The End of the Rehydration Era
The Problem of Sharing Harmful Twitter Research Data

Dennis Assenmacher, Indira Sen, Leon Fröhling, Claudia Wagner
GESIS - Leibniz Institute for the Social Sciences
{dennis.assenmacher, indira.sen, leon.froehling, claudia.wagner}@gesis.org

Abstract

Social media research is currently confronted with a data-sharing problem, as social media platforms prohibit full data distribution in their terms of service. Until recent changes to the platform, Twitter was an exception, allowing academics to legally share Tweet and user IDs with peers, which could then be re-collected using the Academic API endpoints. This work investigates how Twitter data is currently shared in two domains of harmful online communication — abusive language and social bot detection. We find that the currently frequently utilized intermediate strategy of sharing Twitter IDs suffers from substantial data loss, leading to the incomparability of computational results. Moreover, recent changes in the API result in additional expenses and an increased collection time that may have an impact on the feasibility of research projects. All of these aspects further fuel the reproducibility crisis that social media analytics currently faces. To improve the current situation, we propose several best practices for research projects utilizing ID-based datasets for their experiments and provide recommendations for researchers who want to share their Twitter data with peers.

Introduction

Computational social scientists are concerned with investigating social and socio-technical phenomena through the utilization of computational methods and digital trace data, often from large social media platforms. Despite millions of new data instances being produced on these platforms within seconds, data access and distribution are currently inherently restricted by the platform's Terms of Services (ToS)[1]. Since some platforms are more restrictive than others, we currently observe an inherent platform bias in social media analytics (Tufekci 2014). Most of the current research endeavors in different Computational Social Science sub-domains, such as detecting abusive content, focus on data originating from the Twitter platform (Vidgen and Derczynski 2021). Twitter offers an external API that allows researchers to programmatically access the platform either by specifying certain queries or by monitoring a presumable random sample of the real-time stream of content posted on Twitter. However, according to Twitter's ToS, it is not allowed to share full data collections of these observations with other researchers publicly.

As discussed in Assenmacher et al. (2021), this leads to an inherent problem of limited reproducibility for research that relies on this type of social media data. Different approaches are prevalent in the community to circumvent the data-sharing problem (e.g., access via request). One method stands out in this context, especially for Twitter: `Rehydration`. Instead of sharing Tweets and associated meta-data directly, only unique Tweet identifiers (IDs) are shared. Using lists of IDs, the original Tweets and user objects can be retrieved from the Twitter API, allowing researchers to recreate the original data collection in theory. However, Twitter provides no access to Tweets or user profiles that have been removed from the platform (by the users themselves or removed by Twitter, e.g., because of a ToS violation). Investigating harmful online communication patterns in social media leads to a situation where a substantial amount of observations will become inaccessible over time, as this content is prone to be removed from the platform, either by the platform itself (moderation) or user-initiated deletion. When it comes to the evaluation of new methodological approaches (e.g., detection mechanisms for abusive language) and the comparison of empirical findings on social or socio-technical phenomena (e.g., fake news), this does not make for an acceptable status quo as it inherently hinders the reproducibility of research that is based on social media data. Recently - after Elon Musk's acquisition of Twitter - the platform announced fundamental changes to its APIs, including the Academic one.[2] ID-based Tweet retrieval will be strictly limited to 10,000 Tweets per month at a price-point of $100, introducing even more practical problems when it comes to the actual feasibility of rehydrating large corpora of Twitter data.

While the problem of data decay is already known to the research community (Yang et al. 2020; Pamungkas, Basile, and Patti 2023), it is comprehensively discussed only for general Twitter-based research (Zubiaga 2018) or controversial issues (Elmas 2023). A full discussion of the con-

[1] https://developer.twitter.com/en/developer-terms/agreement-and-policy

[2] https://developer.twitter.com/en/docs/twitter-api

sequences of data decay for the critical domains of abusive language and social bots is still missing. Furthermore, while there are several surveys about general characteristics of data quality of abusive language (Vidgen and Derczynski 2021; Pamungkas, Basile, and Patti 2023) and social bot datasets (Cresci 2020; Samper-Escalante et al. 2021; Hays et al. 2023), there is no focused analysis on data-sharing strategies and detailed information regarding the decay of Twitter-based datasets for harmful communication. We systematically shed light on the scale of the problem for two domains of harmful online communication that are particularly affected: social bots and abusive language. We review 45 existing datasets in both domains and reveal how Twitter data is usually shared and what extent of information loss is associated with the rehydration process.

Harmful Communication on Twitter

Social media platforms like Twitter, Instagram, and Facebook allow people worldwide to share content and communicate with small or large audiences, promoting the democratic concept of freedom of expression. However, these platforms and their users have become the subject of targeted harmful communication patterns alongside these significant benefits. As a result, recent research has concentrated on comprehending these hazards and devising machine learning-based countermeasures, such as techniques for identifying abusive language or disinformation.

This work focuses on two prominent research areas that aim to understand harmful communication patterns. Both heavily rely on Twitter data and require different meta-information: abusive language and social bot research. Abusive language (and related concepts like hate-speech) research is concerned with the identification of (mostly textual) content (Tweets) that *"attacks or diminishes, that incites violence or hate against groups, based on specific characteristics ... even in subtle forms or when humour is used"* (Fortuna and Nunes 2018). Social bots are described in research as automated actors (Twitter accounts) on social media websites that try to mimic human behavior and could potentially deceive others (Wagner et al. 2012). Sometimes social bots may even manipulate public opinion (Cresci 2020). While there is a controversial academic discussion on the degree of intelligence of such automatons, they are heavily researched, especially in the context of large-scale political manipulation, e.g., during elections in the US or the Brexit vote (Assenmacher et al. 2020; González-Bailón and De Domenico 2021). In contrast to abusive language, social bots are considered a transportation medium for harmful communication patterns. While abusive language detection mainly focuses on the Tweet level, social bot detection mechanisms often require more information and rely on detailed user/account data, including personal Tweet history.

While other types of harmful online communication such as fake news exist, a thorough investigation of all subcategories is outside the scope of this work. Therefore, we focus on the two prominent representatives relying heavily on Twitter data, covering both Tweet- and user-based datasets. We further argue that a specific focus on these two domains of harmful online communication - especially in the context of the recent change in the leadership of the platform - is justified, as there have been first results suggesting that changes in moderation practices have already led to measurable increases in the prevalence of both abusive language and bots on Twitter (Hickey et al. 2023).

Datasets

Abusive Language Datasets

Several abusive language datasets have been curated across many different languages and modalities. Vidgen and Derczynski (2021) carry out an extensive literature survey on abusive language, analyzing the characteristics of 63 datasets which led to the creation of hatespeechdata.com, a repository of hate speech and abusive langauge datasets. We use this frequently updated website (abbreviated as HSD in this work)[3] to aid in our search for finding prominent abusive language datasets.

HSD does not actually host the datasets themselves but curates a list of them with pointers to the original dataset as made available by the respective creators. As of 10.02.2022, there were 97 datasets listed in HSD. To narrow our scope, we collect the information of all English and German Twitter datasets amounting to 26 datasets, summarized in Table 1.[4] Despite its name, HSD enumerates hate speech datasets as well as datasets modeling related constructs like abusive and offensive language, harassment, sexism, and racism. All datasets are purely text-based; except the dataset by (Gomez et al. 2020) which contains potentially hateful memes and text-image pairs.

Social Bot Datasets

The domain of social bot research suffers from data scarcity. One of the central hubs for social bot datasets is the comprehensive data repository curated by the University of Indiana.[5] Besides providing a platform for datasets, they also developed one of the more popular bot classifiers called Botometer (formerly known as BotOrNot), which was improved over several iterations in recent years (Sayyadiharikandeh et al. 2020). Currently, 19 different datasets are hosted on their platform, ranging from verified collections of human user accounts over manually annotated political bot accounts to fake followers. The identification of bot accounts was achieved by different approaches, varying between datasets, ranging from manual labeling over honeypot deployment (Lee, Eoff, and Caverlee 2021) up to the purchase of content polluter accounts (Yang et al. 2019). All datasets in the Botometer repository are Twitter datasets predominantly modeling English-speaking users. To the best of our knowledge, no non-twitter social bot dataset is available to the research community, which further underlines the importance of research on sharing Twitter data. As previously mentioned, abusive language detection mainly focuses on the Tweet level, whereas social bot detection mechanisms

[3] https://github.com/leondz/hatespeechdata

[4] Some datasets like Ousidhoum et al. (2019) also have tweets in languages other than English or German.

[5] https://botometer.osome.iu.edu/bot-repository/datasets.html

Dataset	Lang	Size	Avail	Construct	Sharing	Hosting
`Waseem2016` (Waseem and Hovy 2016)	EN	6,909	91/54%	HS	IDs	GitHub
`WaseemHovy2016` (Waseem and Hovy 2016)	EN	16,791	64/50%	HS (sexism, racism)	IDs	GitHub
`BenevolentSexism` (Jha and Mamidi 2017)	EN	7,205	-/33%	HS (sexism)	IDs	GitHub
`Davidson2017` (Davidson et al. 2017)	EN	25,296	●	HS & OL	Tweets	GitHub
`Golbeck2017` (Golbeck et al. 2017)	EN	35,000	●	Harassment	Tweets on req	SH
`Ross2017` (Ross et al. 2016)	GER	477	●	HS (immigrants)	Tweets	GitHub
`Bohra2018` (Bohra et al. 2018)	EN,HI	4,067	69/69%	HS	IDs	GitHub
`ElSherief2018` (ElSherief et al. 2018)	EN	28,498	-/32%	HS	IDs	GitHub
`Founta2018` (Founta et al. 2018)	EN,HI	79,894	61/39%	HS	IDs / Tweets on req	SH / Zenodo
`GermEval2018` (Wiegand, Siegel, and Ruppenhofer 2018)	EN,HI	8,541	●	OL	Tweets	GitHub
`IberEval2018` (Fersini, Rosso, and Anzovino 2018)	EN	3,977	●	Sexism, misogyny	Tweets on req	SH
`Rezvan2018` (Rezvan et al. 2018)	EN/HI	24,189	●	Harassment	Tweets on req	SH
`Ribeiro2018` (Ribeiro et al. 2018)	EN	4,972	●	HS	Network(req)	GitHub / Kaggle
`HASOC19` (Mandl et al. 2020)	EN,HI,GER	7,005	●	HS & OL	Tweets	SH
`HatEval` (Basile et al. 2019)	EN,SP	13,000	●	HS	Tweets	SH
`OLID` (Zampieri et al. 2019)	EN	14,100	●	OL	Tweets	SH
`Ousidhoum2019` (Ousidhoum et al. 2019)	EN,AR	5,647	●	HS	Tweets	GitHub
`Toosi2019` (Toosi 2019)	EN	31,961	●	HS sentiment	Tweets	Kaggle
`ALONE` (Wijesiriwardene et al. 2020)	EN	688	●	HS/OL (Toxicity)	Tweets on req	SH
`Gomez2020` (Gomez et al. 2020)	EN	149,823	48/44%	Multimodal HS	Tweets, IDs, Img	SH
`MeTooMA` (Gautam et al. 2020)	EN	9,973	78/78%	HS	IDs	Dataverse
`CMSB` (Samory et al. 2021)	EN	2,743	87/85%	Sexism	Tweets on req	Datorium
`Covid2021` (Wich, Räther, and Groh 2021)	GER	4,960	78/70%	HS (Sexism)	IDs	GitHub
`SWAD` (Pamungkas, Basile, and Patti 2020)	EN	1,675	●	AL	Tweets	GitHub
`DeTox` (Demus et al. 2022)	GER	10,278	78/65%	HS/OL	IDs	GitHub
`LSHSCDT` (Toraman, Şahinuç, and Yilmaz 2022)	EN,TR	128,907	82/72%	HS	IDs	GitHub

Table 1: Overview of abusive language datasets analyzed. We refer to the datasets with the original names provided by the authors, or if not available the name of the first author + publication year. In terms of relative data availability (**Avail**), we differentiate between non-harmful/harmful observations. Datasets that are not shared via IDs are always 100% available (●). We also rehydrate those datasets that make an effort to share Tweet texts in addition to the Tweet ID, as this allows us to better study the data decay on Twitter.

Dataset	**Lang**	**Size**	**Avail**	**Construct**	**Sharing**	**Hosting**
`Caverlee2011` (Lee, Eoff, and Caverlee 2021)	EN	41,411	75/64%	Social Bots	ALL	SH
`Cresci2015` (Cresci et al. 2015)	EN	5,301	83/23%	Social Bots	ALL	SH / Boto
`Cresci2017` (Cresci et al. 2017)	EN	14,368	77/57%	Social Bots	U-ID, U-Obj., T-ID, T-Obj.	SH / Boto
`Gilani2017` (Gilani et al. 2017)	EN	2,614	89/88%	Social Bots	U-ID, U-Obj.	SH / Boto
`Varol2017` (Varol et al. 2017)	EN	2,573	81/80%	Social Bots	U-ID	Boto
`Celebrity` (Yang et al. 2019)	EN	5,970	95/-%	Social Bots	U-ID, U-Obj.	Boto
`CresciRtbust2019` (Mazza et al. 2019)	EN, IT	759	83/80%	Social Bots	U-ID, U-Obj., T-ID	Zenodo
`CresciStock2019` (Cresci et al. 2019)	EN	25,987	77/36%	Social Bots	U-ID, U-Obj., T-ID, T-Obj.	Zenodo / Boto
`Feedback` (Yang et al. 2019)	EN	527	78/65%	Social Bots	U-ID, U-Obj.	Boto
`Political` (Yang et al. 2019)	EN	62	-/21%	Social Bots	U-ID, U-Obj.	Boto
`Pronbots` (Yang et al. 2019)	EN	21,964	-/9%	Social Bots	U-ID, U-Obj.	GitHub / Boto
`VendorPurchased` (Yang et al. 2019)	EN	1,088	-/63%	Social Bots	U-ID, U-Obj.	Boto
`Astroturf` (Sayyadiharikandeh et al. 2020)	EN	585	-/31%	Social Bots	U-ID	Boto
`Botwiki` (Yang et al. 2020)	EN	704	-/91%	Social Bots	U-ID, U-Obj.	Boto
`Midterm` (Yang et al. 2020)	EN	5,0538	87/0%	Social Bots	U-ID, U-Obj.	Boto
`Rauchfleisch` (Rauchfleisch and Kaiser 2020)	EN, GER	1,971	94/90%	Social Bots	U-ID	Dataverse / Boto
`TwiBot20` (Feng et al. 2021)	EN	11,826	95/90%	Social Bots	ALL	SH
`Verified` (Yang et al. 2020)	EN	2,000	97/-%	Social Bots	U-ID, U-Obj.	Boto
`TwiBot22` (Feng et al. 2022)	EN	1,000,000	96/94%	Social Bots	ALL	SH

Table 2: Overview of bot datasets analyzed. We refer to the datasets with the original names provided by the authors, or if not available the name of the first author + publication year. In terms of relative data availability (**Avail**), we differentiate between non-bot/bot observations. In the column **Sharing**, *U-ID* and *U-Obj.* stand for user ID and user object, and *T-ID* and *T-Obj.* for Tweet ID and Tweet object, respectively. *ALL* is reserved for datasets that share network information in addition to the aforementioned information. We also rehydrate those datasets that make an effort to share user information in addition to the user ID, as it is oftentimes not the full user object available from the Twitter API that is shared, but a select subset relevant to the original dataset creation purpose. This furthermore allows us to better study the data decay on Twitter.

often require more information and rely on detailed account data. Early detection approaches only utilized a broad user feature set returned when user data was queried from Twitter (e.g., number of Tweets or friends). More state-of-the-art approaches incorporate additional information, such as the follower network and the account's Tweet content. All considered datasets and their characteristics are summarized in Table 2.

Results

Sharing Approaches We generally observe two major sharing patterns in both domains: either the Tweets are distributed via ID lists, or the Tweets' contents are published. In the latter case, much effort is put into the pseudonymization of the Tweets by removing the associated IDs and substituting user mentions and URLs (Zampieri et al. 2019). For all abusive language datasets that we identified in our study, 8 (31%) shared only Tweet IDs. The majority of the remaining datasets containing complete Tweets are often abusive language detection competitions such as HateEval, OLID, IberEVAL, and HASOC, which usually provide access only after registration. In this space, CodaLab[6], an open-source web-based competition platform, proved to be the *de facto* standard. The process of accessing data is mainly automated, and researchers are often required to sign a usage agreement that prohibits the further distribution of Tweets. For the remaining datasets that were not constructed in the context of a specific challenge, data access was granted on request, after contacting one of the work's authors. We positively report that most of the authors (11 out of 14 for abusive language, 6 out of 7 authors for social bots) we contacted replied within several days and provided us with links and passwords to the data repositories (e.g., Dropbox). This holds for both sharing approaches (IDs and Tweets). However, for the three abusive language datasets that only consisted of Tweets IDs, the authors either did not reply or replied saying that they could not make the missing data available, thus making it impossible to reconstruct the whole dataset.

For the social bot domain, the approach of sharing only IDs (here, the IDs of Twitter accounts) is even more prevalent, with only 31% of the analyzed datasets being completely available (including all meta-data required to replicate ML experiments). However, it is important to emphasize that an additional sharing approach can be observed in this domain. Instead of only providing references to the datasets, as HSD does for abusive language datasets, the Botometer website not only refers to these external sources but additionally hosts the datasets, consisting of both IDs and the corresponding user objects. These user objects usually contain meta-information such as the accounts screen name, number of followers, or number of Tweets posted by the account at the time of retrieval. While this seems to be an improvement over the sharing of IDs-only, which is always associated with the rehydration loss, it is still not a real solution to the reproducibility problem; most of the existing social bot detection mechanisms not only consider the user object as input features but also additional meta-data such as Tweet content or network characteristics. To replicate experimental results, it is important that all relevant information is shared. Examples of datasets for which all information that might be relevant for a classifier is made available are the `TwiBot`-datasets, which share both IDs and full objects of Tweets and users, as well as additional network information. Unfortunately, for the datasets that are hosted solely on Botometer, these additional input features are not available, as the authors raise concerns that the sharing of actual Tweet contents would violate Twitter's ToS.

Dataset Hosting and Availability We have identified a heterogeneous landscape of different hosting platforms across both investigated domains. Sharing strategies range from using centralized platforms like GitHub, Botometer, and Zenodo, to self-hosting through artisanal platforms or by sending the data (typically as a zipped folder) over email. For abusive language datasets, a majority of the datasets are shared via GitHub (14 out of 26), while self-hosting is the next popular option (9). When data is shared via GitHub, it is typically in the form of IDs with pointers to scripts for automatic rehydration. On the other hand, datasets containing tweets or other metadata are shared through self-hosted platforms. Self-hosting also allows dataset creators to make the data available if requesters sign a user agreement. In contrast, for social bot datasets, GitHub is rarely used for sharing bot data. Botometer is the most popular platform for sharing bot data in a standardized manner. Some notable dataset creators like Cresci et al. (2019), in addition to making the data available in the Botometer-friendly format of sharing only IDs and user objects, additionally share the data on their own, providing additional features like tweet IDs and contents or network information. While they followed this approach with a positive intention, it resulted in a dual state where it may not be clear which versions of the dataset were used by researchers for conducting their experiments.

However, using dedicated archives like Zenodo that comply with the ideas of open science and long-term availability are rather the exception than the norm, leading to a status-quo in which research depends on the responsiveness of the original dataset creators.

Rehydration Results We rehydrated 12 abusive language and 19 social bot datasets. Figure 1 summarizes the main results for each domain individually. We display the total availability of Tweets or user accounts on a dataset level in blue and additionally differentiate between the availability of harmful content/bots (red) and non-harmful content/non-bots (green). First, it is evident that a substantial amount of content is not accessible anymore in both domains, with a median availability of 75% for social bot data and 69% for abusive language data. Unsurprisingly this issue is even more amplified when focusing on the availability of the positive class (the harmful content/bot class). Here, we observe a median availability of 64% for the social bot data and only 54% for the abusive language data. We also observe a high variance in data available for rehydration across the different datasets, especially for the social bot domain. We assume that this is because not all bot accounts are harmful and are not prone to be removed from Twitter.

[6] https://codalab.org/

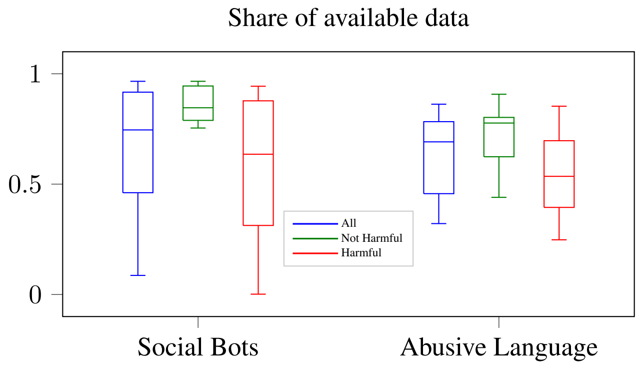

Figure 1: Rehydration results showing the fraction of available data per dataset.

A bot account that is explicitly identified as such and regularly shares weather information, for example, would not conflict with the Twitter ToS and should, therefore, not be expected to be removed. We observe that for social bot datasets, the highest rehydration loss is associated with data originating from a political context (`Political`, `Astroturf`, `Midterm`), while more general bot collections (`Varol2017`, `CresciRtbust2019`, `Botwiki`) tend to exhibit a higher account availability. The worst example of rehydration loss is found in the `Midterm` dataset, in which all accounts are permanently banned/removed from the platform, making it infeasible to recollect any data on social bots at all.

Zombie Tweets and Users We conducted our rehydration experiments at two distinct points in time to shed light not only on the at least anecdotally known phenomenon of Twitter data decay but also to be able to explore the underlying dynamics in greater detail. The first round of rehydrations was conducted in January and February 2022, and the second round was done more than a year later, in March 2023. While we mainly expected to observe that the decay of the datasets would have progressed with more time elapsed between the two rounds of rehydrations, the results are less clear (see Tables 3 and 4). In line with our expectation to observe signs of a progressing decay, we could not rehydrate a non-negligible share of objects in the second round that was still available in the first round of rehydrations. However, we also observe what we refer to as *Zombie Tweets* and *Zombie Users*, Tweets and users that we could not rehydrate in the first round but which were again available for rehydration during the second round. We thus find an effect that works in the opposite direction of the decay problem, thereby obfuscating its gravity. For the abusive language datasets, the decay of 11,859 Tweets that were available in Round 1 and are not available anymore in Round 2 is thus countered by 5,343 Tweets that were not available in Round 1 but could be rehydrated again in Round 2. What would have been equal to a further decay of 7.3% based on the total number of Tweets that were still available in Round 1 becomes a decay of only 4.0% if simultaneously accounting for the number of Tweets coming back online between rehydration rounds 1 and 2. On the user level and for the bot datasets, this effect also exists but is less pronounced for the Tweets, with 1,780 user accounts that were available in Round 1 not being available anymore in Round 2 and 263 user accounts that were not available in Round 1 being available again in Round 2. The reinstatement of previously banned accounts as announced by the `TwitterSafety` account between the two rehydration rounds could at least partially be responsible for the Zombie Tweets and Zombie Users observed here.[7] Together with banned users allowed back to the platform, their formerly removed Tweets would also have been made available again, potentially explaining the elevated numbers of both Tweets and users again available for rehydration in Round 2.

Table 3: Confusion matrix, showing the number of Tweets from the abusive language datasets that were available for rehydration in rounds 1 and 2.

	Avail R1	Not Avail R1
Avail R2	150,474	5,343
Not Avail R2	11,859	140,444

Table 4: Confusion matrix, showing the number of user accounts from the bot datasets that were available for rehydration in rounds 1 and 2.

	Avail R1	Not Avail R1
Avail R2	76,706	263
Not Avail R2	1,780	98,914

Pre-Sharing Rehydration Loss For the datasets made available via the Botometer repository, another dimension of the rehydration problem occurs. Some of the datasets included in the repository have originally been collected by other authors, before they were finally recollected by the maintainers of the Botometer repository, for the development of their own detection models and for inclusion in the repository. Because the other authors did not collect and share all the information necessary for the Botometer detection models, the user IDs had to be rehydrated before being added to the repository. However, since this round of rehydrations apparently happened a while later than the original creation of these datasets, the datasets included in the repository suffer from a (significant) loss of information for the exact same rehydration issue that is explored and discussed here. This is also why the dataset size presented in the publication that originally presented the dataset differs from the size of the dataset as introduced in the Botometer publications and the repository. This issue is acknowledged in Yang et al. (2020) and handled by sharing both the full lists of user IDs (as presented in the original datasets) and the list of user objects (as available when the dataset was rehydrated for inclusion in the repository). For some datasets, the loss in rehydration at this earlier stage is quite significant. While the

[7] https://twitter.com/TwitterSafety/status/1619125112716005376?s=20 (Accessed May 9, 2023)

number of user IDs with associated label available from the Botometer repository for the CresciStock2019 dataset match the 25,987 instances shared by the creator of the dataset, the number of user objects available from Botometer for the same dataset is only 13,276 – a loss of almost 50% of the original dataset instances even before the data is shared in the repository. Adding to the problem of the dual state of datasets already introduced above, this deviation in dataset size makes it even more challenging to keep track of the different dataset versions that are publicly available.

Rehydration Costs After Twitter's acquisition by Elon Musk, the platform underwent significant changes, including modifications to its API tiers. Prior to the acquisition, Twitter allowed researchers to use the platform's academic API to access up to 10,000,000 historic tweets (if still available) per month, completely free of charge. However, with the recent introduction of new API tiers, this has fundamentally changed.[8] While the new *Free* tier does not allow to retrieve Tweets from the platform at all, the *Basic* variant enables the rehydration of 10,000 Tweets per month at a cost of $100. These changes have significantly impacted the feasibility of benchmarking studies. For instance, if NLP researchers want to conduct a comprehensive model performance study on all abusive language datasets, they would need a budget of $4,500 to retrieve all the datasets for which only IDs were shared. However, budget constraints are not the only challenge hindering data-driven research. As the *Basic* tier only allows for the collection of 10,000 tweets per month, it would take 45 months to rehydrate all of the previously mentioned datasets (assuming that only one paid API developer account is used). This could become a major restriction for researchers working in small-sized research labs who do not wish to violate Twitter's ToS by creating multiple developer accounts.

Challenges and Best Practices

In our study, we identified several challenges in the domain of harmful online communication when it comes to sharing data via Twitter IDs. As shown in Figure 1, a significant amount of data decay occurs in both bot and abusive language datasets, with the decay affecting the harmful cases more. Therefore, the sharing of only IDs is not enough, as it leads to systematic data loss. We also observed other factors that hinder researchers from reproducing results from previous studies. Firstly, we cannot assume continuous data decay, as tweets or users that were previously unavailable may reappear on the platform at later times. Secondly, alterations to Twitter's API and pricing policy lead to higher financial and temporal costs for research institutions. Lastly, the existence of various versions of the same dataset across different hosting locations, rehydrated at different times and thus with different compositions, adds to the complexity of reproducing experimental results. Given Twitter's ToS, we acknowledge that sharing IDs is the *de facto* most straightforward and convenient way of sharing Twitter data. On the other hand, keeping in mind the importance of studying harmful communication, especially for its societal impact, there are valid reasons to question and reconsider Twitter's ToS (Freelon and Hargittai 2020). Sharing Tweets instead of IDs might seem like the logical solution, however, on top of violating Twitter's ToS, it can also lead to loss of privacy (Fiesler and Proferes 2018) and vulnerability to attack, especially in the context of sensitive topics like abusive communication.

Best practices for dataset creators

Keeping the previously mentioned challenges as well as ethical and privacy issues in mind, while attempting to balance transparency and reproducibility, we recommend dataset creators to do the following:

1. Releasing the full data via request that includes stipulations about protecting data subjects, such as conditions that require the data to not be shared further and the data subjects to not be contacted. Several papers in this survey do so (Golbeck et al. 2017; Rezvan et al. 2018)
2. Releasing the Tweet text (for abusive language datasets) or user information (their bios or network for social bots) in an anonymized way with personally identifiable information scrubbed
3. Releasing the features used for training automated methods with or without the actual Tweet texts or users
4. Releasing the data in a differentially private manner (D'Orazio, Honaker, and King 2015)
5. Releasing data only in one hosting location, preferably a dedicated and secure data archive (to avoid discrepancies in versioning)

A combination of two or more of these approaches is also possible (Samory et al. 2021). While all five options are more time-consuming for dataset creators, rather than just releasing lists of IDs, and are not without limitations of their own (Oberski and Kreuter 2020), they better ensure both reproducibility and the welfare of data subjects. Open science platforms also reduce reliance on individual researchers for data access. Finally, we recommend that new studies utilizing existing datasets should clearly state (a) whether and which subset of IDs was used in the context of their experiments and (b) which additional information was used (e.g. in the bot domain the most recent Tweets for each account).

Best practices for dataset users

We recommend the following steps for anyone intending to use a Twitter-based dataset: If the dataset is only shared via Tweet IDs, it is advisable to contact the original dataset creators to explore potential collaboration possibilities before embarking on any rehydration efforts. In our investigations, we found that a substantial majority of the dataset creators were responsive and often granted us access to their data. However, if gaining access to the original data is not feasible, researchers should report which Tweet IDs were successfully rehydrated and used for subsequent experiments. This reporting ensures a certain level of comparability between different studies. It is essential to document how and

[8] https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api (Accessed April 12, 2023)

from which source the experimental data was retrieved in all cases.

Discussion and Outlook

Sharing Tweet and user IDs that researchers not involved with the original data collection effort can potentially rehydrate appears to be a promising solution for the data-sharing issues posed by Twitter's ToS. However, this work showed that a significant amount of data becomes inaccessible over time, especially when harmful content is concerned. The resulting discrepancy between different ground truth variants leads to a situation where evaluation results are no longer comparable. An important factor is keeping in mind the "right to be forgotten" enshrined in privacy legislature such as the General Data Protection Regulation (GDPR) (Tsesis 2014); one that implies that we should respect a data subject's wish to have their content removed from data stores when they delete that content on a platform. However, building on previous calls for closer examination of trade-offs between the right to be forgotten and transparency (Freelon and Hargittai 2020; Tromble and Stockmann 2017), our work raises important questions on how data can be shared for research topics of great societal importance, such as harmful communication.

Ultimately, we want to emphasize that the current situation is not caused by the bad research habits of academics who want to keep their data private. On the contrary, academics are eager to share their datasets but often cannot comply with concerns. Despite the good intentions that academics have when releasing their valuable data as lists of IDs, we argue that it might be counterproductive due to dataset decay as well as the prohibitive monetary and temporal costs of rehydration. It is, therefore, of uttermost importance that researchers systematically elaborate on which information should be shared with peers depending on the underlying problem. Finally, it is important that both researchers and Twitter work together to facilitate data access for academics.

Limitations and Future Work Our work is not without limitations, and there are important follow-up research questions that have to be addressed in upcoming endeavors. First, we only investigated two sub-domains of harmful online communication. While both domains are important dimensions of harmful online communication, especially when it comes to research that heavily relies on Twitter data, future work might investigate other relevant areas, such as misinformation or propaganda detection. It is also vital to analyze languages other than German and English. Future work should also investigate and characterize the content that was removed from Twitter, or perhaps even more interestingly, harmful communication that has been spared because of either mislabeling, because it was too subtle to be detected by internal Twitter detection mechanisms, or because it was not flagged as harmful by other Twitter users. Finally, it would be interesting to investigate how the extent of missing data and the impact of different decay rates for harmful and non-harmful content on the class (im-)balance affects the performance of trained models.

References

Assenmacher, D.; Clever, L.; Frischlich, L.; Quandt, T.; Trautmann, H.; and Grimme, C. 2020. Demystifying social bots: On the intelligence of automated social media actors. *Social Media + Society*.

Assenmacher, D.; Weber, D.; Preuss, M.; Valdez, A. C.; Bradshaw, A.; Ross, B.; Cresci, S.; Trautmann, H.; Neumann, F.; and Grimme, C. 2021. Benchmarking crisis in social media analytics: A solution for the data-sharing problem. *Social Science Computer Review*.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel Pardo, F. M.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Bohra, A.; Vijay, D.; Singh, V.; Akhtar, S. S.; and Shrivastava, M. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. WWW '17 Companion. International World Wide Web Conferences Steering Committee.

Cresci, S.; Lillo, F.; Regoli, D.; Tardelli, S.; and Tesconi, M. 2019. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Trans. Web*.

Cresci, S. 2020. A decade of social bot detection. *Commun. ACM*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Demus, C.; Pitz, J.; Schütz, M.; Probol, N.; Siegel, M.; and Labudde, D. 2022. Detox: A comprehensive dataset for german offensive language and conversation analysis. In *Proceedings of the 6th Workshop on Online Abuse and Harms (WOAH 2022), Association for Computational Linguistics, Online*, 54–61.

D'Orazio, V.; Honaker, J.; and King, G. 2015. Differential privacy for social science inference. *Sloan Foundation Economics Research Paper*.

Elmas, T. 2023. The impact of data persistence bias on social media studies. In *Proceedings of the 15th ACM Web Science Conference 2023*, 196–207.

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018. Peer to peer hate: Hate speech instigators and their targets. *Intl AAAI Conf. Web and Social Media*.

Feng, S.; Wan, H.; Wang, N.; Li, J.; and Luo, M. 2021. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM International Con-*

ference on Information & Knowledge Management, 4485–4494.

Feng, S.; Tan, Z.; Wan, H.; Wang, N.; Chen, Z.; Zhang, B.; Zheng, Q.; Zhang, W.; Lei, Z.; Yang, S.; et al. 2022. Twibot-22: Towards graph-based twitter bot detection. *arXiv preprint arXiv:2206.04564*.

Fersini, E.; Rosso, P.; and Anzovino, M. E. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Fiesler, C., and Proferes, N. 2018. “participant” perceptions of twitter research ethics. *Social Media+ Society*.

Fortuna, P., and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51(4).

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Intl AAAI Conf. Web and Social Media*.

Freelon, D., and Hargittai, E. 2020. Chapter one when social media data disappear. In *Research Exposed*. Columbia University Press. 6–29.

Gautam, A.; Mathur, P.; Gosangi, R.; Mahata, D.; Sawhney, R.; and Shah, R. R. 2020. #metooma: Multi-aspect annotations of tweets related to the metoo movement. *Intl AAAI Conf. Web and Social Media*.

Gilani, Z.; Farahbakhsh, R.; Tyson, G.; Wang, L.; and Crowcroft, J. 2017. Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM ’17. Association for Computing Machinery.

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; Gergory, Q.; Gnanasekaran, R. K.; Gunasekaran, R. R.; Hoffman, K. M.; Hottle, J.; Jienjitlert, V.; Khare, S.; Lau, R.; Martindale, M. J.; Naik, S.; Nixon, H. L.; Ramachandran, P.; Rogers, K. M.; Rogers, L.; Sarin, M. S.; Shahane, G.; Thanki, J.; Vengataraman, P.; Wan, Z.; and Wu, D. M. 2017. A large labeled corpus for online harassment research. In *Web Science*.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

González-Bailón, S., and De Domenico, M. 2021. Bots are less central than verified accounts during contentious political events. *Proceedings of the National Academy of Sciences* 118(11).

Hays, C.; Schutzman, Z.; Raghavan, M.; Walk, E.; and Zimmer, P. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. *arXiv preprint arXiv:2301.07015*.

Hickey, D.; Schmitz, M.; Fessler, D.; Smaldino, P.; Muric, G.; and Burghardt, K. 2023. Auditing elon musk’s impact on hate speech and bots. *arXiv preprint arXiv:2304.04129*.

Jha, A., and Mamidi, R. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Workshop on NLP and Computational Social Science*.

Lee, K.; Eoff, B.; and Caverlee, J. 2021. Seven months with the devils: A long-term study of content polluters on twitter. *ICWSM*.

Mandl, T.; Modha, S.; Kumar M, A.; and Chakravarthi, B. R. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Forum for Information Retrieval Evaluation*.

Mazza, M.; Cresci, S.; Avvenuti, M.; Quattrociocchi, W.; and Tesconi, M. 2019. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Web Science*.

Oberski, D. L., and Kreuter, F. 2020. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review: HDSR*.

Ousidhoum, N.; Lin, Z.; Zhang, H.; Song, Y.; and Yeung, D.-Y. 2019. Multilingual and multi-aspect hate speech analysis. In *Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *Language Resources and Evaluation Conference*. European Language Resources Association.

Pamungkas, E. W.; Basile, V.; and Patti, V. 2023. Towards multidomain and multilingual abusive language detection: a survey. *Personal and Ubiquitous Computing* 27(1):17–43.

Rauchfleisch, A., and Kaiser, J. 2020. The false positive problem of automatic bot detection in social science research. *PLOS ONE* (10).

Rezvan, M.; Shekarpour, S.; Balasuriya, L.; Thirunarayan, K.; Shalin, V. L.; and Sheth, A. 2018. A quality type-aware annotated corpus and lexicon for harassment research. In *Web Science*.

Ribeiro, M.; Calais, P.; Santos, Y.; Almeida, V.; and Meira Jr., W. 2018. Characterizing and detecting hateful users on twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 12(1).

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *3rd Workshop on Natural Language Processing for Computer-Mediated Communication/Social Media*, 6–9.

Samory, M.; Sen, I.; Kohne, J.; Flöck, F.; and Wagner, C. 2021. Call me sexist, but...: Revisiting sexism detection using psychological scales and adversarial samples. In *ICWSM*.

Samper-Escalante, L. D.; Loyola-González, O.; Monroy, R.; and Medina-Pérez, M. A. 2021. Bot datasets on twitter: Analysis and challenges. *Applied Sciences*.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Conference on information & knowledge management*.

Toosi, A. 2019. Twitter sentiment analysis.

Toraman, C.; Şahinuç, F.; and Yilmaz, E. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2215–2225. Marseille, France: European Language Resources Association.

Tromble, R., and Stockmann, D. 2017. Lost umbrellas: Bias and the right to be forgotten in social media research. *Internet research ethics for the social age: New challenges, cases, and contexts* 75–91.

Tsesis, A. 2014. The right to erasure: Privacy, data brokers, and the indefinite retention of data. *Wake Forest L. Rev.*

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM*.

Varol, O.; Ferrara, E.; Davis, C. B.; Menczer, F.; and Flammini, A. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM*.

Vidgen, B., and Derczynski, L. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*.

Wagner, C.; Mitter, S.; Körner, C.; and Strohmaier, M. 2012. When social bots attack: Modeling susceptibility of users in online social networks. In *2nd workshop on Making Sense of Microposts at WWW2012*.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Wich, M.; Räther, S.; and Groh, G. 2021. German abusive language dataset with focus on COVID-19. In *Conference on Natural Language Processing (KONVENS 2021)*.

Wiegand, M.; Siegel, M.; and Ruppenhofer, J. 2018. Overview of the germeval 2018 shared task on the identification of offensive language.

Wijesiriwardene, T.; Inan, H.; Kursuncu, U.; Gaur, M.; Shalin, V. L.; Thirunarayan, K.; Sheth, A.; and Arpinar, I. B. 2020. Alone: A dataset for toxic behavior among adolescents on twitter. In *International Conference on Social Informatics*.

Yang, K.-C.; Varol, O.; Davis, C. A.; Ferrara, E.; Flammini, A.; and Menczer, F. 2019. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*.

Yang, K.-C.; Varol, O.; Hui, P.-M.; and Menczer, F. 2020. Scalable and generalizable social bot detection through data selection. In *ICWSM*.

Zampieri, M.; Malmasi, S.; Nakov, P.; Rosenthal, S.; Farra, N.; and Kumar, R. 2019. Predicting the type and target of offensive posts in social media. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zubiaga, A. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*.