Political Analysis (2024), **32**, 493–506
doi:10.1017/pan.2024.7

ARTICLE

Nonrandom Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies

Andreas Küpfer

Institute for Political Science, Technical University of Darmstadt, Darmstadt, 64283 Hesse, Germany

Corresponding author: Andreas Küpfer; Email: andreas.kuepfer@tu-darmstadt.de

(Received 15 September 2023; revised 1 February 2024; accepted 06 February 2024; published online 17 May 2024)

Abstract

Used by politicians, journalists, and citizens, Twitter has been the most important social media platform to investigate political phenomena such as hate speech, polarization, or terrorism for over a decade. A high proportion of Twitter studies of emotionally charged or controversial content limit their ability to replicate findings due to incomplete Twitter-related replication data and the inability to recrawl their datasets entirely. This paper shows that these Twitter studies and their findings are considerably affected by nonrandom tweet mortality and data access restrictions imposed by the platform. While sensitive datasets suffer a notably higher removal rate than nonsensitive datasets, attempting to replicate key findings of Kim's (2023, *Political Science Research and Methods* 11, 673–695) influential study on the content of violent tweets leads to significantly different results. The results highlight that access to complete replication data is particularly important in light of dynamically changing social media research conditions. Thus, the study raises concerns and potential solutions about the broader implications of nonrandom tweet mortality for future social media research on Twitter and similar platforms.

Keywords: text-as-data; twitter; replication

Edited by: Jeff Gill

1. Introduction

Researchers use Twitter[1] data to explain a broad array of political phenomena. A substantial share of these political science studies involves the analysis of tweets that may contain subjects like violence, racism, or other controversial content (e.g., Keller *et al.* 2020; Kim 2023; Mitts 2019), which I refer to as sensitive content. The replication[2] of findings based on sensitive content is hampered by Twitter's policy that prohibits sharing tweets instead of tweet IDs only and the resulting inability to crawl tweets that have been removed from the platform.[3] This becomes particularly problematic for sensitive content as these tweets lead to potential bias due to nonrandom patterns of tweet removal.

Why should researchers take a deeper look at these nonrandom removal patterns? Social science research relies on replicable datasets as the recent replication crisis in social sciences underlines (e.g.,

[1]Twitter was renamed *X* in July 2023.

[2]I refer to replication as "using the same methods on different data produces comparable results" (Davidson *et al.* 2023).

[3]It is important to note that beside replicability also the validity of certain Twitter studies can be affected by these strict policies. For example, Frimer *et al.* (2023) crawl and analyze tweets created more than 10 years ago. The resulting dataset potentially includes fewer old tweets and higher availability of more recent tweets due to nonrandom tweet mortality. This might affect the findings of papers' studying historical Twitter content.

© The Author(s), 2024. Published by Cambridge University Press on behalf of The Society for Political Methodology.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

Dreber and Johannesson 2019; Key 2016; King 2003; Laitin and Reich 2017). The discipline can confidently build upon and trust findings only if platforms like Twitter offer a representative, stable, and end-to-end replicable data source. The ability to fulfill these requirements may be hampered by the platform's limitations: it prohibits crawling removed tweets and restricts publishing them along with academic papers.

Existing insightful studies on how tweets are removed are based on rather general datasets focusing on random or issue-related samples. Some find no alarming patterns for replicability (Pfeffer *et al.* 2023; Zubiaga 2018). However, recent research on datasets yielded from the 1% Streaming Twitter API shows that emotionally charged or potentially controversial datasets behave differently than nonsensitive datasets (Elmas 2023). As sensitive datasets belong to very frequently studied Twitter content by political scientists, it is crucial to elaborate on how the removals of tweets impact research findings and datasets.

To investigate potential nonrandom removal patterns of tweets and how these affect replicating journal articles, I first conduct a systematic study of Twitter papers published in seven top political science journals. A high share of papers are based on sensitive content, and political scientists need a unified way to share their Twitter replication data. Recrawling the content of both nonsensitive and sensitive datasets implies that tweets belonging to the latter category are removed at a noticeably higher rate. To show the impact of these nonrandom removal patterns on sensitive dataset findings, I attempt to replicate central findings reported in a recent *Political Science Research and Methods* article by Kim (2023). The availability of only less than 20% compared to the original number of tweets suggests that such an incomplete sensitive dataset compromises both descriptive and statistical findings. To understand why tweets become unavailable, it helps aggregating whether the platform or the user is in charge of these tweet removals. The platform is largely responsible for over half of all tweet removal decisions in the case study dataset. However, the other half originates from direct user actions: Users can remove an individual tweet, protect their account to make tweets only visible to their followers, or deactivate their account.

Especially when using social media data, researchers should focus on two important questions: What are the reasons that previously available observations might become unavailable later, and what are the implications for replicating studies that rely on them?

This paper first emphasizes the high relevance of Twitter research to political science, particularly regarding sensitive datasets. Second, it raises awareness of how Twitter hampers replicable research and how this affects actual research findings. Disentangling underlying mechanisms of this data foundation allows for a more critical and evidence-driven process when deciding which data sources to leverage in political science studies. The article contributes to both existing threads of literature by taking a rather practical-oriented point of view, which is particularly valuable for scientists studying social media platforms. In light of the dynamic changes in these platforms, I draw attention to the challenges of replicating social media studies. The paper formulates potential solutions for accessing social media data in the post-API era to tackle these challenges, giving a perspective for making future social media research replicable.

2. From Replication Crisis to the Persistence of Twitter Data

Publishing replicable research is a fundamental pillar of science. Authors, as well as journals within political science and beyond, continuously work on the revision of policy standards, adding the replication of data and code to publications (Key 2016; King 1995, 2003; Laitin and Reich 2017). However, while these revisions address the ongoing replication crisis in the social sciences, they cannot solve it. I argue that a major reason is the need for more knowledge and awareness about datasets researchers use in their studies.

While code availability is essential to replicate findings, underlying data forms the deepest research layer. Diverse data sources, like surveys, experiments, and social media, can be subject to biases, errors, and methodological issues. This means that researchers must make complex decisions and assumptions influencing the data collection process. In the worst case, these decisions lead to inconsistent results

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

due to incomplete replication data. The interaction between authors and journals is one opportunity to elaborate ways of circumventing replication issues (Laitin and Reich 2017). However, especially proprietary datasets, that is, limited access to original data and important ethical data privacy concerns further complicate the replication process.

Commercial social media platforms—Twitter in particular—are prominent drivers for studies leading to proprietary datasets. While there are many platforms, 39.70% of social media researchers use Twitter as a data source for their projects (Hemphill, Hedstrom, and Leonard 2021). The frequent use of Twitter data by social scientists is related to the platform, presenting an ideal combination of size, international reach, and—compared with other social media platforms—good data accessibility making it the preferred platform for social media research (Steinert-Threlkeld 2018). Another aspect is that in 2020 Twitter rebuilt its API (Twitter 2020) to allow access to its full tweet archive for academic purposes, which, however, got suspended in its known form in June 2023.

On the one hand, researchers tremendously benefited from the suspended API, and developing solutions that allow researchers to continue working on Twitter studies is important as data from this platform is part of much insightful research. On the other hand, the platform's policies bound scientists. The major technical limitation is the inability to crawl removed tweets by their unique identifier. Researchers cannot replicate findings based on the complete set of tweets as it is only allowed to publish the ID of a tweet but not its textual content—leading to unavailable tweets when trying to recrawl tweet IDs.[4]

The attrition rate as an established metric for unavailable tweets helps to understand the process of and its impact on representational aspects of a dataset (Almuhimedi *et al.* 2013; Elmas 2023; Hai and Fu 2015; Noonan 2022; Pfeffer *et al.* 2023; Zubiaga 2018). While studies are analyzing the attrition rate, unfortunately, many of the datasets studied represent Twitter as a whole but do not distinguish between specific issue domains and sentiment types of tweets that are of high interest in political science. Other work on different issue domains focuses on rather general keyword-generated datasets between 2012 and 2016. Recollected datasets are still representative to a large extent in terms of their textual content but are not stable on metadata (Zubiaga 2018). Metadata involves further descriptive information about a tweet or user, such as the number of likes or retweets. However, as metadata can be published without violating the policies of Twitter, at least this aspect should play only a minor role in replication issues.

Previous studies argue that even though the recrawling ratio of tweets may drop below 70%, the content of tweets in their datasets is still representative. However, looking at the sentiment of tweets might explain the underlying mechanism of tweet removals more comprehensively. This is important as sentiment and other latent text features are crucial for many projects. Recrawled controversial datasets show considerable differences from the original ones in various metrics relevant to political scientists (Elmas 2023). These include shifts in political orientation, trending topics, and harmful content. The difference between the share of collectible tweets at a later time and the original dataset is even larger for controversial datasets in particular. The reason that a tweet in a sensitive dataset is not available for recollection anymore is mainly due to account and tweet suspensions initiated by Twitter itself (e.g., due to violating policies) which holds specifically for controversial datasets (Elmas 2023). These indicators for sensitive datasets suggest that one has to assume nonrandom removal patterns leading to incomplete replication datasets and, thus, inconsistent findings.

An emerging body of research examines extreme sentiment expressed in tweets (e.g., Alrababah *et al.* 2021; Kim 2023; Muchlinski *et al.* 2021). However, it remains unclear how Twitter researchers address the subsequent issue of replicability in their replication archives. Furthermore, no prior studies have investigated the implications of replicating the findings of published political science studies and real-world datasets that focus on sensitive content. It is necessary to measure tweet attrition more fine-grained when judging replicability of studies containing sensitive datasets.

[4]While Twitter's policy previously allowed sharing the content of up to 50,000 tweets per day, they recently decreased this number to 500. Most importantly, all tweets in the shared dataset must still be available on the platform, leading to an incomplete dataset.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

3. Tweet Sharing and Mortality in Political Science Studies

How do researchers share Twitter data? In this section, I outline both how researchers share their Twitter datasets and examine nonrandom deletion patterns dependent on whether a dataset is sensitive or nonsensitive.

3.1. How the Discipline Shares Tweets

In some cases, researchers may be allowed to release the entire dataset (e.g., Twitter itself offers a selection of publicly available datasets), but in others, restrictions imposed by national laws and social media platforms—such as the right to be forgotten—try to prevent this. In the light of this, researchers handle the data-sharing process in various ways. Moreover, different requirements, replication policies, university restrictions through the Institutional Review Board, and journal integrity checks lead to manifold decisions during the data-sharing process.

I conduct an empirical analysis crawling all 151 papers that mention the keyword "Twitter" published between January 2015 and September 2022 in seven major political science journals *AJPS*, *APSR*, *BJPS*, *JOP*, *PA*, *PolComm*, and *PSRM*.[5] I keep only those that systematically analyze the content of tweets, as the textual content is the most problematic part of a typical Twitter dataset to share. Finally, I annotate the remaining dataset of 50 papers with additional information on the topic of the Twitter dataset.

Of these papers, 30.00% study sensitive Twitter content.[6] Figure 1 shows that in general, less than half of all papers publish either tweet IDs, the content of the tweets, or both. A proportion of 20.00% of the replication archives contains tweet IDs only, which I assume, in many cases, might be insufficient for successful end-to-end replication. Furthermore, a high percentage of papers (60.00%) share neither tweet IDs nor content, which makes a replication impossible. Surprisingly, almost a fourth share the raw textual content of tweets which technically would violate Twitter policies but is beneficial for the end-to-end replicability of Twitter research. However, this is the only way of replicating Twitter studies without paying the current fees for using the Twitter API and recrawling still available tweets from their IDs.

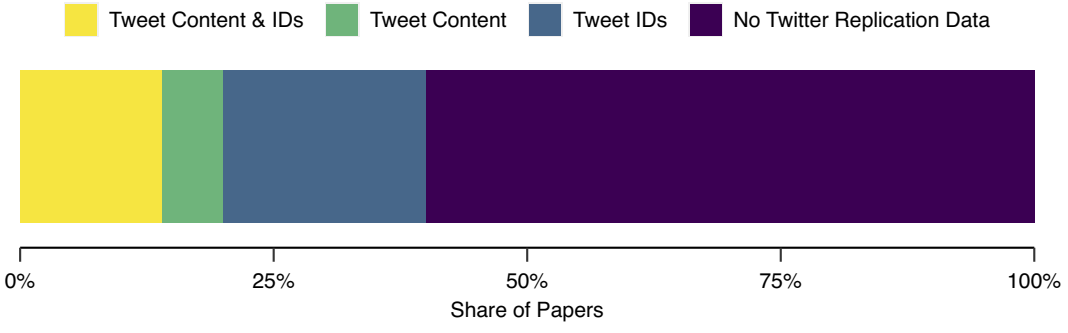


Figure 1. Different ways of how political scientists share Twitter datasets in replication archives among all 50 papers analyzing the content of tweets in seven major political science journals.

[5]To crawl these papers, I use Google Scholar. The selection of journals covers a broad range of high-impact journals, from substantial and methodological research to the field of political communication. By that, it gives a thorough overview of Twitter research in political science.

[6]My definition of sensitive content is partly derived from Elmas (2023) and includes papers explicitly studying the content of tweets in datasets containing fake news/disinformation, hate speech/violence/terrorism, or bots. This sensitive area of study collectively emphasizes the dynamics and impact of harmful online behaviors and their propagation through social media platforms. The Supplementary Material details my annotation approach and provides an overview of all papers.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

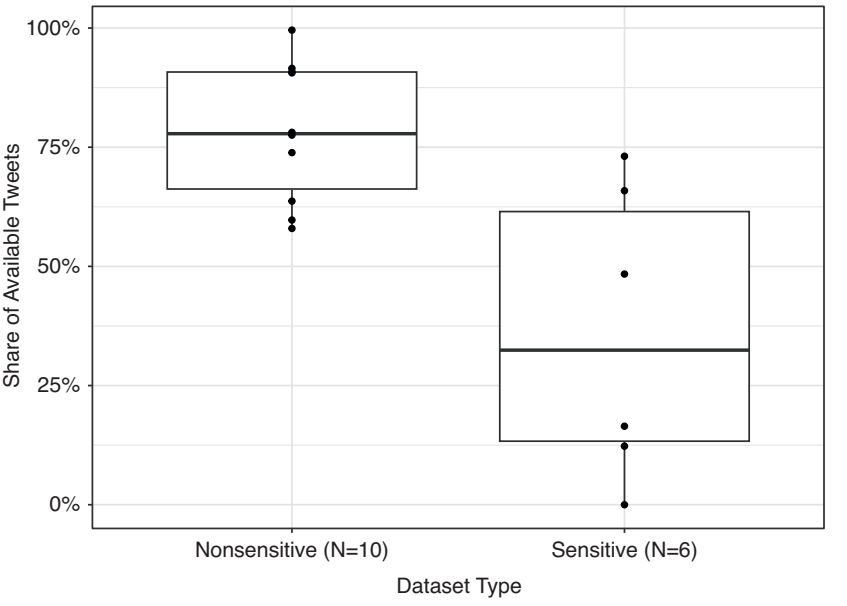


Figure 2. Availability rate of a random sample of up to 10,000 tweets from each of the 16 sensitive and nonsensitive paper datasets which shared at least their tweet IDs. Retrieving all tweets was attempted in cases where the original dataset contained fewer than 10,000 tweets. Temporão *et al.* (2018) shared user IDs instead of tweet IDs, as the authors crawled all users' tweets. Thus, I checked the availability of these user accounts instead of tweets. Data was recrawled on May 17, 2023.

3.2. Nonrandom Deletion Patterns of Sensitive Datasets

Previous studies show that one should expect differences in the availability of tweets when looking at sensitive and nonsensitive datasets in isolation. The overall substantial share of 30.00% of sensitive Twitter datasets suggests that there are enough replication datasets to study available tweets in both dataset types. To analyze the decay of tweets dependent on the dataset type, I can rely on the fraction of replication archives sharing at least their tweet IDs. The literature overview results in 16 papers[7] sharing tweet IDs. Ten of these papers work with nonsensitive datasets, representing 28.57% of all datasets annotated as nonsensitive. In contrast, six papers utilize sensitive datasets, comprising 40.00% of all datasets classified as sensitive.

Figure 2 depicts the proportion of accessible tweets of these papers.[8] Indeed, the descriptive analysis shows clear differences between both types of tweets. In a random sample of 10,000 tweets per dataset, an average of 78.34% of nonsensitive dataset tweets remains accessible, starkly contrasting to only 36.02% in sensitive datasets.[9] Within Twitter replication datasets, it appears that datasets marked as sensitive have a higher chance of mortality.

Relying on a data basis of more than three quarters of still available tweets in nonsensitive datasets sounds convincing to initiate a replication attempt. However, replicating studies could become challenging with only a third of the original tweets retrievable in sensitive datasets and without knowledge about the decision-making process of those removing the data. It is important to note that this issue is not confined solely to sensitive datasets: many nonsensitive datasets also include sensitive tweets.

[7] One more paper by Brie and Dufresne (2020) shares tweet IDs that, however, are corrupted and cannot be further used for rehydrating tweets.

[8] The datasets were crawled on May 17, 2023, using the official Twitter Academic Research Track API accessed via the R package `academictwitteR` (Barrie and Ho 2021).

[9] It is essential to note that while older datasets may suffer from an increased mortality rate, the age of a dataset is no systematic bias in this sample as the average year of publication among the recrawled datasets is 2020 for both sensitive and nonsensitive datasets.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

Recrawled versions of these published datasets might also generate bias to a certain extent, as sensitive tweets are more likely to be removed—and thus become unavailable to researchers.

While the retrieval rate of tweets from an MP or voters' opinions on policies on Twitter appears to be closer to the original population (i.e., nonsensitive dataset), hate speech or extreme ideological datasets endure a significant loss in tweets (i.e., sensitive dataset). This bias must be highlighted as it is critical for replication. A high share of tweet removals is not explicitly caused by the authors of tweets (Almuhimedi *et al.* 2013). Letting users report tweets and accounts certainly impacts the platform's decision to remove them. However, Twitter's content moderation takes the final decision on whether to remove tweets and suspend accounts or keep them on Twitter (Alizadeh *et al.* 2022; Pierri, Luceri, and Ferrara 2022). In result, the platform introduces a nontransparent layer of nonrandom tweet mortality directly impacting the data basis.

4. Case Study: Implications for Replicating Sensitive Twitter Studies

Sensitive datasets suffer from a notably higher loss of tweets than nonsensitive datasets, affecting replication. Kim (2023) is one example of a study that works with a sensitive dataset. The paper demonstrates how violent tweets surrounding the 2020 U.S. Presidential election reflect the real world and spotlights the groups targeted by violent content.[10]

There are several reasons for considering this study for replication. Among all sensitive Twitter studies sharing tweet IDs, this paper does not only analyze the well-researched U.S. election in 2020 on social media but combines three methodological and datawise characteristics well-suited for an insightful replication. First, it studies violent tweets and compares them with nonviolent ones, which supports analyzing differences between the behavior across both dataset types. Second, it studies rather aggregated data and provides a longitudinal perspective. Third, the replication archive offers much data beyond tweet IDs (e.g., document-frequency matrices or hashtag frequencies), which supports comparing the original findings with the replication. Other potential replication candidates reflect these selection criteria only partially.[11]

The study's initial population of more than 300 million tweet IDs processed in a data collection pipeline is not publicly available. However, the replication archive allows access to all tweet IDs classified by the article's deep-learning algorithm as containing violent content.[12] This set of IDs ranges from September 23, 2020 to January 8, 2021 and consists of 215,923 unique tweet IDs. As of November 15, 2022, there are only 35,552 (16.47%) of the original number of tweets retrievable via the API.[13] The reported values are even lower than the numbers for other controversial datasets (Elmas 2023), thus underlining the evidence that sensitive tweet removals are not random.

What are the learnings from unavailable tweets and their authors? Twitter's reasons for unavailable tweets are manifold. The compliance endpoint of the Twitter V2 API (Twitter 2021) helps examine them based on users tweeting violent content.[14] Over half of the tweets (52.90%) in the dataset are removed

[10] While the paper contains important substantial findings, I focus on the methodological aspects of replicating its findings. Moreover, as sharing the content of tweets is problematic, I mainly replicate findings that involve the direct analysis of tweets and, thus, do not consider replicating the study of user networks in Chapters 4.3–4.5 in the original paper. More details on substantial aspects and the methodology of creating the original dataset are described in the Supplementary Material.

[11] See Appendix 2 of the Supplementary Material for detailed information on other replication candidates.

[12] Efforts were made to obtain the original data from the author. Unfortunately, they remained unsuccessful.

[13] After that, I recrawled the dataset three more times on November 25, 2022, December 15, 2022, and March 30, 2023, expecting an increasing retrieval rate due to the takeover by Elon Musk and the reinstation of Donald Trump's Twitter account. However, there are no significant shifts in the rate of retrievable tweets. Kim (2023) is aware of shrinking retrieval rates in general, too, which is reflected in the following statement in a ReadMe file in his replication archive: "Note that, as the tweets included in the dataset are highly likely to violate Twitter's rules [. . .], some of the tweets might have already been taken down or the related account might have been suspended."

[14] Although I do have both, the tweet IDs and a list of user IDs who tweeted violent content, I do not know which tweet ID belongs to a certain user which is necessary for the compliance endpoint when working with tweet IDs. Hence, I rely on retrieving the compliance status of all user IDs weighted by their total number of tweets to get an estimate for removal reasons.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

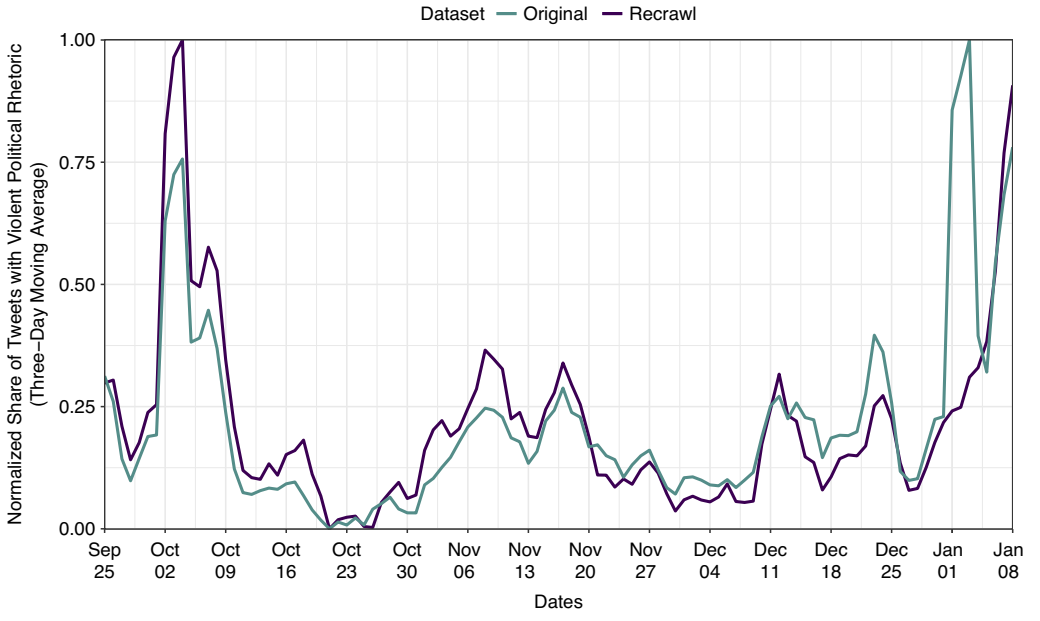


Figure 3. Timeline comparison of normalized proportion of violent political rhetoric tweets during the U.S. election 2020 for both the original and recrawled datasets (replication of Kim 2023). Proportions are based on aggregated information on 215,923 original and 35,552 recrawled tweets.

due to user suspensions. It is important to note that these decisions are taken by Twitter, for example, their systematic content moderation based on controversial trends or hashtags, or user reports of a particular tweet or account. Actions originating on the user side—deleted, protected, or deactivated accounts—are responsible for the remaining unavailable tweets. The Supplementary Material depicts detailed proportions in Figure A.2.

Compared with the original data, essential aspects of the recrawled data are no longer representative. Even without access to the full data, I can still rely on a random subset of 5,000 violent tweets aggregated in a document-frequency matrix openly distributed by the author. I approach the representativity of different textual features compared with an equally sized random sample of the recrawled violent tweets using Welch's *t*-test (Zubiaga 2018). The basis for the analysis is the word frequencies independently generated from both samples. The *t*-test results show that the 95% confidence intervals for textual content and hashtags do not contain zero, indicating that these features are different in both datasets.[15] This is not the case for user mentions that seem representative based on the random sample. However, this does not ensure that findings related to specific groups of user mentions remain unaffected by replication issues. The metric only looks at the frequency of all user mentions in both datasets and by that, gives an overall picture, potentially overlooking group-specific dynamics.

4.1. Replication: Descriptive Analysis

Describing social media datasets frequently involves looking at how data change over time. Figure 3 (based on Figure 3 in Kim 2023) shows peaks of tweet counts containing violent political rhetoric over time in the original dataset (teal line) and the recrawl (purple line). It becomes clear that the curve does not behave as expected when assuming a random removal of tweets. This is especially highlighted through early January 2021 during the power transition after the election and when the Capitol Riot

[15] The exact numbers are in Table A.2 in the Supplementary Material.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

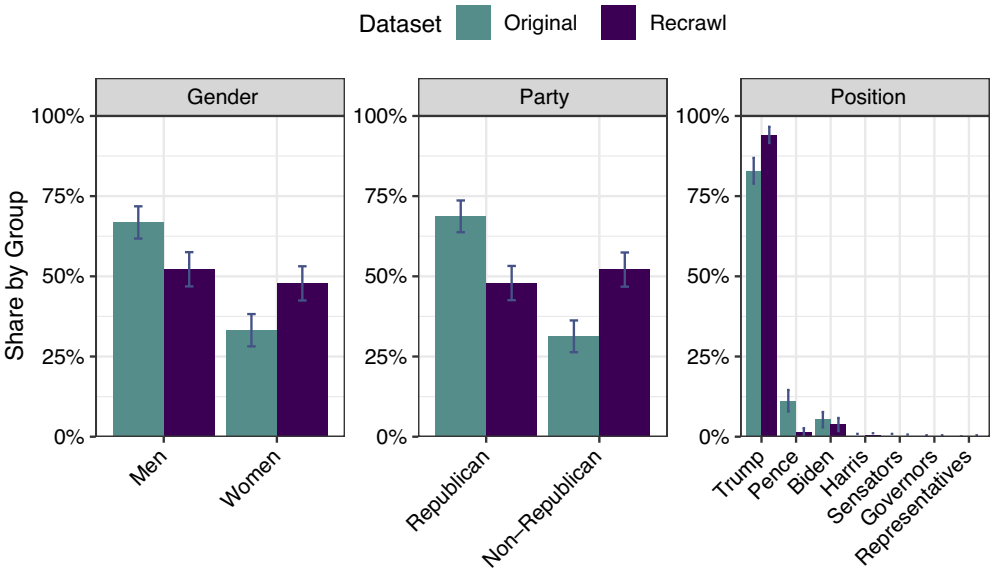


Figure 4. Comparative distribution of mean mentions of accounts in tweets containing violent political rhetoric by gender, party, and position in the original and recrawled datasets. Uncertainty displays the 95% confidence interval of each group. Proportions are based on aggregated information on 215,923 original and 35,552 recrawled tweets.

happened. While one of the key findings of the author is to demonstrate that offline events are mirrored on social media, the recrawled data behave differently and fail to mirror the original data in their most important aspects. Accordingly, nonrandom tweet removals hamper the longitudinal representation of the dataset and the findings based on it.

Hashtags are a core feature on Twitter and are vital to spreading ideas and sparking conversations. Therefore, it is crucial to also examine Kim's study of frequent hashtags. Reusing Table 2 in the original paper published with hashtag frequencies, I retrieve the original counts of hashtags. As one would expect, all counts are much lower in the recrawled dataset than in the original one. However, the sorting of hashtags also differs clearly between the datasets, which is just another visual argument that the retrieved tweets do not represent the same distribution of hashtags as the original dataset. Most importantly, the top three most frequent hashtags in the original dataset (*#wethepeople*, *#1*, and *#pencecard*) are either absent from the revised top ranking or are indistinguishable from other hashtags due to their low count. Table A.3 in the Supplementary Material reports the usage of hashtags in violent tweets during the complete election period.

While hashtags show discrepancies when being recrawled, how are different groups represented in the recrawled dataset? Discussions on Twitter occur between different actors. Utilizing the actor's characteristics allows assigning them to groups. Only if the distribution across these groups remains consistent during replication, should further analysis consider their utilization. In Table 3 in the original paper, Kim (2023) summarizes the count of account mentions in violent political rhetoric and nonviolent tweets into three groups (Gender, party, and position). Reusing the author's proposed group assignment allows for calculating the recrawled dataset's proportions. Both proportions are depicted in Figure 4. The most important aspect is not necessarily the raw numbers but the proportions within the grouping characteristics. The original dataset has a disproportionate distribution of political party (69% Republican, 31% non-Republican) and gender (33% Women, 67% Men). However, in the recrawled set of tweets, party and gender are distributed evenly. While Trump remains the leader without substantial variation in the position group, the proportion of Pence-related user mentions shrinks close to zero. The findings demonstrate nonrandom removal patterns where tweets referring to women and Republicans are more likely to be removed than those referencing men and non-Republicans.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

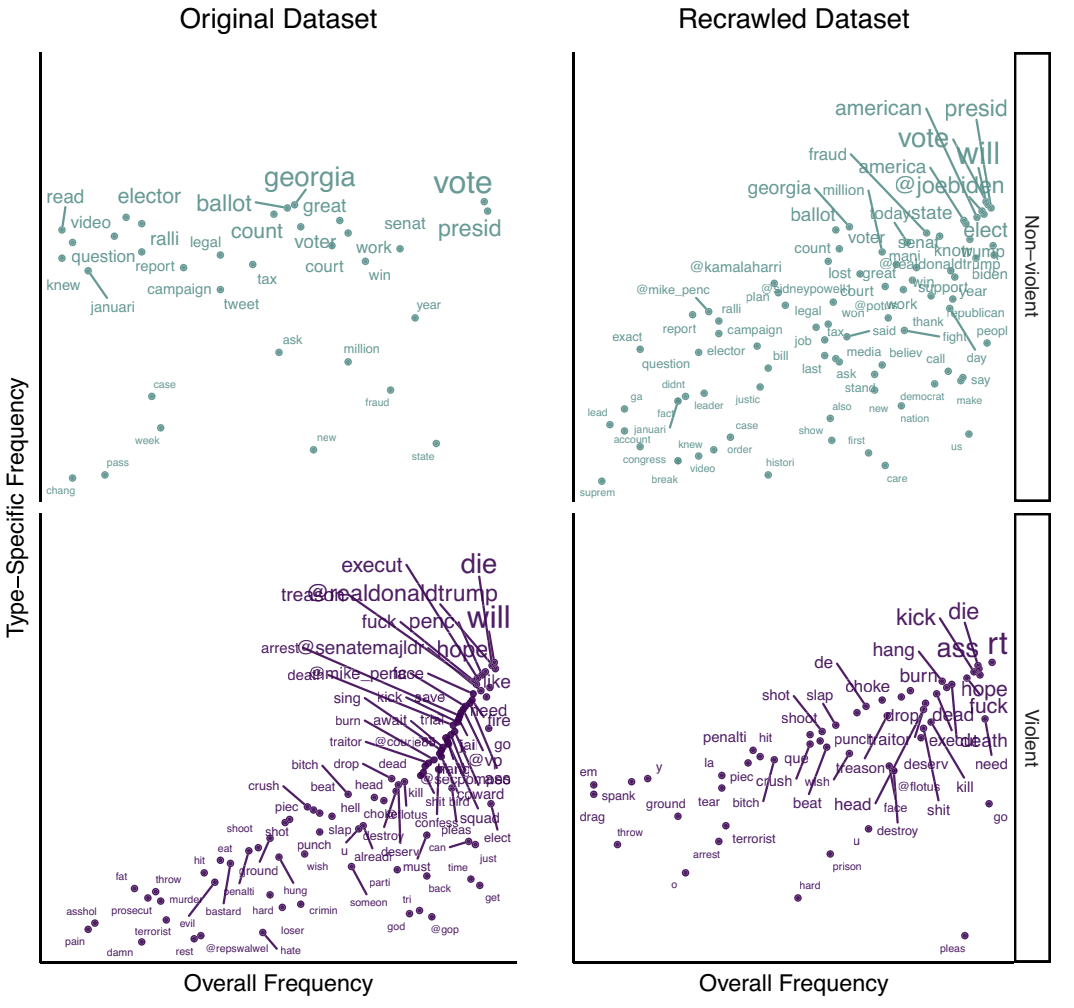


Figure 5. Comparison of terms grouped by violent (teal) and nonviolent (purple) tweets (a random sample of 5,000 tweets for each group) for both the original (left plot and upper-right panel) and the recrawled datasets (bottom-right panel). The *x*-axis shows the overall frequency of words in the dataset. The *y*-axis and the size of a word represent the frequency of words within a group. Following Kim (2023), several preprocessing techniques, such as lowercasing, stopword removal, and stemming, were applied to the tweets' content.

4.2. Replication: Statistical Model Findings

Knowing that the tweet content of both datasets is not representative anymore is one characteristic of irretrievable violent tweets. What are the implications for the overall distribution of words? In Figure 2 in their original paper, Kim focuses on a frequency comparison of words between and within violent and nonviolent tweets based on the Fightin' Words algorithm (Monroe, Colaresi, and Quinn 2008). The algorithm measures differences in word occurrences across groups by reducing (or increasing) the importance of very frequent (or infrequent) words.[16]

Figure 5 replicates the results for both datasets. The original analysis (left plot) reveals that violent tweets (lower panel) very often mention political actors like Donald Trump, Mike Pence, and Mike Pompeo. Nonviolent tweets (upper-left panel) do not show this trend. The recrawled dataset shows

[16] More details on the implementation of the algorithm are detailed in the Supplementary Material.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

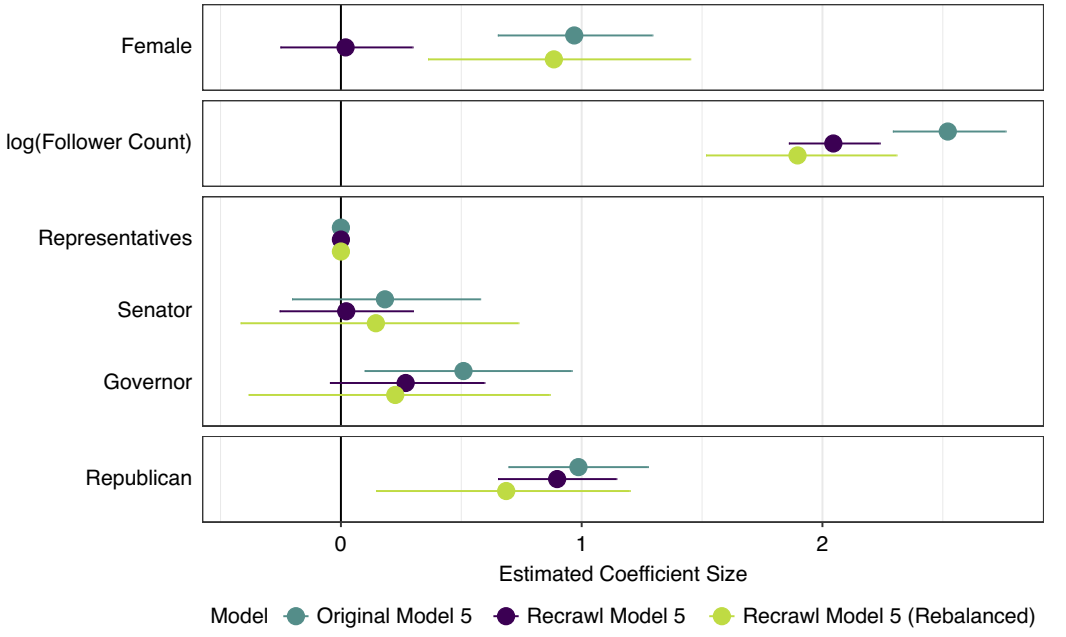


Figure 6. Comparison of regression model coefficients based on the original, recrawled, and resampled dataset with their 95% confidence intervals. The resampled regression model is a simulation based on rebalanced party and gender ratios following the original dataset distributions.

only one user mention (Michelle Obama) in the most significant words of the recrawled violent tweets. Beyond that, her prominence is somewhat limited to violent tweets, according to the Fightin' Words algorithm. Comparing recrawled violent tweets with the set of nonviolent frequencies reveals that the recrawled dataset characterizes itself by many user mentions in nonviolent tweets. Furthermore, even nondirect mentions and party names appear frequently in the nonviolent keywords (such as Trump, Biden, Republican, or Democrat). This comparison indicates a significant shift in the behavior of both groups between the original and recrawled datasets, reversing the original face validity outcome.

What are the implications of non-replicable descriptive findings on statistical models? The author calculates five negative binomial regressions to estimate the number of mentions of a political account in violent tweets (Table 4 in the original paper). The different model specifications include the position of a political account (representative, governor, or senator), whether an account represents a woman, a party dummy (Republican or non-Republican), as well as its logged follower count.[17]

Comparing the original model 5 with regressions using the recrawled dataset (see Figure 6) reveals that one of the paper's key findings of having more women targeted by violent tweets does not hold anymore. The estimate of the effect is close to zero, and the coefficient's 95% confidence interval of the recrawled model widely includes zero, too. In the recrawled data, it is still most likely that Republicans find more mentions in violent tweets. However, the new parameter estimate is nearly 10% smaller than in the original model. In a similar vein, the Senators' position estimate shrinks to only 11% of its original size. In addition to that, the Governors' parameter estimate reduces to half of its original value.

The recrawled regression model displays differing results, especially in the effect of gender. What patterns in the recrawled dataset drive these results? I leverage the original distribution of party and

[17] As it is uncritical for Twitter policies to share details about the number of followers of an account, I only consider the original data provided by the author for the number of followers in the recalculation of the regression models, while using the recrawled dataset for the remaining variables. Please note that while I focus on model specification 5, the Supplementary Material compares all models in Tables A.4 and A.5.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

gender to resample the distribution of recrawled tweets. The redistributed dataset allows me to calculate another regression model. Simulating the resampling and reestimation process of the model 1,000 times reduces randomness and generates uncertainty intervals in the resulting coefficients. The averaged coefficients and their lowest/highest 95% confidence intervals are shown in the third regression level (lime color). While all of their confidence intervals are much wider than the original and recrawled data, the rebalanced Female coefficient shows no significant difference from the original one according to the 95% confidence interval. Correspondingly, the rebalanced regression model depicts important removal patterns on the differing group shares displayed in Figure 4. The results indicate that, most likely, tweets mentioning male Republicans were more often removed from the dataset than tweets mentioning female Republicans or male or female Democrats.

Twitter's policies and API dismantling hinder the replication of research studies that involve the content of tweets, especially those containing sensitive content. While there are further findings in the original paper focusing on the follower network rather than the textual content, replicating the text-related steps of the study gives an idea of the implications of nonrandom tweet removal within sensitive datasets as several groups (of words, hashtags, and political actors) are no longer equally represented in the recrawled dataset compared to the original one. This unequal representation leads to descriptive and statistical model findings that differ considerably from the published figures. One could expect similar behavior on other sensitive Twitter datasets used by researchers within the discipline and beyond.

5. Data Access in the Post-API Era

While the replication issues related to Kim (2023) seem one case out of many, it highlights the issues inherent in replicating social media data studies. Following the suspension of Twitter's Academic Research Track API, many researchers avoid studying Twitter or are forced to cancel their ongoing Twitter projects (Davidson *et al.* 2023). Although Twitter still offers an API, it comes for 5,000 USD/month (Twitter 2023), which is not affordable for most researchers. Even if some can afford the new subscription plans, this does not solve the issue of non-replicable research, as the conditions concerning Twitter's restrictive data-sharing policies remain unchanged.

How can the research community respond effectively? Exploring ways to circumvent the restrictive behavior of commercial platforms and making work less dependent on their policies seems promising. One possible way out could be data donations (Davidson *et al.* 2023). Social media users can, by law, request a full copy of their data or install an app that collects it in real time and donate it for research. While this is a straightforward procedure that could be handled by centrally organized data donation platforms, researchers can only analyze the data of users they reach and those who consent. For studies that require analyzing sensitive datasets, researchers often cannot ask users for their consent. In these situations, a combination of approaches might lead to a promising way out: Public institutions can use their responsibility to archive data of public interest, including social media data. For example, when the Academic Research Track API was still available, the German National Library launched a data donation initiative to archive all German tweets. As proposed in Davidson *et al.* (2023), automatic crawlers could update these archives without needing an API. However, one must carefully evaluate this step, as crawling social media platforms might be a legal gray zone. The library intends to make its collected data available "within the German National Library's infrastructure."[18]

However, even if researchers have partial access to the data within institutions, Twitter's policy still prohibits researchers from directly sharing the raw content of tweets. Under these circumstances, sharing the tweets' one-way hashed content could be an option. One-way hash algorithms are designed to securely transform the original data into an encrypted version (Naor and Yung 1989). The one-way aspect of this well-established computer science technique prevents rehydrating tweets' raw content. Instead, a corresponding replication pipeline can reproduce the original results using the encrypted

[18]Data donation initiative of the German National Library: https://www.dnb.de/EN/Professionell/Sammeln/Sammlung_Websites/twitterArchiv.html.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

data (Bost *et al.* 2014). This could act as proof for reproducible research but would not replace direct access to social media data archives—still, though, hindering transparent replication.

Ultimately, academic journals are also responsible for ensuring a smooth and reliable review and replication process. Paying more attention to the origin and characteristics of data during review leads to higher-quality replication processes. This goes hand in hand with developing guidelines around the types of data that can be legally shared for scientific purposes.

In situations when none of the above approaches lead to replicable research, the discipline should broaden its scope to foster other data sources. While social media platforms provide a wealth of data, there are research questions about where alternative sources might lead to reliable *and* replicable results. Alternative sources especially include publicly available databases from institutional organizations. As a result, this leads to a diversification of data sources and less dependence on commercial platforms.

6. Conclusion

Even though Twitter experienced a lot of ups and downs due to the takeover of Elon Musk in October 2022, it still holds valuable data, which certainly keeps the platform essential for studying a wide range of social phenomena. While 75.00% of published Twitter studies in seven major political science journals might be potentially impeded by difficulties replicating the results due to missing replication data, this is especially alarming for 30.00% of all papers analyzing sensitive Twitter content. Based on tweet IDs in their replication archives, I demonstrate that only a third of the tweets in sensitive datasets are still available through the Twitter API. As this share is substantially lower than it is for nonsensitive datasets, it amplifies the worthiness and importance of understanding the tweet removal process on a more fine-grained level. In most cases, removed tweets do not result from an explicit user action but the final decision of Twitter's content moderation department. Hence, nonrandom tweet removal is not a direct phenomenon controlled by the users. Instead, it is Twitter itself that potentially affects the outcomes of replicating political science studies. I replicate some of the central findings of Kim (2023) based on a recrawled sensitive dataset established on tweets to illustrate. The case study suggests that irretrievable tweets might not only lead to a drastically reduced corpus size of less than 20.00% compared with the original dataset, but also nonrandom tweet mortality undermines some of the paper's fundamental descriptive and statistical model findings.

There is no easily feasible option for crawling tweets via the official Twitter API, making these results even more critical for replicable research. Foreseeing upcoming changes in the API is impossible, so the discipline needs to find alternatives to tackle both challenges: unavailable tweets due to removal and inaccessible tweets due to extensive API fees. This article presents a first outlook on data access possibilities in the post-API era, ranging from data donation to institutional obligations. Although platforms other than Twitter have not yet started to apply extensive fees for scientifically using their API, the issues and potential solutions raised in this paper are likely to also apply to other commercial social media platforms like TikTok, Instagram, or Facebook. This holds especially in light of recent changes to their APIs, which raise barriers to free, open, and easily replicable academic research. To give one example, TikTok tries to force users of their research API to update their collected dataset at least every 15 days to remove data points that were previously available but have since become unavailable (TikTok 2023)—and by that favoring a compromised replication instead of encouraging replicable research.

Acknowledgements. Christian Arnold, Brian Boyle, Christian Stecker, and the COMPTEXT 2023 audience provided very insightful comments on earlier versions of the manuscript. I thank six anonymous reviewers and the editor for their extremely helpful feedback. I also thank Leon Siefken for his excellent research assistance.

Funding Statement. There are no funding sources to report for this article.

Competing Interests. The authors have no competing interest to declare.

Data Availability Statement. Replication code for this article has been published in Code Ocean, a computational reproducibility platform that enables users to run the code and can be viewed interactively at https://doi.org/10.24433/CO.1624743.v3.

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

A preservation copy of the same code and data can also be accessed via Harvard Dataverse at https://doi.org/10.7910/DVN/UUDNM7 (Küpfer 2024a, 2024b).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.7.

References

Alizadeh, M., F. Gilardi, E. Hoes, K. J. Klüser, M. Kubli, and N. Marchal. 2022. "Content Moderation as a Political Issue: The Twitter Discourse around Trump's Ban." *Journal of Quantitative Description: Digital Media* 2 (October). https://doi.org/10.51685/jqd.2022.023. https://journalqd.org/article/view/3424.

Almuhimedi, H., S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. 2013. "Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 897–908. New York: Association for Computing Machinery. https://doi.org/10.1145/2441776.2441878

Alrababah, A., W. Marble, S. Mousa, and A. Siegel. 2021. "Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes." *American Political Science Review* 115 (4): 1111–1128.

Barrie, C., and J. C.-t. Ho. 2021. "Academictwitter: An R Package to Access the Twitter Academic Research Product Track V2 API Endpoint." *Journal of Open Source Software* 6 (62): 3272. https://doi.org/10.21105/joss.03272

Bost, R., R. A. Popa, S. Tu, and S. Goldwasser. 2014. Machine learning classification over encrypted data. *Cryptology ePrint Archive*.

Brie, E., and Y. Dufresne. 2020. "Tones from a Narrowing Race: Polling and Online Political Communication during the 2014 Scottish Referendum Campaign." *British Journal of Political Science* 50 (2): 497–509. https://doi.org/10.1017/S0007123417000606

Davidson, B. I., et al. 2023. Platform-Controlled Social Media APIs Threaten Open Science. *Nature Human Behaviour* 7: 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Dreber, A., and M. Johannesson. 2019. "Statistical Significance and the Replication Crisis in the Social Sciences." In *Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190625979.013.461

Elmas, T. 2023. "The Impact of Data Persistence Bias on Social Media Studies." In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci '23)*, 196–207. Austin, TX: Association for Computing Machinery. https://doi.org/10.1145/3578503.3583630

Frimer, J. A., et al. 2023. "Incivility Is Rising among American Politicians on Twitter." *Social Psychological and Personality Science* 14 (2): 259–269.

Hai, L., and K.-w. Fu. 2015. "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science." *PLoS One* 10 (8): 1–14. https://doi.org/10.1371/journal.pone.0134270

Hemphill, L., M. L. Hedstrom, and S. H. Leonard. 2021. "Saving Social Media Data: Understanding Data Management Practices among Social Media Researchers and Their Implications for Archives." *Journal of the Association for Information Science and Technology* 72: 109–197. https://doi.org/10.1002/asi.24368

Keller, F. B., D. Schoch, S. Stier, and J. Yang. 2020. "Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign." *Political Communication* 37 (2): 256–280. https://doi.org/10.1080/10584609.2019.1661888

Key, E. M. 2016. "How Are We Doing? Data Access and Replication in Political Science." *PS: Political Science and Politics* 49: 268–272. https://doi.org/10.1017/S1049096516000184

Kim, T. 2023. "Violent Political Rhetoric on Twitter." *Political Science Research and Methods* 11 (4): 673–695. https://doi.org/10.1017/psrm.2022.12

King, G. 1995. "Replication, Replication." *PS: Political Science and Politics* 28 (3): 444–452. https://doi.org/10.2307/420301

King, G. 2003. "The Future of Replication." *International Studies Perspectives* 4: 443–499. https://doi.org/10.1111/1528-3577.04105

Küpfer, A. 2024a. "Replication Data for: Non-Random Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies." https://doi.org/10.24433/CO.1624743.v3

Küpfer, A. 2024b. "Replication Data for: Non-Random Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies." https://doi.org/10.7910/DVN/UUDNM7.

Laitin, D. D., and R. Reich. 2017. "Trust, Transparency, and Replication in Political Science." *PS: Political Science and Politics* 50 (1): 172–175. https://doi.org/10.1017/S1049096516002365

Mitts, T. 2019. "From Isolation to Radicalization: Anti-Muslim Hostility and Support for ISIS in the West." *American Political Science Review* 113 (1): 173–194. https://doi.org/10.1017/S0003055418000618

Monroe, B., M. Colaresi, and K. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Funding Information: National Science Foundation (Grant BCS 05-27513 and BCS 07-14688)." *Political Analysis* 16 (4 SPEC. ISS.): 372–403. https://doi.org/10.1093/pan/mpn018

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press

Muchlinski, D., X. Yang, S. Birch, C. Macdonald, and I. Ounis. 2021. "We Need to Go Deeper: Measuring Electoral Violence Using Convolutional Neural Networks and Social Media." *Political Science Research and Methods* 9 (1): 122–139.

Naor, M., and M. Yung. 1989. "Universal One-Way Hash Functions and Their Cryptographic Applications." In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing (STOC '89)*, 33–43. Seattle, WA: Association for Computing Machinery. https://doi.org/10.1145/73007.73011

Noonan, J. 2022. "'Where Did Their Tweets Go?': A Quantitative Analysis of Parliamentarians 'Missing Tweets' in Western Europe." Independent thesis Advanced level (degree of Master (Two Years)), 20 credits/30 HE credits, Uppsala University, Disciplinary Domain of Humanities and Social Sciences, Faculty of Social Sciences, Department of Government. https://doi.org/10.48550/arXiv.2211.16506

Pfeffer, J., A. Mooseder, J. Lasser, L. Hammer, O. Stritzel, and D. Garcia. 2023. "This Sample Seems to Be Good Enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API." *Proceedings of the International AAAI Conference on Web and Social Media* 17, no. 1 (June): 720–729. https://doi.org/10.1609/icwsm.v17i1.22182; https://ojs.aaai.org/index.php/ICWSM/article/view/22182.

Pierri, F., L. Luceri, and E. Ferrara. 2022. "How Does Twitter Account Moderation Work? Dynamics of Account Creation and Suspension during Major Geopolitical Events." https://doi.org/10.48550/arXiv.2209.07614

Steinert-Threlkeld, Z. C. 2018. *Twitter as Data*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108529327

Temporão, M., C. V. Kerckhove, C. van Der Linden, Y. Dufresne, and J. M. Hendrickx. 2018. "Ideological Scaling of Social Media Users: A Dynamic Lexicon Approach." *Political Analysis* 26 (4): 457–473.

TikTok. 2023. *TikTok Research API Services Terms of Service*. https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en (accessed July 21, 2023).

Twitter. 2020. "Announcing Early Access to the Next Generation of the Twitter API." Accessed January 1, 2024. https://devcommunity.x.com/t/announcing-early-access-to-the-next-generation-of-the-twitter-api/139612.

Twitter. 2021. *Batch Compliance*. https://developer.twitter.com/en/docs/twitterapi/compliance/batch-compliance (accessed July 19, 2023).

Twitter. 2023. *Twitter API Tiers*. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api (accessed November 15, 2023).

Zubiaga, A. 2018. "A Longitudinal Assessment of the Persistence of Twitter Datasets." *Journal of the Association for Information Science and Technology* 69 (8): 974–984. https://doi.org/10.1002/asi.24026

Cite this article: Küpfer, A. (2024). Nonrandom Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies. *Political Analysis*, 493–506. https://doi.org/10.1017/pan.2024.7

https://doi.org/10.1017/pan.2024.7 Published online by Cambridge University Press