three years. Table 2 compares post types in the original and rehydrated samples, indicating that 3,074 original posts and 4,487 reposts could not be re-collected in 2023, resulting in an overall absolute data erosion of 7,561 posts, comprising 44.92% of the initial sample. Overall, in absolute terms, more reposts than original posts became inaccessible over the years. Figure 1 illustrates the data loss in post composition. The data loss in the rehydrated sample occurred particularly during the most heated phases of the discussion, as can be seen in Figure 2. Altogether, 1,934 posts were lost around the time of

Table 1. Sample changes from 2020 to 2023.

Example	2020	2023
Online	73.12%	55.08%
Deleted	12.24%	28.32%
Protected	1.30%	5.58%
Suspended	13.34%	10.62%
Deactivated	—	0.39%

Table 2. Changes in post composition from 2020 to 2023.

Post type	2020 (% of the original sample)	Data Loss (% compared to the original sample	2023 (% of the rehydrated sample)
Original posts	7,555 (44.89%)	−3,074 (40.69%)	4,481 (48.34%)
Re-posts	9,276 (55.11%)	−4,487 (48.37%)	4,789 (51.66%)
Sum	16,831 (100%)	−7,561 (44.92%)	9,270 (100%)

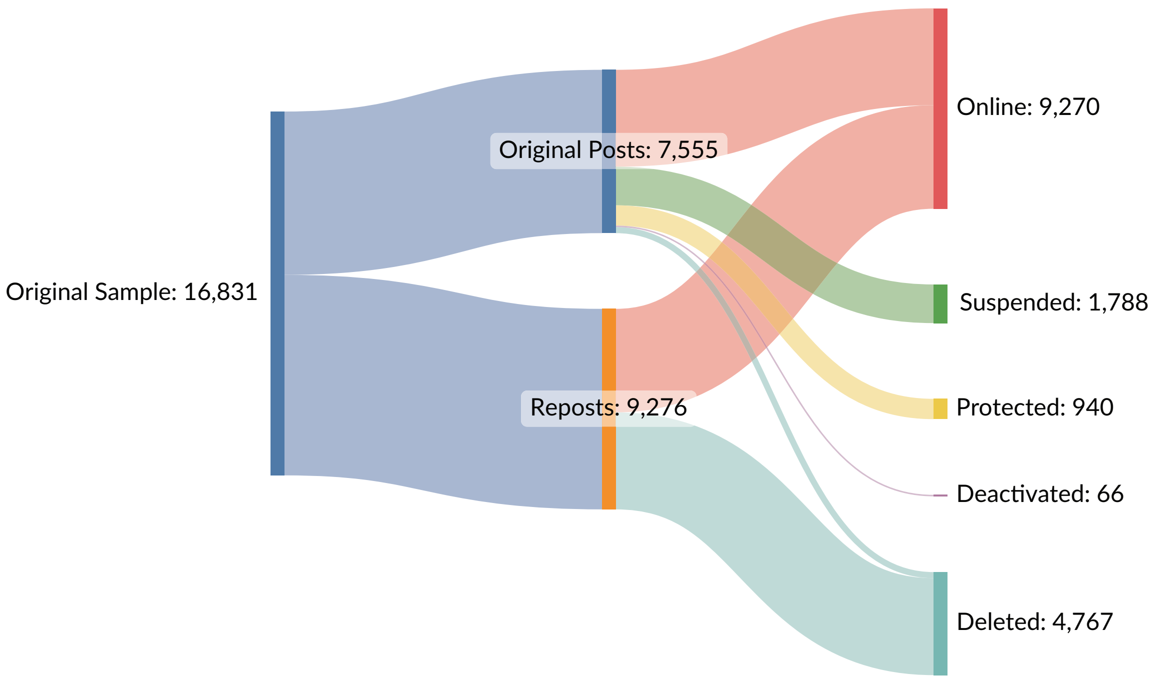

Figure 1. Sankey-diagram of the sample's structure from the original sample to the rehydrated sample.

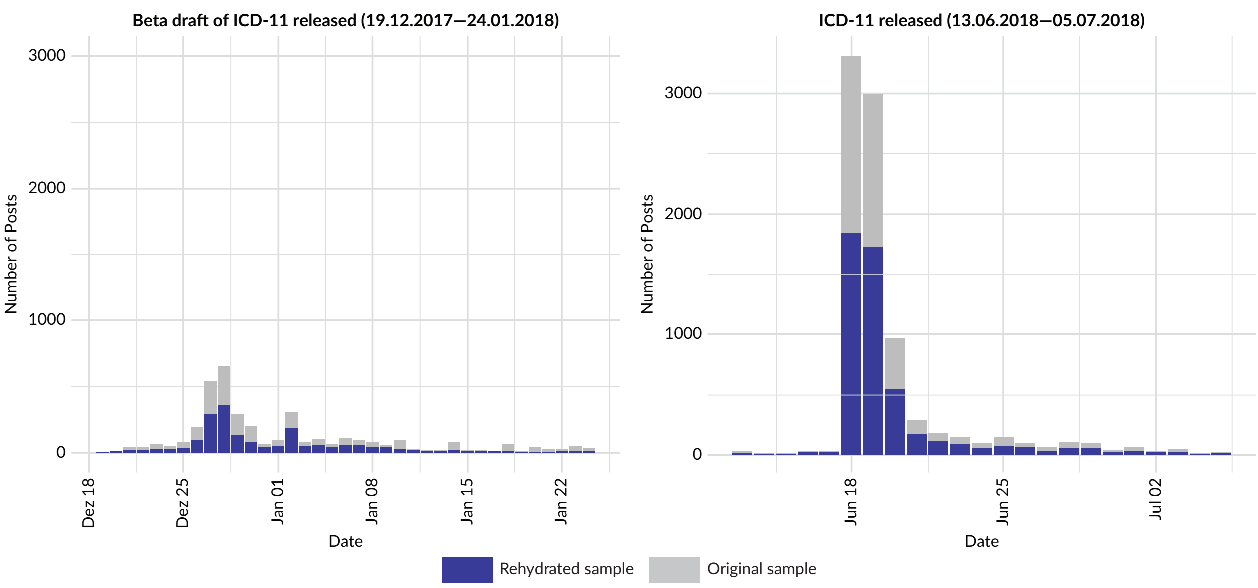

Figure 2. Sample loss over time.

the ICD-11 beta draft release (December 1, 2017–February 1, 2018) and 5,227 during the ICD-11 release (June 1, 2018–July 30, 2018), coinciding with the heated period of the discussion. Most data points were missing from the discussion's central segments on the days the ICD-11 beta draft was discussed (December 26, 2017) and on the day of the introduction (June 18, 2018). In the original sample, 15,845 unique users altogether participated in the discussion. However, in the rehydrated sample, data from only 8,621 of those individuals could be retrieved successfully.

4.2. Changes in the Analysis of Hashtags and Main Actors in the Debate

To better understand the key topics and participants in the debate, Schatto-Eckrodt et al. conducted an initial analysis that focused on influential actors and topics. Table 3, based on the original and rehydrated data, indicates a re-creation of Schatto-Eckrodt et al.'s Table 3, which highlighted top users based on the extent of their participation in the debate. Four of the top 10 participants were suspended due to X terms of service violations in the rehydrated sample. The reasons for these suspensions were not made public. The number of posts in the rehydrated data set stayed relatively stable for the remaining top users. Almost all posts from users who are still on the platform were rehydrated successfully. Most data attrition for the main actors in the debate happens on a user level. Regarding perspectives on the World Health Organization's (WHO) decision, no discernible patterns emerged in the data loss favoring or opposing the decision.

Table 4 indicates systematic differences between the original rehydrated data for the top users based on their reach, as measured as reposts. While the top 10 users with the most reposts remained the same, their ranking order shifted. In the rehydrated data set, significant data loss of approximately 30–40% was found for each user. CNN, the only traditional media outlet in the top 10, exhibited the most prominent reduction in repost support, with a decrease of 40.44%. The loss of repost data appeared to be homogenous across users. Notably, users with the highest reach registered a relative increase in sample size when comparing the original sample, suggesting that data loss disproportionately affects users with a smaller reach.

Table 3. Re-created top users according to their extent of participation.

User	Posts by the user (in the rehydrated sample)	% of all posts (in the rehydrated sample)	Status	# Followers in September 2023	View on the WHO decision	Notes
MommyNooz	32 (0)	0.19 (0)	Suspended	n/a	Supporting	Parenting blog
camerondare	31 (0)	0.18 (0)	Suspended	n/a	Supporting	Activist
AvocateforEd	25 (25)	0.15 (0.27)	Active	29,272	Supporting	Blog
Lynch39083	22 (22)	0.13 (0.24)	Active	43,298	Supporting	Scholar/activist
techedvocate	22 (22)	0.13 (0.24)	Active	22,082	Supporting	Tech blog
eplayuk	16 (16)	0.10 (0.17)	Active	284	Opposing	Gaming blog
Gamescosplay	14 (14)	0.08 (0.15)	Active	627	Opposing	Gaming blog
Gamingthemind	14 (12)	0.08 (0.13)	Active	3,758	Opposing	NGO/activists
Pairsonnalites	13 (0)	0.08 (0)	Suspended	n/a	Opposing	NGO
HealthyWrld	12 (0)	0.07 (0)	Suspended	n/a	Neutral	Health blog

We also replicated Schatto-Eckrodt et al.'s analysis of the most prominent topics in the debate, i.e., the examination of the most salient hashtags in the discourse. Table 5 presents our results. Out of the initial 3,017 hashtags, we successfully rehydrated posts with 1,691 hashtags, illuminating a consistent data erosion

Table 4. Re-created top users according to their reach.

User	Times the user was reposted (of the rehydrated sample)	% of all posts (of the rehydrated sample)	View on the WHO decision	# Followers	Notes
CNN	403 (240)	2.39 (2.59)	Neutral	61,766,081	Media
Deadmau5	385 (257)	2.29 (2.77)	Opposing	3,287,491	Musician
GaijinGoombah	318 (221)	1.89 (2.38)	Opposing	71,334	Youtube content creator
BrendoTGB	274 (153)	1.63 (1.65)	Opposing	294	Regular user
LEGIQN	245 (158)	1.46 (1.70)	Opposing	309,287	Twitch content creator
Pamaj	236 (159)	1.40 (1.72)	Opposing	1,191,142	E-Sports athlete
NoahJ456	233 (158)	1.38 (1.70)	Opposing	1,198,840	Youtube content creator
TheSmithPlays	227 (131)	1.35 (1.41)	Opposing	229,963	Youtube content creator
Boogie2988	182 (114)	1.08 (1.23)	Opposing	482,982	Youtube content creator
CaptainSparklez	170 (112)	1.01 (1.21)	Opposing	4,257,088	Youtube content creator

Table 5. Re-created top hashtags.

Hashtag	Number of occurrences (in the rehydrated sample)	% of hashtag occurrences (in the rehydrated sample)
Gaming	269 (150)	8.92 (8.87)
Gamingdisorder	123 (78)	4.08 (4.61)
Datascience	121 (2)	4.01 (0.12)
Addiction	113 (51)	3.75 (3.02)
Health	102 (50)	3.38 (2.96)
Icd11	97 (69)	3.22 (4.08)
Mentalhealth	97 (76)	3.22 (4.49)
Bbcbreakfast	70 (0)	2.32 (0)
Videogames	61 (38)	2.02 (2.25)
Parenting	59 (8)	1.96 (0.47)
Who	47 (38)	1.56 (2.25)
Gamingaddiction	46 (32)	1.52 (1.89)
Children	40 (5)	1.33 (0.30)
News	36 (19)	1.19 (1.12)
tech	32 (21)	1.06 (1.24)

across the top 15 hashtags. Remarkably, certain hashtags—such as "datascience," "parenting," and "bbcbreakfast"—experienced substantial data loss, reaching 98.35%, 86.44%, and 100%, respectively. This data attrition pattern suggests the absence of entire conversational threads within the discourse related to specific hashtags in the rehydrated data set. Otherwise, data loss occurred systematically across topical hashtags and variations thereof, as well as for individual hashtags most likely associated with journalistic reporting ("#bbcbreakfast" and "#datascience") and parenting ("#parenting"). Even though the omitted hashtags within the top 15 have been substituted by hashtags with a similar thematic tone ("edtech," "games," and "worldhealthorganization"), they represent very different conversation threads. While the rehydrated versions largely maintained unity with the original table, the process of rehydrating revealed that a considerable number of subconversations on the discussion surrounding the inclusion of "gaming disorder" in the ICD-11 led by certain hashtags was lost.

Overall, we found mixed success when replicating the explorative analyses in Schatto-Eckrodt et al. Although many of the replicated tables exhibited high congruence with their counterparts in the initial study, pronounced data loss was found across the rehydrated data set, which effectively affected the analysis of the most prominent actors and topics in the debate. Specifically, we noted that data decay in the context of original posts tends to manifest as binary absence, with all post data per user unattainable.

4.3. Changes in Sentiment Analysis

To investigate the prevailing sentiment within the discourse, the original authors conducted a sentiment analysis with 141,908 sentiment-labeled tokens, revealing an overall negative tone in the discussions. While acknowledging data pre-processing's considerable influence on sentiment analysis outcomes, particularly in the context of X data (Krouska et al., 2016), we replicated the initial sentiment analysis by adhering to the same preprocessing procedures as in Schatto-Eckrodt et al. Our replication sentiment analysis, with 78,661 sentiment tokens, reaffirmed this observation, indicating a consistently negative sentiment. Moreover, upon a detailed examination, we observed a marginal increase in the negativity of sentiments within the rehydrated data set compared with the original sample. Building on the methodology used by Schatto-Eckrodt et al., who examined weekly sentiment fluctuations to follow the discourse's evolving tone over time, our analysis revealed disparities in the absolute weekly sentiment measures between the original and rehydrated samples, reaching statistical significance at the 10% level (t [88] = 1.93, $p < 0.057$, $d = 0.12$). Notably, these differences stemmed from variations in the magnitude of sentiments, rather than the fundamental nature of the sentiments themselves. Furthermore, our analysis did not reveal any statistically significant differences in sentiment between Segment 1 (March 16, 2017–November 30, 2017) and the combined sentiment in Segments 2 (December 1, 2017–June 14, 2018) and 3 (June 15, 2018–November 15, 2018; t [48] = 1.49, $p < 0.1427$, $d = 0.36$), as was found in the original analysis. While we could not completely replicate the original study's results, our replicated sentiment analysis indicated extremely similar patterns observed by Schatto-Eckrodt et al., which is impressive considering the significant data loss of 44.57% in sentiment-labeled tokens.

4.4. Changes in the Topic Model

Finally, we calculated a structured topic model (Roberts et al., 2019) while adhering to the same pre-processing steps that the original authors used. The original sample comprised 5,378 documents, of which we could replicate only 3,266 documents in the rehydrated sample. We used the same rule of thumb as the original authors to choose the number of topics, i.e., the elbow method on the semantic coherence and held-out likelihood. Semantic coherence measures topics' "interpretability." A higher semantic coherence suggests that topics are more interpretable by humans, as terms within a topic are related closely. The measure aids in understanding topics' meaningfulness. Held-out likelihood evaluates the model's ability to predict unseen data. A higher held-out likelihood indicates that the model captured underlying structures and patterns in the text data. Table 6 presents the results from these two measures for a topic model calculated on the rehydrated data set. In our case, a topic model with three topics ($K = 3$) had the highest held-out likelihood and semantic coherence, and it seemed to be the best choice based on the two metrics, although admittedly, the difference in semantic coherence between topic models was small. Notably, a topic model originally chosen by Schatto-Eckrodt et al. (see their Table 6), featuring $K = 5$, proved to be an ill-suited fit for our data. This discrepancy likely arose from the sample size's impact. Larger data sets tend to produce more stable and robust topic models. By reducing the sample size, we introduced more variability in topic assignments, leading to less reliable results in the replicated topic model. Moreover, we found no change in the prevalence of topics during certain phases of the sample. All three topics, highlighted in Table 7, were discussed to an equal extent over the sample period. However, Topics 1 and 2 were discussed considerably more than Topic 3. To sum up, replication of the topic model did not reflect the original study's results, as it arrived at a different optimal number of topics ($K = 3$) than the original study ($K = 5$), and we

Table 6. Topic model metrics for the rehydrated sample.

	$K = 3$	$K = 4$	$K = 5$
Held-out likelihood	−2.260403	−2.27303	−2.271619
Semantic coherence	−91.21908	−97.38662	−105.8436

Table 7. Re-created topic model: Description, top terms, and representative quote of the topics.

Topic	Description	Terms	Quote
1	This topic focuses on discussions of health aspects related to gaming disorder	health, YouTube, condition, addiction, mental, official, disease	"US Health News: If Gaming Addiction Is Now a Mental Health Disorder, How Can We Fight It? #MentalHealth" "Playing too many video games can cause a mental health disorder, says World Health Organization"
2	This topic revolves around issues related to the classification and recognition of gaming addiction	addiction, mental, condition, organization, recognize, classification, health	"WHO to classify 'Gaming Disorder' as a mental health condition in 2018" "Mental Health Experts Warn Against World Health Organization's Definition of 'Gaming Disorder'"
3	This topic involves discussions about the influence of video games, gaming habits, and their impact on individuals and society	video, classification, play, people, time, disease, official	"So, gaming disorder only is a thing if you play games so much that you are physically, mentally, and socially atrophying...like every other disorder built around doing too much of something" "Imo it is an overreaction especially when it's mostly aimed at kids. It's the time in their life for them to be able to play games as much as they do, but now it is gonna be classed as a mental disorder if you play too much"

could not replicate the topics that Schatto-Eckrodt et al. found semantically. Reasons for the latter could very well be the reduced rehydrated sample size.

5. Discussion

The reproduction exercise's results demonstrated a high success rate in reproducing empirical claims from the original study, with 88.46% of the claims successfully reproduced. Challenges arose primarily from issues related to deterministic reproducibility in network illustration methods and a minor error in reporting the correct *t*-statistic. These results emphasize the importance of following proper coding practices and documentation standards in CCS research to maintain research outcomes' integrity, reproducibility, and reliability.

Replicating the original analysis introduced unique challenges, particularly social media data's volatile nature and X API's changing access policies and limitations. The rehydrated sample, collected under strict, new X API data access restrictions, revealed that only 55.08% of the initial sample remains accessible today—a data loss that affected our replication results significantly. Data attrition becomes more pronounced during intense discussion phases and is characterized by a complete presence or absence of data for a user. We have demonstrated that data attrition can be linked to user deletions, post-privacy settings, and account suspensions. Thus, data loss over extended periods—in our case, five years—when dealing with social media data is a consequence of the evolving nature of data on social media platforms. This raises concerns about the long-term sustainability and feasibility of conducting replications based on social media analyses, particularly when working with nonrecent data. Replications may fail for several reasons, such as an initially incorrect finding by the original authors, changes in the measured phenomenon over time, or a lack of generalizability of a finding because it is population-specific (e.g., Dienlin et al., 2021). Our examination revealed that the use of volatile data (e.g., constant changes in user behaviors, account statuses, and data accessibility) also can have a significant impact on studies' replication success.

Our study is subject to several constraints. First, the employed rehydration strategy does not entail drawing a completely new sample. This would be the ideal approach for generalizing Schatto-Eckrodt et al.'s findings. Instead, we reacquired the original data, potentially perpetuating any biases or shortcomings present in the original data set. Another important aspect to note is that X currently manages post edits by deleting the original post ID and assigning a new one (authors' note: Post editing, at the time this was published, was available exclusively to X users who are subscribed to X Premium, a paid service). Consequently, the rehydration method does not capture edited posts, which may have led to a slight overestimation of data loss in our sample. Furthermore, we had to work with X API's technical and financial restrictions, so the sample size in the replication is relatively small, limiting the findings' overall generalizability.

Notably, we replicated an exploratory analysis, which possesses lower evidentiary weight, as its findings are less definitive and more preliminary in nature. Future research should examine how data erosion can affect predictive and inferential analyses. Given that inferential analyses frequently form the foundation for policy recommendations and decision-making, understanding the extent of their vulnerability to data loss in dynamic data environments is important.

Our replication serves as a hypothetical scenario illustrating how a reproduction could have been undertaken in a scenario with no access to the original data. Our replication results indicate that we would not have been able to replicate most of Schatto-Eckrodt et al.'s findings without the data. This inability to replicate—caused predominantly by data fluctuations, rather than improper documentation—underscores data sharing's crucial role in replication efforts. Transparency and accessibility of data and code are almost prerequisites for conducting replications (Marsden & Pingry, 2018). Unfortunately, data sharing is not yet the prevailing practice in CCS, and in some instances, it is legally and technically unfeasible. This problem is worsened by the unpredictability of reacquiring data effectively, as our study has demonstrated. Furthermore, recent increases in financial and access barriers imposed by major social media platform providers have exacerbated this situation. When significant barriers hinder access to materials needed for replication, conducting replication studies becomes more unlikely. In cases in which direct replication is improbable, and conceptual replication is difficult, researchers must hope for regulatory changes in platform access, such as those that the EU's Digital Services Act (DSA) has been promising. This raises a fundamental question about the core value of a scientific analysis when technical and financial obstacles hinder or prevent other researchers from replicating the analysis. The value of a scientific analysis traditionally lies in its potential to contribute to the body of knowledge and to be subjected to testing and validation through replication (National Academies of Sciences, Engineering, and Medicine, 2019). However, when substantial barriers disrupt replication efforts, the value of the analysis itself should be brought into question. In such challenging circumstances, it is imperative for the CCS community to promote practices that mitigate these barriers actively. Encouraging widespread data sharing, advocating for greater transparency, more collaboration between researchers, and establishing standardized protocols for replication attempts can help bridge the gap between original studies and their validation (Dienlin et al., 2021). Moreover, regulatory institutions are crucial in guaranteeing access to social media platforms for CCS researchers at a time when social media platform APIs are threatening open scientific endeavors (Davidson et al., 2023). Through implementation and enforcement of policies that foster scientific access, these regulatory bodies not only can support CCS research, but can also help overcome barriers to replicability. Addressing these barriers is essential to upholding the core tenets of scientific inquiry and safeguarding the integrity of valuable research through rigorous examination.

6. Conclusion

This study highlights the importance of replication in enhancing the validity and reliability of CCS research. Our replication and computational reproduction efforts provide insights into the challenges and complexities of conducting and replicating CCS research, particularly in the context of data access and third-party platform providers. Our results emphasize the need for researchers to consider data loss and changes in data availability over time when conducting replication and reproduction studies in the CCS field, particularly when working with social media data.

While the practical recommendations drawn from the present study, urging researchers to prioritize reproducibility in the realm of volatile data access, can enhance CCS studies' overall quality and robustness, recognizing that the individual researcher is just one component within a broader framework is crucial: Regulatory bodies, social media platforms and their users, and journals and academic institutions also wield significant influence in this context. Thus, replicability in CCS is also a political and institutional issue. For example, enhancing CCS replicability can be achieved by providing researchers with increased access and control over their data, particularly in the realm of social media platform research. The DSA promises to play a central role in this context, as it offers, among many other benefits to researchers, clear regulations concerning access to data of "very large online platforms," i.e., platforms with more than 45 million monthly active EU users. The DSA could establish the foundation for transparent social media and platform research, aiming to ensure that "socially relevant aspects of digitization can be investigated appropriately, consistently, and independently" (Klinger & Ohme, 2023, p. 3). Under Article 40 of the DSA, researchers affiliated with a research institution and independent of commercial interests will be able to request data from these platforms through a Digital Services Coordinator to conduct research on systemic risks in the EU. Systemic risks, as outlined in the DSA under Article 34(1), include negative effects on civic discourse and electoral processes, protecting public health, and dissemination of illegal content (European Centre for Algorithmic Transparency, 2023). A comprehensive scientific examination of digitization's socially relevant aspects necessitates researchers' ability to reproduce and replicate their findings within this context. Without strong supranational regulations, such as the DSA, many researchers' individual efforts to improve their work's reproducibility and replicability would be conducted in vain. In addition to regulatory actions' effects, the users who create the data being researched also should be considered: Analyzing content that users have deleted also poses ethical research questions that must be considered case by case. Examining actively harmful or anti-democratic entities' actions may justify the analysis of data deleted by users, whereas, in other instances, researchers may need to respect individual users' decisions to withdraw their content from public access. Contemplations on CCS research replicability should extend beyond individual researchers' efforts and include all pertinent stakeholders in the evaluation process. Thus, recognizing that incentive structures established by academic journals and publishers, alongside those of academic institutions, play a pivotal role in advancing the broader goal of enhancing research replicability is crucial. This includes incorporating systematic code and data review as integral components of the peer-review process, establishing infrastructure for responsible sharing of code and data in compliance with data privacy and access regulations, and allocating funding to bolster scientific results' robustness and reproducibility.

Increasing access to data is paramount in addressing reproducibility and replication challenges in CCS. Researchers' ability to access and analyze data directly impacts replication efforts' feasibility and robustness. Policies and regulatory changes, such as the DSA, play a pivotal role in facilitating such access, as they require platforms to provide data access to researchers, particularly concerning socially relevant digitization aspects. Improving data accessibility stands as a crucial measure in tackling reproducibility and replication hurdles in CCS. Moreover, overcoming barriers related to data sharing, transparency, and technical constraints is important to preserving the value of scientific analysis and promoting robust replication practices in CCS. We have demonstrated that reacquiring platform data is both resource-intensive and comes with no guarantee of fully regaining the primary sample. Thus, our study highlights vital aspects in establishing sustainable practices for reproducibility and replicability in CCS, such as data accessibility, data sharing, transparency, and comprehensive documentation of research methods and materials.

Funding

This research was supported by the German Research Foundation priority program META-REP: A Meta-Scientific Programme to Analyse and Optimise Replicability in the Behavioural, Social, and Cognitive Sciences (project number 441890184), with the subproject What Defines and Affects Replicability in Computational Communication Science? (project number 464291459).

Conflict of Interests

All reproduction analyses and the evaluation thereof were done by Philipp Knöpfle. The authors declare no conflict of interests.

Data Availability

The replication package for this study, containing all shareable data, code, and materials, is available at: https://osf.io/2jb9m

Supplementary Material

All supplementary material such as replication materials, software bibliography, reproduction protocol, etc. can be found in the OSF repository (https://osf.io/2jb9m).

References

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets are forever: A large-scale quantitative analysis of deleted tweets. In A. Bruckman & S. Counts (Eds.), *CSCW '13: Proceedings of the 2013 conference on computer supported cooperative work* (pp. 897–908). ACM. https://doi.org/10/gfzgsd

Alston, J. M., & Rick, J. A. (2021). A beginner's guide to conducting reproducible research. *The Bulletin of the Ecological Society of America*, *102*(2), Article e01801. https://doi.org/10.1002/bes2.1801

Benoit, W. L., & Holbert, R. L. (2008). Empirical intersections in communication research: Replication, multiple quantitative methods, and bridging the quantitative–qualitative divide. *Journal of Communication*, *58*(4), 615–628. https://doi.org/10.1111/j.1460-2466.2008.00404.x

Bhattacharya, P., & Ganguly, N. (2021). Characterizing deleted tweets and their authors. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 547–550. https://doi.org/10.1609/icwsm.v10i1.14803

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566, https://doi.org/10.1080/1369118X.2019.1637447

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980. https://doi.org/10.1257/jel.20171350

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y. L., Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., & Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., Van Der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*, 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., . . . De Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26.

Dreber, A., & Johannesson, M. (2023). *A framework for evaluating reproducibility and replicability in economics*. SSRN. http://doi.org/10.2139/ssrn.4458153

European Centre for Algorithmic Transparency. (2023). *FAQs: DSA data access for researchers*. https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en

Freiling, I., Krause, N. M., Scheufele, D. A., & Chen, K. (2021). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*, *71*(5), 686–714. https://doi.org/10.1093/joc/jqab032

Haim, M. (2023). *Computational communication science: Eine Einführung*. Springer. https://doi.org/10.1007/978-3-658-40171-9

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8), 180448. https://doi.org/10.1098/rsos.180448

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239.

Klinger, U., & Ohme, J. (2023). *What the scientific community needs from data access under Art. 40 DSA: 20 Points on infrastructures, participation, transparency, and funding* (Weizenbaum Policy Paper No. 8). Weizenbaum Institute for the Networked Society; The German Internet Institute. https://doi.org/10.34669/WI.WPP/8.2

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on Twitter sentiment analysis. In N. Bourbakis, G. Tsihrintzis, M. Virvou, & D. Kavraki (Eds.), *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (p. 144). IEEE. https://doi.org/10.1109/IISA.2016.7785373

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *New England Journal of Medicine*, *374*(3), 276–277. https://doi.org/10.1056/NEJMe1516564

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems*, *115*, A1–A7. https://doi.org/10.1016/j.dss.2018.10.007

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, Article 0021. https://doi.org/10.1038/s41562-016-0021

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. National Academies Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. http://doi.org/10.1177/1745691612459058

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Peng, R. D., & Hicks, S. C. (2021). Reproducible research: A retrospective. *Annual Review of Public Health*, *42*(1), 79–93. https://doi.org/10.1146/annurev-publhealth-012420-105110

Petrovic, S., Osborne, M., & Lavrenko, V. (2013). *I wish i didn't say that! Analyzing and predicting deleted messages in Twitter*. arXiv. http://arxiv.org/abs/1305.3107

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Rosenberg, D. E., Filion, Y., Teasley, R., Sandoval-Solis, S., Hecht, J. S., van Zyl, J. E., McMahon, G. F., Horsburgh, J. S., Kasprzyk, J. R., & Tarboton, D. G. (2020). The next frontier: Making research more reproducible. *Journal of Water Resources Planning and Management*, *146*(6), Article 1820002. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001215

Rosenthal, R. (1991). Replication in behavioral research. *Journal of Social Behavior and Personality*, *4*(4), 1–30.

Schatto-Eckrodt, T. (2022). Hidden biases—The effects of unavailable content on Twitter on sampling quality. In J. Jünger, U. Gochermann, C. Peter, & M. Bachl (Eds.), *Grenzen, Probleme und Lösungen bei der Stichprobenziehung* (pp. 178–195). Herbert von Halem Verlag.

Schatto-Eckrodt, T., Janzik, R., Reer, F., Boberg, S., & Quandt, T. (2020). A computational approach to analyzing the Twitter debate on gaming disorder. *Media and Communication*, *8*(3), 205–218. https://doi.org/10.17645/mac.v8i3.3128

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing reproducible research*. CRC Press.

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2/3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Computational communication science: Toward open computational communication science—A practical road map for reusable data and code. *International Journal of Communication*, *13*, 3935–3954. https://ijoc.org/index.php/ijoc/article/view/10631

Wei, W., Joseph, K., Liu, H., & Carley, K. M. (2016). Exploring characteristics of suspended users and network stability on Twitter. *Social Network Analysis and Mining*, *6*(1), 1–18. https://doi.org/10/gc92c9

Zhou, L., Wang, W., & Chen, K. (2016). Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In J. Bourdeau, J. A. Hendler, R. N. Nkambou (Eds.), *WWW '16: Proceedings of the 25th International Conference on World Wide Web* (pp. 603–612). ACM. https://doi.org/10/gf4hpq

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: Bioinformatics and beyond. *Briefings in Bioinformatics*, *24*(6), Article bbad375. https://doi.org/10.1093/bib/bbad375

About the Authors

Philipp Knöpfle is a research associate at the Department of Media and Communication at Ludwig-Maximilian University of Munich (LMU). He obtained his MSc degree in Finance and Information Management from the Technical University of Munich, University of Augsburg, and University of Bayreuth. Currently, he is working at the Computational Communication Research unit under the supervision of Professor Haim at LMU. His primary research interests encompass meta science, open science, and computational methods.

Tim Schatto-Eckrodt is a research associate at the Department of Journalism Studies and Communication at the University of Hamburg, Germany. He holds an MA degree in Communication Studies from the University of Münster. He is currently working as the chair for Digital Communication and Sustainability at Hamburg University. His further research interests include computational methods and online propaganda.