Non-random Tweet Mortality and Data Access Restrictions: Compromising the Replication of Sensitive Twitter Studies

Andreas Küpfer
Technical University Darmstadt
andreas.kuepfer@tu-darmstadt.de

Date: February 6th, 2024

Accepted for publication in Political Analysis

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 non-random tweet mortality and data access restrictions imposed by the platform. While sensitive datasets suffer a notably higher removal rate than non-sensitive datasets, attempting to replicate key findings of Kim's (2023) 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 non-random tweet mortality for future social media research on Twitter and similar platforms.

Keywords: text-as-data · twitter · replication

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 non-random patterns of tweet removal.

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

[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 ten years ago. The resulting dataset potentially includes fewer old tweets and higher availability of more recent tweets due to non-random tweet mortality. This might affect the findings of papers' studying historical Twitter content.

To investigate potential non-random 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 non-sensitive and sensitive datasets implies that tweets belonging to the latter category are removed at a noticeably higher rate. To show the impact of these non-random 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 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, i.e., 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 (Developers, 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; wa Liang 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 non-random removal patterns leading to incomplete replication datasets and, thus, inconsistent findings.

[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.

An emerging body of research examines extreme sentiment expressed in tweets (e.g., Alrababah et al., 2019; 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.

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 non-random deletion patterns dependent on whether a dataset is sensitive or non-sensitive.

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.

[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.

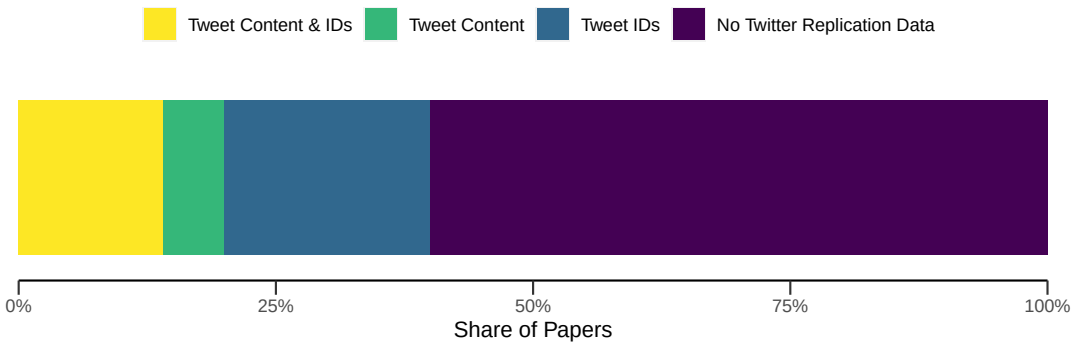

Figure 1: Different ways of how political scientists share Twitter datasets in replication archives among all 50 papers analyzing the content of tweets in seven major political science journals.

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.

3.2 Non-Random Deletion Patterns of Sensitive Datasets

Previous studies show that one should expect differences in the availability of tweets when looking at sensitive and non-sensitive 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

[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 Appendix details my annotation approach and provides an overview of all papers.

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

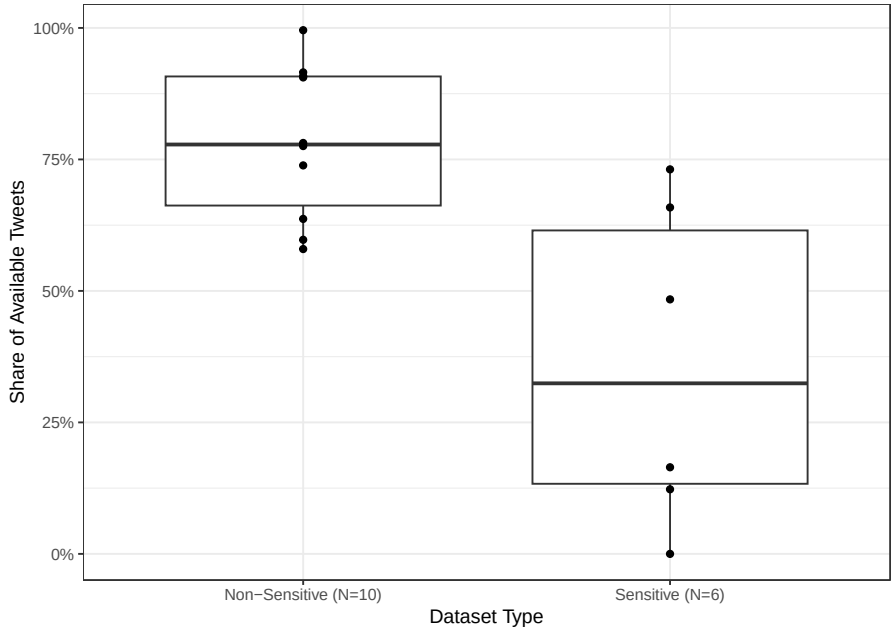

Figure 2: Availability rate of a random sample of up to 10,000 tweets from each of the 16 sensitive and non-sensitive paper datasets which shared at least their tweet IDs. Retrieving all tweets was attempted in cases where the original dataset contained fewer than 10,000 tweets. Temporão et al. (2018) 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.

tweet IDs. Ten of these papers work with non-sensitive datasets, representing 28.57% of all datasets annotated as non-sensitive. 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 non-sensitive 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 rate of mortality. If a dataset is sensitive, then this is associated with its mortality chance.

Relying on a data basis of more than three-quarters of still available tweets in non-sensitive datasets sounds convincing to initiate a replication attempt. However, replicating studies could become chal-

[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 ting 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 amongst the recrawled datasets is 2020 for both sensitive and non-sensitive datasets.

lenging 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 non-sensitive datasets also include sensitive tweets. 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., non-sensitive 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 non-random tweet mortality directly impacting our data basis.

4 Case Study: Implications for Replicating Sensitive Twitter Studies

Sensitive datasets suffer from a notably higher loss of tweets than non-sensitive 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 US 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 US election in 2020 on social media but combines three methodological and data-wise characteristics well-suited for an insightful replication. First, it studies violent tweets and compares them with non-violent ones, which supports analyzing differences between the behavior across both dataset types. Second, it studies rather

[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, 4.4, and 4.5 in the original paper. More details on substantial aspects and the methodology of creating the original dataset are described in the Appendix.

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 due to user suspensions. It is important to note that these decisions are taken by Twitter, e.g., 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 Appendix 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 5000 violent tweets aggregated in a document-frequency matrix openly distributed by the author. I approach the represen-

[11] See Appendix 2 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 data set 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] Altough 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.

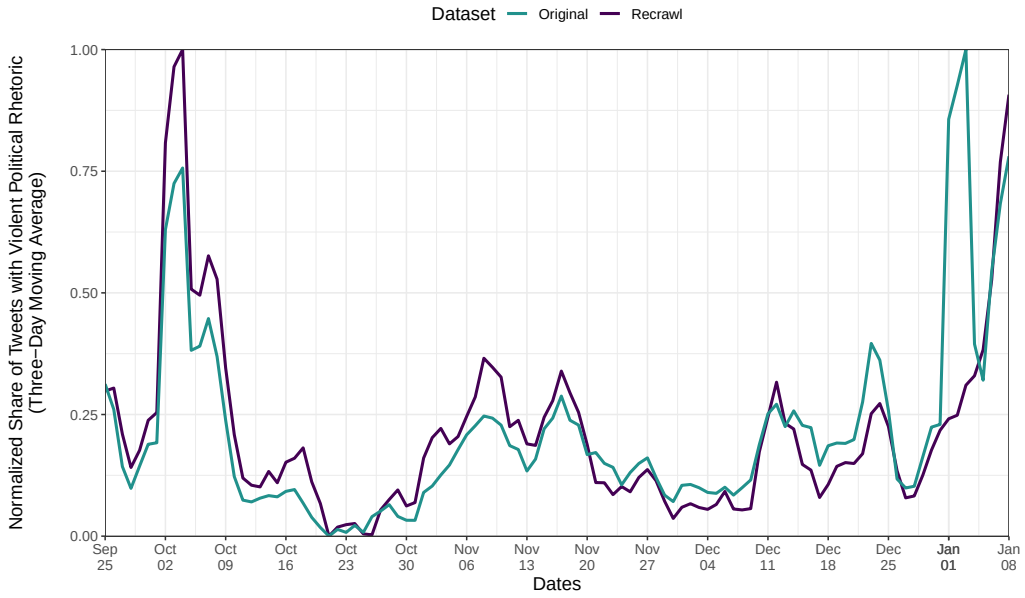

Figure 3: Timeline comparison of normalized proportion of violent political rhetoric tweets during the US election 2020 for both the original and recrawled datasets (replication of Kim (2023)). Proportions are based on aggregated information on 215,923 original and 35,552 recrawled tweets.

tativity 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 changes 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 the Appendix in Table A.2.

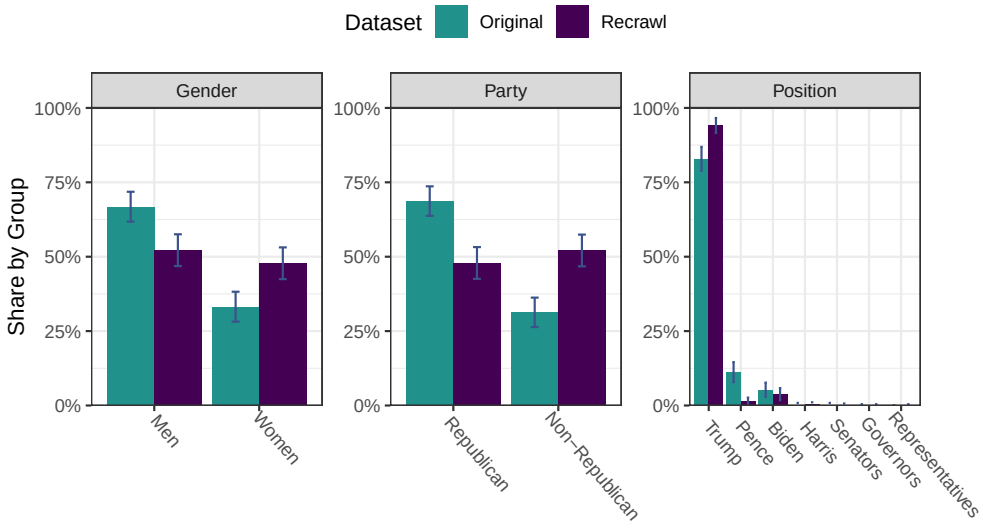

Figure 4: Comparative distribution of mean mentions of accounts in tweets containing violent political rhetoric by gender, party, and position in the original and recrawled datasets. Uncertainty displays the 95% confidence interval of each group. Proportions are based on aggregated information on 215,923 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 behaves differently and fails to mirror the original data in its most important aspects. Accordingly, non-random 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 Appendix 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 non-violent 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 non-random removal patterns where tweets referring to women and Republicans are more likely to be removed than those referencing men and non-Republicans.

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 non-violent 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. Non-violent tweets (upper-left panel) do not show this trend. The recrawled dataset shows 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 non-violent frequencies reveals that the

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

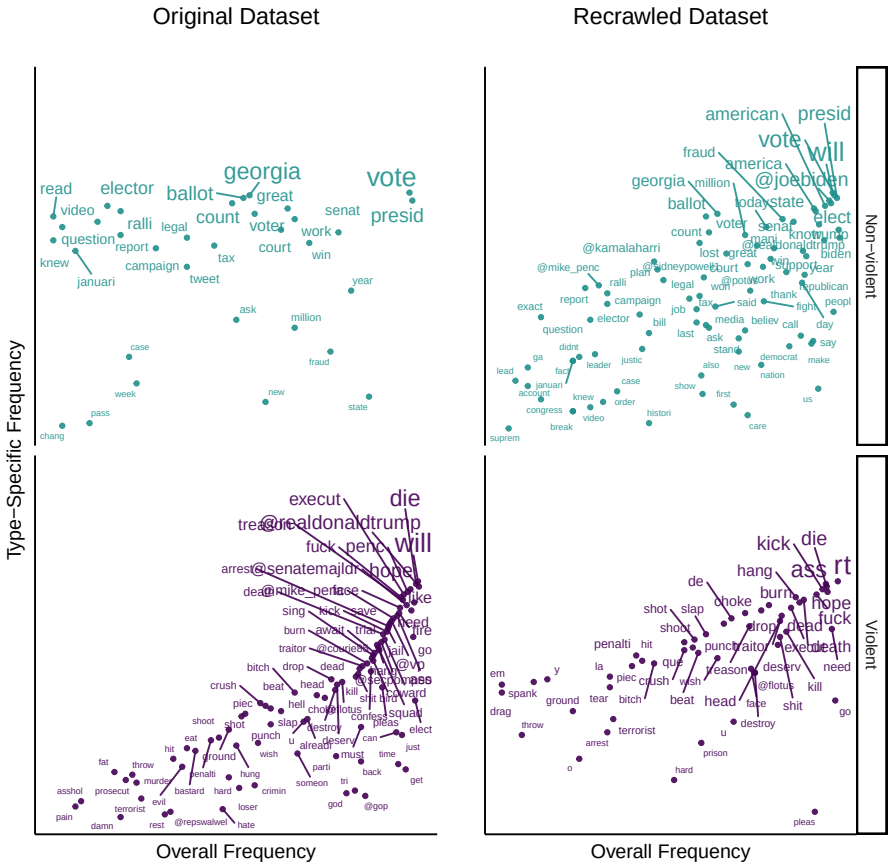

Figure 5: Comparison of terms grouped by violent (teal) and non-violent (purple) tweets (a random sample of 5000 tweets for each group) for both the original (left plot and upper-right panel) and the recrawled dataset (bottom-right panel). The x-axis shows the overall frequency of words in the dataset. 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.

recrawled dataset characterizes itself by many user mentions in non-violent tweets. Furthermore, even non-direct mentions and party names appear frequently in the non-violent 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

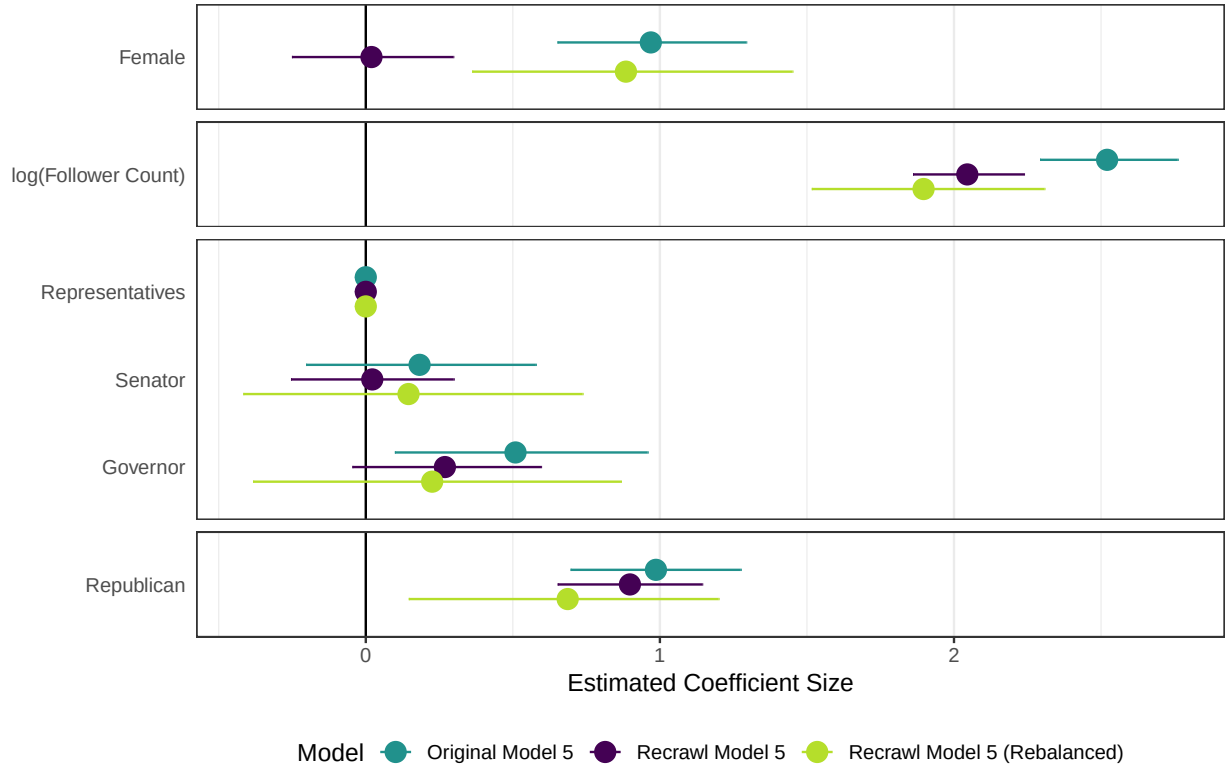

Figure 6: Comparison of regression model coefficients based on the original, recrawled, and resampled dataset with their 95% confidence intervals. The resampled regression model is a simulation based on rebalanced party and gender ratios following the original dataset distributions.

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.

[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 Appendix compares all models in Tables A.4 and A.5.

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 gender to resample the distribution of recrawled tweets. The redistributed dataset allows me to calculate another regression model. Simulating the resampling and re-estimation process of the model 1000 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 non-random 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 sub-

scription 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 grey 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 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

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

to higher-quality replication processes. This goes hand in hand with developing guidelines around the type 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, diversifying data sources and reducing the field's 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 non-sensitive 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, non-random 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 non-random 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.

Acknowledgement

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.

References

Alizadeh, Meysam, Fabrizio Gilardi, Emma Hoes, K. Jonathan Klüser, Maël Kubli and Nahema Marchal. 2022. "Content Moderation As a Political Issue: The Twitter Discourse Around Trump's Ban." *Journal of Quantitative Description: Digital Media* 2.
URL: *https://journalqd.org/article/view/3424*

Almuhimedi, Hazim, Shomir Wilson, Bin Liu, Norman Sadeh and Alessandro Acquisti. 2013. Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets. Association for Computing Machinery pp. 897–908.

Alrababah, Ala, William Marble, Salma Mousa and Alexandra Siegel. 2019. "Can exposure to celebrities reduce prejudice? The effect of Mohamed Salah on Islamophobic behaviors and attitudes." *American Political Science Review* .

Barrie, Christopher and Justin Chun ting 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.
URL: *https://doi.org/10.21105/joss.03272*

Bost, Raphael, Raluca Ada Popa, Stephen Tu and Shafi Goldwasser. 2014. "Machine learning classification over encrypted data." *Cryptology ePrint Archive* .

Brie, Evelyne and Yannick 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.

Davidson, Brittany I., Darja Wischerath, Daniel Racek, Douglas A. Parry, Emily Godwin, Joanne Hinds, Dirk van der Linden, Jonathan F. Roscoe, Laura Ayravainen and Alicia G. Cork. 2023. "Platform-controlled social media APIs threaten open science." *Nature Human Behaviour* .

Developers, Twitter. 2020. "Announcing Early Access to the next generation of the Twitter API.". Accessed: 2024-01-22.
URL: *https://devcommunity.x.com/t/announcing-early-access-to-the-next-generation-of-the-twitter-api/139612*

Dreber, Anna and Magnus Johannesson. 2019. "Statistical Significance and the Replication Crisis in the Social Sciences.".

Elmas, Tugrulcan. 2023. The Impact of Data Persistence Bias on Social Media Studies. In *Proceedings of the 15th ACM Web Science Conference 2023*. WebSci '23 New York, NY, USA: Association for Computing Machinery p. 196–207.

Frimer, Jeremy A, Harinder Aujla, Matthew Feinberg, Linda J Skitka, Karl Aquino, Johannes C Eichstaedt and Robb Willer. 2023. "Incivility is rising among American politicians on Twitter." *Social Psychological and Personality Science* 14(2):259–269.

Hemphill, Libby, Margaret L Hedstrom and Susan Hautaniemi 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* 72:109–97.

Keller, Franziska B, David Schoch, Sebastian Stier and JungHwan Yang. 2020. "Political astroturfing on Twitter: How to coordinate a disinformation campaign." *Political communication* 37(2):256–280.

Key, Ellen M. 2016. "How Are We Doing? Data Access and Replication in Political Science." *PS: Political Science & Politics* 49:268–272.

Kim, Taegyoon. 2023. "Violent political rhetoric on Twitter." *Political Science Research and Methods* 11(4):673–695.

King, Gary. 1995. "Replication, Replication." *PS: Political Science & Politics* 28:444–452.

King, Gary. 2003. "The future of replication." *International Studies Perspectives* .

Laitin, David D and Rob Reich. 2017. "Trust, Transparency, and Replication in Political Science." *PS: Political Science & Politics* 50:172–175.

Mitts, Tamar. 2019. "From isolation to radicalization: Anti-Muslim hostility and support for ISIS in the West." *American Political Science Review* 113(1):173–194.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403. Funding Information: National Science Foundation (grant BCS 05-27513 and BCS 07-14688).

Muchlinski, David, Xiao Yang, Sarah Birch, Craig Macdonald and Iadh 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 New York, NY, USA: Association for Computing Machinery p. 33–43.
URL: *https://doi.org/10.1145/73007.73011*

Noonan, Joseph. 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.

Pfeffer, Jürgen, Angelina Mooseder, Jana Lasser, Luca Hammer, Oliver Stritzel and David 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(1):720–729.
URL: *https://ojs.aaai.org/index.php/ICWSM/article/view/22182*

Pierri, Francesco, Luca Luceri and Emilio Ferrara. 2022. "How Does Twitter Account Moderation Work? Dynamics of Account Creation and Suspension During Major Geopolitical Events.".

Steinert-Threlkeld, Zachary C. 2018. *Twitter as Data*. Cambridge University Press.

Temporão, Mickael, Corentin Vande Kerckhove, Clifton van Der Linden, Yannick Dufresne and Julien 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: 2023-07-21.

Twitter. 2021. "Batch Compliance." https://developer.twitter.com/en/docs/twitterapi/compliance/batch-compliance. Accessed: 2023-07-19.

Twitter. 2023. "Twitter API Tiers." https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api. Accessed: 2023-11-15.

wa Liang Hai, King and Fu. 2015. "Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science." *PLOS ONE* 10:1–14.

Zubiaga, Arkaitz. 2018. "A Longitudinal Assessment of the Persistence of Twitter Datasets." *Journal of the Association for Information Science and Technology* 69.

Supplementary Material

Andreas Küpfer
Technical University Darmstadt
andreas.kuepfer@tu-darmstadt.de

A Paper Study: Detailed Information

To determine whether a paper analyzes Twitter data, I first looked for relevant methods or dataset descriptions in the paper itself or its appendix. For the remaining portion of the papers, I automatically downloaded all publicly available official replication repositories. This lets me identify if someone shared their dataset as tweet IDs by searching for the following pattern:

(^|,|;|”|’|)(\d{8,})($|,|;|”|’|)

This regular expression looks for numeric sequences of at least eight numbers embedded in typical delimiters such as spaces, quotes, or semicolons and is applied on all files which are either text-files or readable by the R package `rio` (hong Chan et al., 2021). It turned out that this approach acts as a good heuristic to solve this task. I then manually checked all downloaded replication archives again, looking for false positives. After that, another identifier is added, which represents whether the replication archive contains datasets with raw textual tweet content data.

Table A.1 lists the papers that systematically analyze the content of tweets. The columns Tweet IDs and Tweet Content contain a ✓ if they share data for the respective categories and their replication archives. Sensitive content is marked with a ✓ for papers that explicitly analyze the content of tweets in datasets belonging to at least one of the following three areas (inspired by Elmas (2023)):

- Fake News/Disinformation
- Hate Speech/Violence/Terrorism
- Bots

In many cases, these categories are strongly connected with each other (e.g., bots often share fake news, or terrorism is strongly related to disinformation). My definition of sensitive datasets does not mark a dataset of tweets by Donald Trump as being sensitive, even if one would expect a few tweets containing fake news and disinformation, as most of these tweets is general political content.

Table A.1: Papers from seven political science journals mentioning the keyword "Twitter". The columns `Tweet IDs` and `Tweet content` show whether the authors shared the respective data in their replication archive. `Is sensitive content?` marks if a paper analyzes sensitive Twitter datasets.

Paper	Journal	Tweet IDs	Tweet content	Is sensitive content?
Beauchamp (2017)	AJPS	-	-	-
Benton and Philips (2020)	AJPS	-	-	-
Fong and Grimmer (2021)	AJPS	-	-	-
King, Lam and Roberts (2017)	AJPS	✓	✓	-
Nielsen (2020)	AJPS	✓	-	-
Alrababah et al. (2019)	APSR	✓	-	✓
Barberá et al. (2019)	APSR	-	-	-
Mitts (2019)	APSR	-	-	✓
Osmundsen et al. (2021)	APSR	-	-	✓
Pan and Siegel (2020)	APSR	-	-	✓
Silva and Proksch (2021)	APSR	✓	-	-
Sobolev et al. (2020)	APSR	-	-	-
Stukal et al. (2022)	APSR	✓	-	✓
Brie and Dufresne (2020)	BJPS	✓	✓	-
Clarke and Kocak (2020)	BJPS	✓	-	-
Jones and Mattiacci (2019)	BJPS	-	-	-
Munger et al. (2022)	BJPS	-	✓	-
Bisbee and Lee (2022)	JOP	-	-	-
Boucher and Thies (2019)	JOP	-	-	-
Das et al. (2022)	JOP	-	-	-
Mitts, Phillips and Walter (2022)	JOP	-	-	✓
Skytte (2022)	JOP	-	-	-
Bestvater and Monroe (2022)	PA	-	✓	-
Kubinec and Owen (2021)	PA	-	✓	-
Miller, Linder and Mebane (2020)	PA	✓	✓	-
Temporão et al. (2018)	PA	✓	-	-
Castanho Silva, Proksch et al. (2022)	PSRM	✓	-	-
Cirone and Hobbs (2023)	PSRM	✓	✓	✓
Kim (2023)	PSRM	✓	-	✓
Muchlinski et al. (2021)	PSRM	✓	-	✓
Munger et al. (2019)	PSRM	-	-	-
Settle et al. (2016)	PSRM	-	-	-
Bradshaw et al. (2020)	PolComm	-	-	✓
Cassell (2021)	PolComm	-	-	-
DiResta, Grossman and Siegel (2022)	PolComm	-	-	✓
Gilardi et al. (2022)	PolComm	✓	-	-
Guess et al. (2019)	PolComm	-	-	-
Kang et al. (2018)	PolComm	-	-	-
Keller and Klinger (2019)	PolComm	-	-	✓
Keller et al. (2020)	PolComm	-	-	✓
Ketelaars and Sevenans (2021)	PolComm	✓	✓	-

Kligler-Vilenchik et al. (2021)	PolComm	-	-	-
Kobayashi and Ichifuji (2015)	PolComm	-	-	-
Konitzer et al. (2019)	PolComm	-	-	-
Linvill and Warren (2020)	PolComm	-	-	✓
Margolin, Hannak and Weber (2018)	PolComm	✓	✓	✓
Muddiman, McGregor and Stroud (2019)	PolComm	-	-	-
Popa et al. (2020)	PolComm	-	-	-
Stier et al. (2018)	PolComm	-	-	-
Yarchi, Baden and Kligler-Vilenchik (2021)	PolComm	✓	✓	-

A.1 Results per Journal

Figure A.1 depicts the proportions of each Twitter-related replication category grouped by the respective journal. It highlights that there are a few journals where no Twitter replication data exists for more than half of their Twitter-related papers. Authors of Twitter research in AJPS and PA either share their Tweet IDs, or the content of the analyzed tweets, or even both. On the contrary, Twitter papers published in JOP do not contain any Twitter-related replication data.

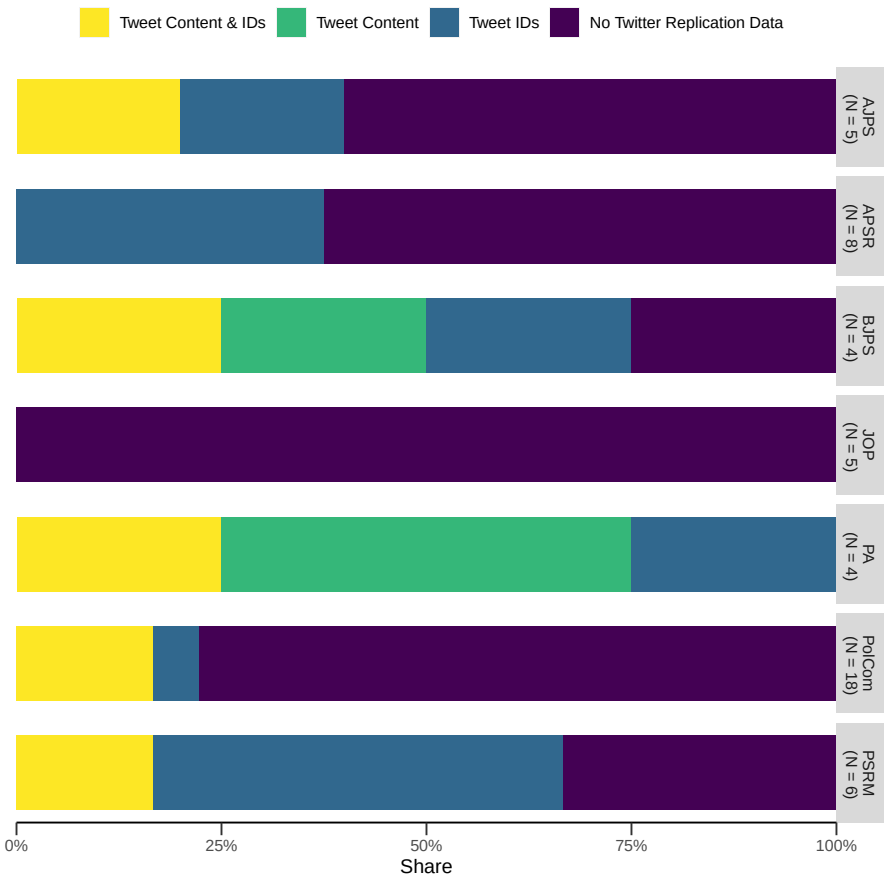

Figure A.1: Different methods of publishing Twitter datasets in papers, categorized by the journal.

B Kim (2023) Data Leveraging Pipeline and Replication Study

Choosing to replicate Kim (2023) in favor of other potential replication candidates is based on valid grounds. While Alrababah et al. (2019); Muchlinski et al. (2021) do not provide much further replication data related to tweet content beyond code and tweet IDs, Stukal et al. (2022); Cirone and Hobbs (2023) do not focus on the comparison of sensitive and non-sensitive studies. Finally, Margolin, Hannak and Weber (2018) do not study longitudinal aspects of their dataset.

Kim's study examines violent political rhetoric on social media and its relationship with offline political violence, focusing on the Capitol Riot. The author introduces a new automated method to identify violent rhetoric on Twitter and finds that users who engage in such rhetoric are ideologically extreme and located on the fringe of the communication network. The tweets are more frequently targeted at women and Republican politicians and are often shared across the ideological divide, creating the potential for co-radicalization.

The database for these findings grounds a random proportion of 1% of all tweets in real-time by taking advantage of the Streaming API by Twitter. These tweets are then processed in a pipeline containing several keyword-based filtering approaches and finally classified as holding political violence or not, using a transformer model.

B.1 Reasons for Unavailable Tweets

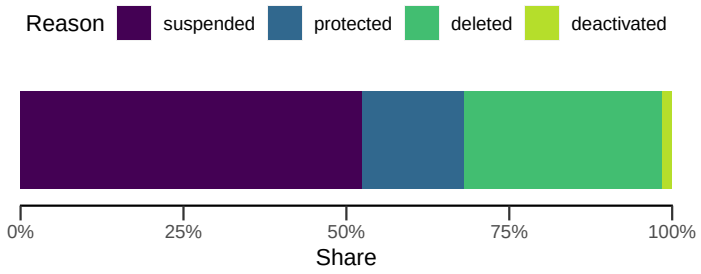

Figure A.2: Reasons for unavailable users in Kim (2023) that posted violent tweets according to Twitters' Compliance API endpoint. Results are weighted with the number of tweets each user posted according to aggregated information in the replication files of Kim (2023). The purple area shows the share of suspended users due to platform actions, while the other sections (visualized in brighter colors) highlight reasons for unavailable users due to explicit user actions.

While A.2 depicts the reasons for removed tweets based on removed user accounts weighted by their total number of tweets, there are also tweet removals without the complete user becoming unavailable. However, according to the Compliance API, only 10,425 tweets were removed explicitly, which is 5.78% of all unavailable tweets. This means that 169,946 (94.22%) tweets were removed due to account suspensions, deletions, protections, or deactivations.

B.2 Representativity of Recrawled Content

Table A.2: Results of Welch's t-test comparing different features between a sample of 5000 violent political rhetoric tweets of the original population (published by the author along with the replication files) and the recrawled dataset. A 95% confidence interval excluding zero is an indicator that a feature is different in both datasets.

Type	95% confidence interval
Textual Content	$[-3.94; -1.31]$
Hashtag	$[-1.72; -1.17]$
User Mentions	$[-5.44; 0.96]$

B.3 Hashtag Frequency

Table A.3: Hashtag frequency comparison of the original dataset and recrawl. Arrows indicate the direction of rank change, dashes show no difference, and question marks reflect that a hashtag is not available anymore. The different colors depict the intensity (red = considerable change; green = slight change).

	Count		Rank		
Hashtag	Original	Recrawl	Original		Recrawl
#wethepeople	1511	8	1	⇓	24
#1	1398	-	2	?	
#pencecard	1341	3	3	⇓	29
#maga	881	55	4	—	4
#fightback	702	1	5	⇓	31
#1776again	672	-	6	?	
#antifaarefascists	607	2	7	⇓	30
#blmareracists	607	1	7	⇓	31
#covid19	606	83	8	⇑	1
#treason	555	18	9	⇓	14
#vote	498	13	10	⇓	19
#trump	452	26	11	⇑	9
#trump2020	434	42	12	⇑	5
#walterreed	428	62	13	⇑	3
#savebrandonbernard	421	78	14	⇑	2
#pardonsnowden	365	1	15	⇓	31
#traitortrump	358	15	16	⇓	17
#freeassange	356	-	17	?	
#punkaf	354	-	18	?	
#godwins	244	1	19	⇓	31

B.4 Fightin' Words Algorithm

Fightin' Words (Monroe, Colaresi and Quinn, 2008) is a lexical feature selection algorithm that helps in determining which terms are most distinctively characteristic of a particular textual group's (sensitive

versus non-sensitive tweets) language usage. The calculation of the word importance yielded by the algorithm is implemented as follows:

$$\hat{w} = normalize(normalize(log(g\_dtm))')' \tag{1a}$$

In 1a, g_dtm describes the document-term frequency per group. Log transformation and normalization[1] of these frequencies leads to $\hat{w}$, which builds the basis for the final group-wise word importance.

$$w\_se = \sqrt{\frac{1}{g\_dtm} + \frac{1}{g\_dtm\_w} + \frac{1}{g\_dtm\_k} + \frac{1}{g\_dtm\_kw}} \tag{1b}$$

Equation 1b calculates the standard error for each word w per group. The suffix $\_w$ is about the usage of other terms in the same group k, whereas the suffix $\_k$ describes the frequency of the current word w spoken by groups other than k. Finally, $\_kw$ is the total number of words not spoken by group k other than the specific word w.

$$\hat{w}_{zeta} = \frac{\hat{w}}{w\_se} \tag{1c}$$

Finally, $\hat{w}$ is divided by the corresponding standard error. The resulting zeta scores in $\hat{w}_{zeta}$ represent how distinctive a term is for a particular group. Figure 5 leverages these scores per group and word. It contains very dense information about the group-dependent relative frequency of each keyword on the x-axis. At the same time, the y-axis (and the size of a specific word) displays the extent to which a keyword is associated with a group.

B.5 Regression Models

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	4.02	5.00	4.47	−8.79	−9.44
	(0.10)	(0.10)	(0.12)	(0.52)	(0.59)
Position: Governors	1.08				0.51
	(0.30)				(0.22)
Position: Senators	2.15				0.18
	(0.23)				(0.18)
Female		−0.38			0.97
		(0.21)			(0.15)
Republican			0.78		0.99
			(0.18)		(0.13)
Follower Count (log)				2.57	2.52
				(0.11)	(0.13)
AIC	5255.01	5364.26	5348.76	4689.54	4636.13
Num. obs.	585	585	585	562	562

Table A.4: Negative binomial regression models (original dataset)

[1] $normalization$ is applied two times to normalize both within-group and across-groups.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	2.47	2.87	2.82	−7.67	−8.69
	(0.11)	(0.09)	(0.12)	(0.49)	(0.53)
Position: Governors	0.56				0.27
	(0.25)				(0.16)
Position: Senators	0.96				0.02
	(0.19)				(0.14)
Female		0.10			0.02
		(0.20)			(0.14)
Republican			0.14		0.90
			(0.17)		(0.12)
Follower Count (log)				1.96	2.05
				(0.10)	(0.10)
AIC	2380.92	2406.76	2406.32	2017.15	1966.65
Num. obs.	328	328	328	322	322

Table A.5: Negative binomial regression models (recrawled dataset)

References

Alrababah, Ala, William Marble, Salma Mousa and Alexandra Siegel. 2019. "Can exposure to celebrities reduce prejudice? The effect of Mohamed Salah on Islamophobic behaviors and attitudes." *American Political Science Review* .

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost and Joshua A Tucker. 2019. "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data." *American Political Science Review* 113(4):883–901.

Beauchamp, Nicholas. 2017. "Predicting and interpolating state-level polls using Twitter textual data." *American Journal of Political Science* 61(2):490–503.

Benton, Allyson L and Andrew Q Philips. 2020. "Does the realDonaldTrump really matter to financial markets?" *American Journal of Political Science* 64(1):169–190.

Bestvater, Samuel E and Burt L Monroe. 2022. "Sentiment is not stance: target-aware opinion classification for political text analysis." *Political Analysis* pp. 1–22.

Bisbee, James and Diana Da In Lee. 2022. "Objective facts and elite cues: partisan responses to covid-19." *The Journal of Politics* 84(3):1278–1291.

Boucher, Jean-Christophe and Cameron G Thies. 2019. ""I am a tariff man": The power of populist foreign policy rhetoric under President Trump." *The Journal of Politics* 81(2):712–722.

Bradshaw, Samantha, Philip N Howard, Bence Kollanyi and Lisa-Maria Neudert. 2020. "Sourcing and automation of political news and information over social media in the United States, 2016-2018." *Political Communication* 37(2):173–193.

Brie, Evelyne and Yannick 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.

Cassell, Kaitlen J. 2021. "When "following" the leader inspires action: Individuals' receptivity to discursive frame elements on social media." *Political Communication* 38(5):581–603.

Castanho Silva, Bruno, Sven-Oliver Proksch et al. 2022. "Politicians unleashed? Political communication on Twitter and in parliament in Western Europe." *Political Science Research and Methods* 10(4):776–792.

Cirone, Alexandra and William Hobbs. 2023. "Asymmetric flooding as a tool for foreign influence on social media." *Political Science Research and Methods* 11(1):160–171.

Clarke, Killian and Korhan Kocak. 2020. "Launching revolution: Social media and the Egyptian uprising's first movers." *British Journal of Political Science* 50(3):1025–1045.

Das, Sanmay, Betsy Sinclair, Steven W Webster and Hao Yan. 2022. "All (Mayoral) Politics Is Local?" *The Journal of Politics* 84(2):1021–1034.

DiResta, Renée, Shelby Grossman and Alexandra Siegel. 2022. "In-house vs. outsourced trolls: How digital mercenaries shape state influence strategies." *Political Communication* 39(2):222–253.

Elmas, Tugrulcan. 2023. The Impact of Data Persistence Bias on Social Media Studies. In *Proceedings of the 15th ACM Web Science Conference 2023*. WebSci '23 New York, NY, USA: Association for Computing Machinery p. 196–207.

Fong, Christian and Justin Grimmer. 2021. "Causal inference with latent treatments." *American Journal of Political Science* .

Gilardi, Fabrizio, Theresa Gessler, Maël Kubli and Stefan Müller. 2022. "Social media and political agenda setting." *Political Communication* 39(1):39–60.

Guess, Andrew, Kevin Munger, Jonathan Nagler and Joshua Tucker. 2019. "How accurate are survey responses on social media and politics?" *Political Communication* 36(2):241–258.

hong Chan, Chung, Geoffrey CH Chan, Thomas J. Leeper and Jason Becker. 2021. *rio: A Swiss-army knife for data file I/O*. R package version 0.5.29.

Jones, Benjamin T and Eleonora Mattiacci. 2019. "A manifesto, in 140 characters or fewer: Social media as a tool of rebel diplomacy." *British Journal of Political Science* 49(2):739–761.

Kang, Taewoo, Erika Franklin Fowler, Michael M Franz and Travis N Ridout. 2018. "Issue consistency? comparing television advertising, tweets, and e-mail in the 2014 senate campaigns." *Political Communication* 35(1):32–49.

Keller, Franziska B, David Schoch, Sebastian Stier and JungHwan Yang. 2020. "Political astroturfing on Twitter: How to coordinate a disinformation campaign." *Political communication* 37(2):256–280.

Keller, Tobias R and Ulrike Klinger. 2019. "Social bots in election campaigns: Theoretical, empirical, and methodological implications." *Political Communication* 36(1):171–189.

Ketelaars, Pauline and Julie Sevenans. 2021. "It'sa matter of timing. How the timing of politicians' information subsidies affects what becomes news." *Political Communication* 38(3):260–280.

Kim, Taegyoon. 2023. "Violent political rhetoric on Twitter." *Political Science Research and Methods* 11(4):673–695.

King, Gary, Patrick Lam and Margaret E Roberts. 2017. "Computer-assisted keyword and document set discovery from unstructured text." *American Journal of Political Science* 61(4):971–988.

Kligler-Vilenchik, Neta, Maya de Vries Kedem, Daniel Maier and Daniela Stoltenberg. 2021. "Mobilization vs. demobilization discourses on social media." *Political communication* 38(5):561–580.

Kobayashi, Tetsuro and Yu Ichifuji. 2015. "Tweets that matter: Evidence from a randomized field experiment in Japan." *Political Communication* 32(4):574–593.

Konitzer, Tobias, David Rothschild, Shawndra Hill and Kenneth C Wilbur. 2019. "Using big data and algorithms to determine the effect of geographically targeted advertising on vote intention: Evidence from the 2012 US presidential election." *Political Communication* 36(1):1–16.

Kubinec, Robert and John Owen. 2021. "When Groups Fall Apart: Identifying Transnational Polarization during the Arab Uprisings." *Political Analysis* 29(4):522–540.

Linvill, Darren L and Patrick L Warren. 2020. "Troll factories: Manufacturing specialized disinformation on Twitter." *Political Communication* 37(4):447–467.

Margolin, Drew B, Aniko Hannak and Ingmar Weber. 2018. "Political fact-checking on Twitter: When do corrections have an effect?" *Political Communication* 35(2):196–219.

Miller, Blake, Fridolin Linder and Walter R Mebane. 2020. "Active learning approaches for labeling text: review and assessment of the performance of active learning approaches." *Political Analysis* 28(4):532–551.

Mitts, Tamar. 2019. "From isolation to radicalization: Anti-Muslim hostility and support for ISIS in the West." *American Political Science Review* 113(1):173–194.

Mitts, Tamar, Gregoire Phillips and Barbara F Walter. 2022. "Studying the impact of ISIS propaganda campaigns." *The Journal of Politics* 84(2):1220–1225.

Monroe, Burt L., Michael P. Colaresi and Kevin M. Quinn. 2008. "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict." *Political Analysis* 16(4 SPEC. ISS.):372–403. Funding Information: National Science Foundation (grant BCS 05-27513 and BCS 07-14688).

Muchlinski, David, Xiao Yang, Sarah Birch, Craig Macdonald and Iadh 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.

Muddiman, Ashley, Shannon C McGregor and Natalie Jomini Stroud. 2019. "(Re) claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries." *Political Communication* 36(2):214–226.

Munger, Kevin, Patrick J Egan, Jonathan Nagler, Jonathan Ronen and Joshua Tucker. 2022. "Political knowledge and misinformation in the era of social media: Evidence from the 2015 UK election." *British Journal of Political Science* 52(1):107–127.

Munger, Kevin, Richard Bonneau, Jonathan Nagler and Joshua A Tucker. 2019. "Elites tweet to get feet off the streets: Measuring regime social media strategies during protest." *Political Science Research and Methods* 7(4):815–834.

Nielsen, Richard A. 2020. "Women's authority in patriarchal social movements: the case of female Salafi preachers." *American Journal of Political Science* 64(1):52–66.

Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2021. "Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter." *American Political Science Review* 115(3):999–1015.

Pan, Jennifer and Alexandra A Siegel. 2020. "How Saudi crackdowns fail to silence online dissent." *American Political Science Review* 114(1):109–125.

Popa, Sebastian Adrian, Zoltán Fazekas, Daniela Braun and Melanie-Marita Leidecker-Sandmann. 2020. "Informing the public: How party communication builds opportunity structures." *Political Communication* 37(3):329–349.

Settle, Jaime E, Robert M Bond, Lorenzo Coviello, Christopher J Fariss, James H Fowler and Jason J Jones. 2016. "From posting to voting: The effects of political competition on online political engagement." *Political Science Research and Methods* 4(2):361–378.

Silva, Bruno Castanho and Sven-Oliver Proksch. 2021. "Fake it 'til you make it: a natural experiment to identify European politicians' benefit from Twitter bots." *American Political Science Review* 115(1):316–322.

Skytte, Rasmus. 2022. "Degrees of Disrespect: How Only Extreme and Rare Incivility Alienates the Base." *The Journal of Politics* 84(3):1746–1759.

Sobolev, Anton, M Keith Chen, Jungseock Joo and Zachary C Steinert-Threlkeld. 2020. "News and geolocated social media accurately measure protest size variation." *American Political Science Review* 114(4):1343–1351.

Stier, Sebastian, Arnim Bleier, Haiko Lietz and Markus Strohmaier. 2018. "Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter." *Political communication* 35(1):50–74.

Stukal, Denis, Sergey Sanovich, Richard Bonneau and Joshua A Tucker. 2022. "Why botter: how pro-government bots fight opposition in Russia." *American political science review* 116(3):843–857.

Temporão, Mickael, Corentin Vande Kerckhove, Clifton van Der Linden, Yannick Dufresne and Julien M Hendrickx. 2018. "Ideological scaling of social media users: a dynamic lexicon approach." *Political Analysis* 26(4):457–473.

Yarchi, Moran, Christian Baden and Neta Kligler-Vilenchik. 2021. "Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media." *Political Communication* 38(1-2):98–139.