This Sample Seems to Be Good Enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API

Jürgen Pfeffer[1], Angelina Mooseder[1], Jana Lasser[2], Luca Hammer[3],
Oliver Stritzel[4], David Garcia[2]

[1] School of Social Science and Technology, Technical University of Munich, Germany
[2] Faculty of Computer Science and Biomedical Engineering, Graz University of Technology, Austria
[3] Department of Media Studies, University of Siegen, Germany
[4] Faculty of Informatics, Technical University of Vienna, Austria

Abstract

Because of its willingness to share data with academia and industry, Twitter has been the primary social media platform for scientific research as well as for consulting businesses and governments in the last decade. In recent years, a series of publications have studied and criticized Twitter's APIs and Twitter has partially adapted its existing data streams. The newest *Twitter API for Academic Research* allows to "access Twitter's real-time and historical public data with additional features and functionality that support collecting more precise, complete, and unbiased datasets."[1] The main new feature of this API is the possibility of accessing the full archive of all historic Tweets. In this article, we will take a closer look at the Academic API and will try to answer two questions. First, are the datasets collected with the Academic API complete? Secondly, since Twitter's Academic API delivers historic Tweets as represented on Twitter at the time of data collection, we need to understand how much data is lost over time due to Tweet and account removal from the platform. Our work shows evidence that Twitter's Academic API can indeed create (almost) complete samples of Twitter data based on a wide variety of search terms. We also provide evidence that Twitter's data endpoint v2 delivers better samples than the previously used endpoint v1.1. Furthermore, collecting Tweets with the Academic API at the time of studying a phenomenon rather than creating local archives of stored Tweets, allows for a straightforward way of following Twitter's developer agreement. Finally, we will also discuss technical artifacts and implications of the Academic API. We hope that our work can add another layer of understanding of Twitter data collections leading to more reliable studies of human behavior via social media data.

Introduction

In 2013, David Lazer opened his ICWSM (Emre Kiciman et al. 2013) keynote address with the joke "Welcome to the 7th International Conference on Twitter Analysis", referring to the fact that a majority of published papers at this conference were working with Twitter data. While a series of other social media platforms have attracted the attention of academic researchers in recent years, Twitter has remained the dominant data source for social media analysis of human behavior due to its data accessibility via a variety of application programming interfaces (API). Several review articles have tried to create an overview of the countless number of studies that have utilized Twitter data (Yu and Muñoz-Justicia 2020; Williams, Terras, and Warwick 2013; Zimmer and Proferes 2014; Jungherr 2015; Sinnenberg et al. 2017; Karami et al. 2020; Pradyumn, Kapoor, and Tabrizi 2018).

For most of the data driven studies in recent years, two main data access strategies have been applied. Tweets were collected either with filtering queries (e.g. keywords, accounts, geo-locations) or as random 1% or 10% samples of all Tweets. Both data collection strategies relied on real-time data collection. With the Academic API, researchers are now able to perform full-archive searches, which has created a wide variety of new research possibilities.

In this article, we will build on the research that critically reflected Twitter's data sources and sampling algorithms. Similar to Zubiaga (Zubiaga 2018), we will also re-collect historic data. We will then use the re-collected data and the error messages that the Twitter API creates for missing Tweets to assess how many Tweets are actually *detectable* in the new data collection because they were neither deleted nor protected. We will also compare the collected data from the Academic API with data that was collected at the same time with the same search query via the costly Twitter Premium API. And we will perform controlled experiments to gain better insights into the sample creation process as well as into possible technical artifacts. Finally, we will discuss the implications of using the Academic API for studying Twitter activity.

Contributions. The goal of this article is not to reverse engineer Twitter's hardware and software architecture. However, we do think that a certain level of technical insight into the data production process is necessary to assess the quality of the data source. The contributions of the article are:

- We provide background information for researchers interested in using Twitter's Academic API.

- We show evidence that Twitter's Academic API can provide almost complete samples of the overall activity on Twitter related to specific keywords and time periods.

- We argue that Twitter's Academic API utilizing the new API v2 data endpoint delivers the best data samples from

[1]https://developer.twitter.com/en/products/twitter-api/academic-research

Twitter for most application scenarios.

- We study the proportion of Tweets that get lost over time when historic Twitter data is collected with full-archive searches.
- We will discuss technical artifacts and implications when collecting data with the Academic API.
- We offer a set of recommendations for researchers working with the Academic API in their research studies.

Related Work

Twitter's Academic API. Although the Twitter Academic API is relatively new, it has been used in multiple studies. Regarding the times in which the Twitter's Academic API has been released, it is not surprising that many of these studies cover Covid-19 related topics. Before the Academic API was released, researchers had to use services like Brandwatch to analyze exhaustive datasets around the pandemic (Pellert et al. 2020; Metzler et al. 2021). More recently, Aryal and Bhattarai (2021) used the API to collect Tweets by hashtags and geolocation and study sentiments towards Covid-19 vaccination across several Asian countries. Similarly, Butsos et al. (2022) collected Tweets by hashtags to analyze the sentiments and emotions towards Covid-19 vaccination before and after the first launch of a Covid vaccination in the US. Marcec and Likic (2021) compared the sentiments of English Tweets towards different Covid-19 vaccines over time and location, using the Academic API to collect Tweets by keywords, while Biswas et al. (2022) collected Tweets by vaccine-related Arabic hashtags and analysed the sentiments of Arabic Tweets towards vaccination in general before and during Covid-19.

A somewhat different data collection approach was applied by Muric et al. (2021). They first searched antivaccine related user accounts by employing a keyword search on Twitter's Streaming API. Afterwards they used the Academic API to collect historical Tweets of these accounts. Based on this data, they studied the topics and time distribution of anti-Covid-vaccination Tweets as well as the political leaning of accounts publishing such Tweets.

Besides Covid-19 related studies, the Academic API has been applied in a great variety of research topics. Using searches based on hashtags or keywords, data was collected via the API to study topics of haemophilia-related Tweets and categorize users tweeting such content (Chen, Muralidharan, and Samelson-Jones 2022); analyse the time distribution and user network of Tweets about an event of a false terrorist alarm at Oxford circus 2017 (Eriksson Krutrök and Lindgren 2022); examine the topics, hashtags and mentions of polish Tweets about a Polish women's strike at two different timeframes (Paradowski 2021); investigate the topics of and user engagement in Tweets about women's day across three countries (Wallaschek et al. 2022) and research the topics and time distribution of Tweets about the h-index (Thelwall and Kousha 2021).

Sampling quality and technological insights. Working with social media data is known to be challenged by different artifacts coming from data or applied methods (Ruths and Pfeffer 2014). In the past, researchers took a close look into the quality of Twitter data and possible errors, which might arise during data collection (Salvatore, Biffignandi, and Bianchi 2021). They found that the choice of the API and the collection technique can heavily influence the accuracy, representativeness, and reproducibility of the dataset.

A core challenge when using Twitter APIs is how to build a data sample which is representative of the overall activity on Twitter. Moerstaetter et al. (2013) compared datasets collected with the Streaming API and Firehose based on keywords, geolocation, and user account. They found that the representativeness of data collected with the Streaming API heavily depends on the coverage and type of analysis. For example, the results showed that the Streaming API performs worse than random samples of the Firehose datasets when coverage is low, but topical analysis results were similar between the Streaming API and Firehose datasets when coverage was high. Likewise, Wang et al. (2015) used the Twitter Streaming API with Spritzer and Gardenhose access to collect Tweets of specific user accounts and compare them to a dataset of all Tweets of these users collected with the REST API. Their results indicated that data collected with the Streaming API provides enough information to analyse general or content statistics, but only covers small proportions of user interaction, such as mentions or Retweets. Furthermore, Kim et al. (2020) compared datasets collected with the filtered Streaming API, Search API and GNIP historic Powertrack (Firehose). They found that the volume, content and accounts of Tweets within these datasets varied substantially and that the use of the historic API may result in an underestimation of the amount of marketing and bot-generated Tweets and accounts. Also, Alkulaib et al. (2020) compared datasets of event-related Tweets collected with the Streaming API, Search API and Firehose API and found only a small overlap between these datasets. They suggested that a combination of multiple APIs could results in a more representative dataset. This strategy was also recommended by Timoneda (2018), who compared datasets collected with the Streaming and the Search API based on keywords and found that while 20-30% of Tweets did not appear in the Search API, also on average between 2 and 5% of Tweets were lost during data collection with the Streaming API. Similarly, different collection tools can influence the resulting data sample. For example, Weber et al. (2021) compared datasets collected with Twarc, RAPID and Tweepy, all based on the filtered Streaming Twitter API, and found considerable differences between these datasets, which significantly altered the results of subsequent network analyses. Wu et al. (2020) studied Twitter sampling effects across different timescales and different subjects by analyzing the rate limit messages that indicate the cumulative number of missing tweets for an API connection.

One major problem is that there has been little information from Twitter about how data is sampled by the APIs and that biases might be introduced by sampling mechanisms (Chen, Duan, and Yang 2021). Kergel et al. (2014) were instrumental in revealing the sampling procedure of the 1% and the 10% Sample API. They were able to show meta-data extracted from the Tweet ID can tell us whether a Tweet is in one of Twitter's Sample streams or not. Another question is,

whether data collected with Twitter's APIs are reproducible. Zubiaga (2018) collected 30 datasets in 2012-2016 using the Twitter Streaming API. In 2016, he recollected the data, using the API's statuses/lookup method based on Tweet id, to analyze how much of the data can be reproduced. He found a substantial loss of Tweets and user accounts, with some datasets having more than 30% of data that could not be reproduced, as well as a change in aggregated user metadata. Nevertheless, he showed that the datasets were still representative in regard to textual content.

Twitter's Data Streams

Twitter's data APIs. Based on underlying functionality, Twitter offers different types of APIs for accessing data. First, the Filter API collects Tweets which include user defined search terms (e.g. keywords, accounts). Second, the Sample API provides a sample (1% or 10%) of *all* Tweets. It is important to understand that the Sample and the Filter API (together often referred to as the Streaming API) offer real-time data collection, i.e. the Tweets of our data collection are coming as a *stream* through the API immediately after they have been sent. In contrast, the REST and the Search API can be used to collect historic Tweets and other non-real-time information. The REST (Representational State Transfer) API can be used to collect information about accounts and their followers and followees or to send Tweets. Finally, Search APIs (including the Academic API) can search for historic Tweets (either only some days back or all the way back to the first Tweets in 2006). Several of these APIs offer different access levels, e.g., Standard, Premium, Academic, Enterprise, with elevated access restricted to specific user groups or paying clients.

Data endpoints. The Academic API was introduced as part of a new data endpoint, that Twitter calls its API v2. While there are several minor differences, the major difference when collecting data is that with the old API v1.1, a Tweet in JSON format always arrived as a single complete object, i.e. many of the Tweet's fields are delivered automatically. The endpoint API v2 just delivers the id and the text field of a Tweet automatically; all other fields need to be requested via call parameters. Furthermore, account information of the sender and other meta data comes in a separate JSON object. Another obvious difference between these two data endpoints is that users traditionally needed an authentication key consisting of four parts with the v1.1 endpoints while a single longer *bearer token* key is needed for the v2 endpoint.

Implications when collecting historic Tweets. In recent years, most researchers had collected Twitter data with the real-time sample and filter APIs. By utilizing the Academic API, the data collections strategy changes, which comes with certain implications that we will discuss in this paper. First and foremost, when collecting historic Tweets with the Academic API, we will get a representation of Twitter as of the collection time and *not* the time when the Tweets were created. Simply speaking, the Academic API can access the same information that you can access at Twitter's website (e.g. via the search feature). This also implies that Tweets

that have been deleted by users or when an account is removed (e.g. the suspended account @realdonaldtrump) from Twitter, will not be part of a dataset collected with the Academic API.

Another very important aspect comes from the fact that Twitter seems to not store information about the sender of a Tweet (except the account ID) with the Tweet. Consequently, when you collect historic Tweets and also include user fields, these fields are requested from the user database at the time when the Tweet is collected and not at the time when the Tweet was created. This means that the user profile information and even the user name (but not the user ID) could have changed in the meantime.

Experiments to Assess Data Quality

In this section, we present the setup and results of several experiments that were designed to assess the sampling quality of the Academic API. We do this by studying the two key aspects of this API, the full-archive search and the data endpoint v2. Because we do not have direct access to Twitter's database of Tweets, we need to design different experiments to get different insights into the software and hardware black box. However, we start with a controlled experiment for which we actually know the ground truth.

Controlled Experiment

We performed a controlled experiment to identify possible obvious issues with the data streams. With a set of ten Twitter accounts, we sent $3,400$ Tweets within 2.5 hours—one Tweet every 2.5 seconds—with a single random $42-$character string. We had setup two filter API scripts listening for Tweets with the same string, one each utilizing the v1.1 and the v2 data endpoint. After sending the Tweets, we also collected the Tweets from the previous hours with the Academic API, again with a query searching for the specific string. We repeated this search with the Academic API 24 hours later. For all four data collection approaches, we can report that 100% of all Tweets were collected. In other words, if your search queries include a smaller number of Tweets loosely distributed over time, there is a very good chance that you will receive a very high coverage, most likely with all of Twitter's APIs and endpoints.

High Volume Topics

In a subsequent experiment, we were interested in the performance of these data collection approaches when dealing with a topic with a high volume of Tweets. Since our study was conducted during the Russian war in Ukraine, we searched for the term "Ukraine" for exactly five minutes via the v1.1 and v2 data endpoints in parallel on the same computer. We are aware that Twitter tries to prevent a single query collecting a very large number of Tweets and, consequently, inserts rate limits into the Filter API. Independently from rate limits, our primary interest is in coverage, i.e. more is better. Right after the period of live data collection, we collected "Ukraine" Tweets with the Academic API for the exact same five minutes. Surprisingly, the results were very different with this setup (see Figure 1).

While the numbers between the live collection and the historic collection with the Academic API cannot be compared directly in terms of coverage, since Tweets can be protected (made *invisible*) and unprotected or deleted, some of the numbers in figure 1 are surprising. The dynamics as well as different rate limits result in the fact that only 72.8% of Tweets collected with these three approaches occur in all datasets. The difference between the two Filter APIs indicates that the two data endpoints in fact produce different datasets and that the endpoint v2 creates a larger dataset. The relatively high number of Tweets that only appear in the Filter APIs are strong indicators that a significant Twitter volume gets removed from the platform quickly, either because users delete or hide their Tweets or Twitter removes Tweets because of violations of community norms (e.g. toxic language). Another very interesting number in this Venn diagram is the number 0 on the intersection of *Academic v2* and *Filter v1.1*. This means that there was not a single Tweet that was picked up by the *Filter v1.1* that was still available for the later data collection with the *Academic v2* but had not already been picked up by the *Filter v2*. More proof that the two data endpoints are technically different and another indicator that the endpoint v2 is superior because of the higher number of delivered Tweets.
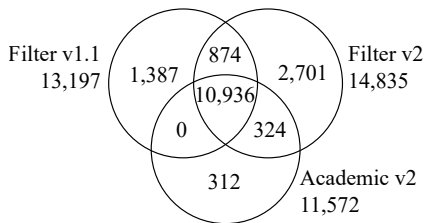

Figure 1: Symbolic Venn diagram of the overlap of three different data sources in a high volume experiment.

Time delay. We can re-construct the arrival time in milliseconds of a Tweet at one of Twitter's servers (Pfeffer, Mayer, and Morstatter 2018; Kergl, Roedler, and Seeber 2014) from the Tweet ID. Having around $40 - 50$ Tweets per second in our datasets, we took a closer look at these timestamps and, despite possible time differences between Twitter's servers and our experimental computers, we made two observations:

1. With the Filter v1.1 API, a Tweet took on average 4.4 seconds $[4.3 - 5.7]$ to get to our computer, with 90% arriving within 4.7 seconds. In contrast, with the Filter v2 it took on average 10.2 seconds $[9.4 - 19.7]$ with almost 90% arriving within $10 + / - 0.5$ seconds so that we can assume that 10 seconds is Twitter's new intended time delay.

2. Tweets do not come in perfect temporal order at either data endpoint. However, with the Filter v1.1 API, less than 1% of Tweets arrived more than 0.2 seconds after a Tweet that they were supposed to precede, with no single Tweet changing the order for more than 1 second. With the Filter v2 API these numbers change dramatically. More than 40% of Tweets arrive more than 0.2 sec-

onds after a Tweet that they were supposed to precede, with a maximum of almost 9 seconds.

The first point is consistent with the minimum time period that needs to pass in order to search for historic Tweets with the Academic API. The second point seems to be another indication that the v2 data endpoint is in fact a different system architecture. We assume that these two aspects combined create a situation in which, for the Filter v2, more Tweets are allowed to get onto the API server before they are picked up by the *real-time* API requests.

We will not speculate here any further but, supported by more experimental runs with different search terms, we can conclude that in general, the Filter v2 API will return more Tweets.

Searching for Historic Tweets

Another test that we performed in order to assess the characteristics of the Academic API is to re-collect data that was collected $8 - 11$ years ago with the 10% Decahose Sample API when studying online firestorms (Lamba, Malik, and Pfeffer 2015). What we expected is that a) we would get much more data with the Academic API today and b) that there should be no Tweet in the old dataset that is not shown today and that has not been deleted or protected. Answering the first point is straightforward. When the dataset was collected in real-time with the then-used Twitter Decahose Sample API, $\mathcal{D} = 165,267$ Tweets were collected for the 20 online firestorms. When collecting Tweets again on March 6, 2022, with the same hashtag/user account queries for the same time periods in $2011 - 2014$ using the Academic API and utilizing Twitter's data API v2, we were able to collect $\mathcal{A}_1 = 799,033$ Tweets.

It is important to understand the above-described differences between these two different data collection methods. We can assume that $\mathcal{D}$ was about 10% of ~ 1.65 million Tweets (see Fig. 2). Comparing this number with $\mathcal{A}_1$ implies that slightly more than half of all Tweets from $\mathcal{D}$ cannot be found on Twitter anymore ~ 10 years later. However, this observation does not answer our question as to whether we were able to collect all Tweets in 2022 that are still on the platform from 2011-2014. Missing Tweets can be due to different reasons. Tweets can be deleted, they can be protected (i.e. not publicly visible), or they could be missing in our sample while still being on Twitter. To distinguish between these three cases, we first compared the two lists of Tweet IDs and identified the 90,908 Tweet IDs that were in the old dataset but were missing in the new dataset. In a second step, we tried to collect the Tweets of these missing Tweet IDs and take advantage of the fact that Twitter delivers different error messages for different reasons for Tweets being untraceable. For 72.9% (66,278 Tweets) we received a "not found error" indicating that a Tweet is no longer on the platform. For 26.6% (24,209 Tweets) we received an "authorization error" pointing to the situation that these Tweets being protected by their users at the time of data collection. Only 421 Tweets (0.5%) from $\mathcal{D}$ that were not in $\mathcal{A}_1$ could be found. These are Tweets that potentially were missed in the Academic API search. We repeated the data collection describe
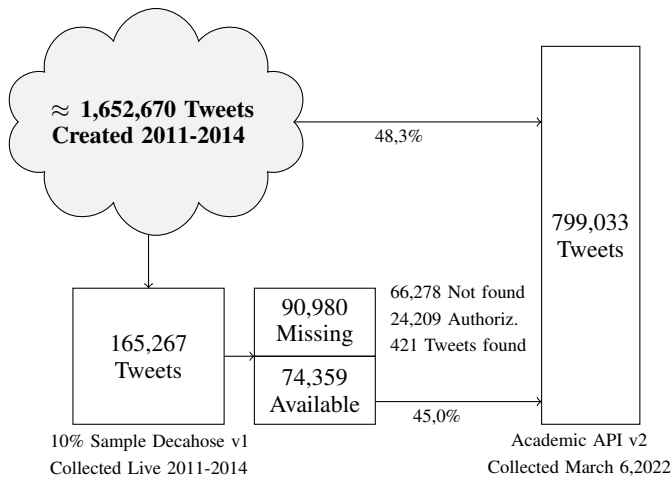
Figure 2: Overview of data samples for comparing Academic API data with historic Sample API data.

Figure 3: Experimental setup for testing Tweet decay over a ten day time period.

Tweet Data Decay

A major feature of the Academic API is that most researchers collect their data at a certain time after an event and not in real time. In the previous section, we could see that about half of all Tweets are lost after 10 years. In this section, we will first take a closer look at the first hours and days after Tweets were sent. Secondly, we will try to understand the long-term Tweet decay that follows a simple exponential decay function.

Our experimental setup for collecting data is depicted in Fig. 3. Our search term was constructed so that we were able (to the best of our knowledge) to collect all Tweets sent within a specific second. The first time-period for data collection was March 9, 2022[2] at 6:42:42am UTC collecting Tweets within the $[0 - 999]$ millisecond range of this second on Twitter. To not bias against individual time zones, we collected the same second of Twitter data every hour for 24 hours. The collection happened 10 seconds after each time window—due to API restrictions the time window of data collection must end a minimum of 10 seconds prior to the request time—and took 7–15 seconds each. We call this the day_0 data. On average, we collected 4,526 Tweets per 1 second of observation period over 24 hours within a range of [3,374–5,953].

On the following day, we re-collected the data from day_0 with the identical search query and time periods exactly 24 hours after the Tweets were created. This initial re-collection of day_0 data was then repeated for ten days to assess the Tweet decay ratio per day. Fig. 4 plots the results of our experimental data collection. The major drop in Tweet availability (-6.1%) happens within the first 24 hours. This is a very intuitive result. Twitter users delete Tweets because of typos or other reasons; Tweets get flagged as toxic and removed from Twitter; etc. Starting with day_2, the daily decay becomes smaller and smaller and after ten days, about 10.8% of Tweets are not available anymore.

Short-Term vs. Long-Term Decay

To get a better estimate of the short-term (hours) and the long-term (months) proportions of Tweets that are lost for

above about 5 hours after the first collection. We performed the same search queries and received $\mathcal{A}_2 = 799,027$ Tweets, i.e. six Tweets less. It is highly likely that six out of the original almost 800k Tweets were deleted within five hours, e.g. via the removal of a user account. Because of the small proportion of missing Tweets, we did not follow this path any longer and conclude that the Academic API can successfully collect all (or at least almost all) Tweets for a search term and time period that are still on the platform at the time of data collection. This example also shows how, in general, historic Sample API data can be used to estimate the number of missing Tweets when collecting data with the Academic API.

Verifying Results with Twitter's Premium API

Another approach of assessing the quality of the dataset collected with the Academic API is by comparing it with a different data source that is supposed to extract the same dataset from Twitter. The Twitter Premium API is a paid-for service for enterprises that gives users access to the "the full history of Twitter data". We were able to access the Premium API and collected the same data as described above, at the same time of collecting $\mathcal{A}_1$. Interestingly, the Premium API delivered only $\mathcal{P} = 774,485$ Tweets. When comparing the Tweet IDs of the two data sources, 24,978 Tweets from $\mathcal{A}_1$ are missing in $\mathcal{P}$, while only 430 Tweets from $\mathcal{P}$ are missing in $\mathcal{D}_1$. At this point, we can only hypothesize that the API endpoint v1.1 has a different approach for searching in Twitter's archival data that leads to different results. Of course, there is a possibility that the endpoint v2 also delivers Tweets that should not be part of the queried result, but most likely the endpoint v1.1 misses Tweets with certain properties due storage or search characteristics—further research can look more closely into these differences. At this point, we can report that the Academic API delivers more Tweets in response to a historic search than the costly Premium API utilizing endpoint v1.1.
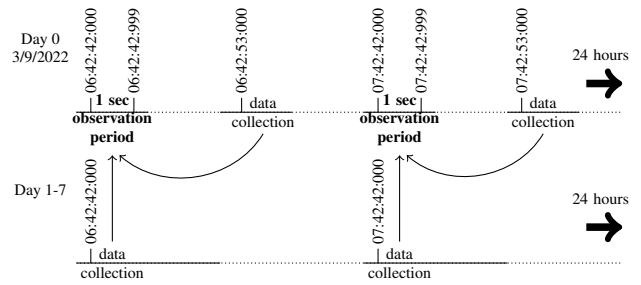
---

[2]The initial dataset was collected March 9 & 10, 2022—notably including the Russian war in Ukraine and Twitter's service blockage in Russia. Since we are not analyzing the content of the Tweets, these and other events should not impact the following analysis.
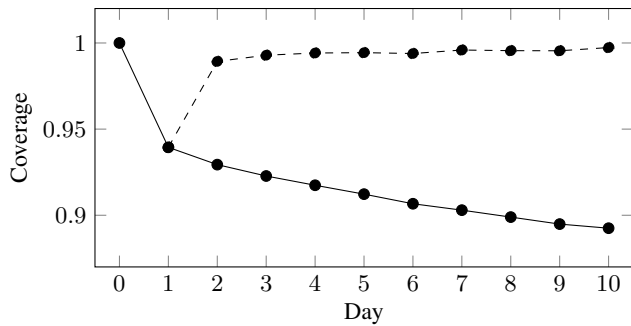
Figure 4: Coverage over time. Proportion of Tweets from day_0 data in day_n data (solid line). Proportion of Tweets from day_{n-1} data in day_n data (dashed line).

an Academic API search, we utilized archived IDs from Twitter's 1% data stream at *archive.com* and another repository (Lasser 2022) over a time span of 5 years. For every data point of the following analysis, we collected all Tweet IDs for the given time span (month, day or hour) and drew five random samples of $100k$ Tweet IDs each. Each sample is then re-hydrated using the twarc2[3] hydrate functionality. The number of available tweets ("found") and unavailable tweets ("not found" and "authorization error") are counted. The data points in the following figures are averages over the five samples for every point in time. The standard deviation for the points is so small ($<0.01\%$) that error bars for these data points are not visible. For the shorter time scales, we re-hydrated the data with a number of API keys in parallel, to reduce artifacts resulting from longer data collection time periods.

Figure 5 shows the results of this analysis by charting the proportion of Tweets that could be found n days after being sent, as well as the proportion of Tweets with *authorization error* and *not found* errors.

Our subsequent intention with these numbers was to fit a simple exponential decay function of available Tweets, so that the coverage of historic data collections can be estimated in future research. Estimating the proportion of avail-

---

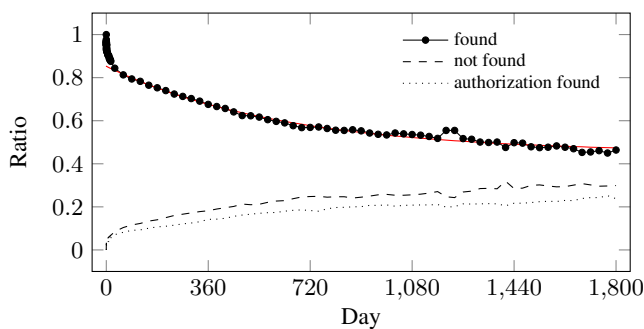[3]https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/

able Tweets can be an important indicator for assessing the fidelity of applied metrics. We have seen in Figure 4, that the daily decay slows down dramatically after the first day and that the proportion of Tweets that remain available from one day to the next reaches 99.7% on day 10. This proportion continues to grow. However, after one month 84.4% of Tweets could be found and this number is reduced on average by about 2% per month for the first 1.5 years. The rate of decay is then reduced to about 0.6% per month.

More precisely, the decay can be fit very well with an exponential decay function if we only consider data points of Tweets 1 month or older. For this month time scale, we employ a function of the form

$$f(x) = A \cdot exp(K \cdot x) + C. \tag{1}$$

With x being the age of Tweets in *days*, we performed a non-linear least squares fit weighted by the standard deviations for each data point to estimate the remaining variables. Since for our data, the constant C in eq. 1 can be interpreted as a lower boundary of Tweet decay, we also deployed a second model with a constant $C = 0$. Interestingly, the model with a variable C clearly creates a better fit by explaining an additional 8% of the variance, while having lower AIC, and even lower BIC, which is more sensitive to additional parameters. Consequently, the monthly decay can be best fit ($R^2 = 0.984, AIC = 260.4, BIC = 266.7$) with:

$$f(days) = 0.40 \cdot exp(-0.0017 \cdot days) + 0.46. \tag{2}$$

While a lower boundary for Tweet decay of $C = 0.46$ is a surprise, a relatively high C is not implausible. We have to keep in mind that at the timescale and time periods we study, we are only capturing some of the processes that can delete a tweet, while systemic events, e.g., a mass exodus of Twitter users to other new platforms, cannot be covered.

To better understand the differences between the short-time and long-time tweet inaccessibility, we also looked into the rate at which tweets vanish on the hour and day time scales (Figure 6). For the hour- and day-time scales, two linear unweighed functions (one for the first 24 hours and one for the first 15 days) showed to be the best fit.

Lastly, we investigated what drives the short-time scale tweet inaccessibility (see Fig. 7). A large number of Tweets

Figure 5: Proportion of Tweets after n days that are found and not found.
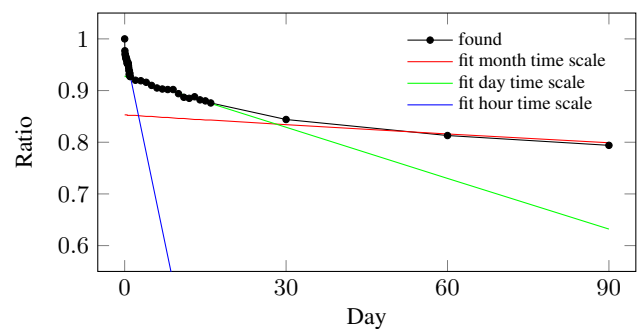

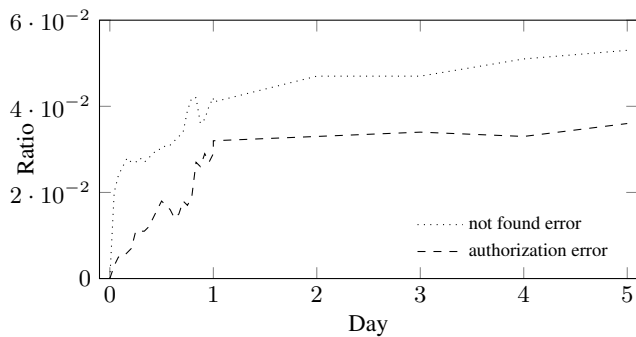Figure 6: Proportion of found tweets and additional fitting of short time (hours, days) data decay.

Figure 7: *Not found error* and *authorization error* per hour within the first 24 hours and per day within the first five days.

Missing	Center 10	11	13	Total
No	27,235	53,332	28,060	108,627
	27,294	53,284	28,049	
	-0.356	+0.209	+0.063	
Yes	1,047	1,881	1,005	3,933
	988	1,929	1,016	
	+1.870	-1.098	-0.33	
Total	**28,282**	**55,213**	**29,065**	**112,560**

Table 1: Cross tabulation of missing Tweets with data center IDs extracted from Tweet IDs: observed and expected number of Tweets, standard residuals.

(2–3%) vanish within the first hour by deletions (i.e. "not found" error). While we assume that many Tweets are deleted immediately by the author because of typos or on second thought, we also hypothesize that Twitter's recently improved efforts to limit toxicity on the platform may also contribute to this large number of quickly removed Tweets. Interestingly, the proportions of missing Tweets that have been deleted ($\bar{x} = 56.3\%, \sigma = 0.011$) and that have been protected ($\bar{x} = 43.7\%, \sigma = 0.011$) are stable over time.

Tweet Data Growth?

As we could show above, one out of 16 sent Tweets is not available anymore after 24 hours. Surprisingly, 3.5% of all Tweets from day_1 do not occur in the day_0 data. In other words, 1 out of 30 Tweets that we collect 24 hours after the Tweets where sent, were not available on the Academic API 10 seconds after they were sent. From our knowledge about Twitter's system, this can have two causes:
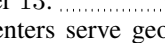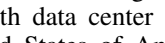
1. Private Tweets are created and then made publicly visible at a later time.

2. This is an unintended side effect from the technical infrastructure of Twitter.

Because of the high number of Tweets that are affected here, we can conclude that option 1 is very unlikely and chose to look closer at option 2. To assess the possibility of technical artifacts that lead to our missing data observations, we extracted some metadata from the Tweet IDs. Kergel et al. (2014) described in great detail the process of creating Tweet IDs. A Linux timestamp with milliseconds (with an offset starting with the first Tweet in 2006), a server ID, a worker ID (a server hosts multiple workers), and a sequence number for elements (created at the same millisecond from the same worker on the same server) are combined into a 64 bit number that then gets converted into the known Tweet ID form. This allows us to decode every Tweet ID back into these four variables. The milliseconds of the Tweet are not provided in the *created_at* variable when collecting Tweets with the API. For the following analysis we compared 108,627 Tweets from day_0 with 3,933 Tweets from day_1 that were not in the day_0 Tweet ID list and we follow the hypothesis that there is a technical reason that causes certain Tweets to arrive later at the API server and are, there-

fore, not collected right after they were created. We studied all aspects of meta-data coded into the Tweet ID.

Location. A cross tabulation of data centers with the two groups of Tweets shows us that there is no significant deviation from the expected numbers. Table **??** shows us that Twitter currently runs three data centers (IDs 10, 11, and 13) and it seems clear that data center number 10 shows some indications for missing Tweets, even though the standard residuals (third lines) when comparing observed (first lines) and expected Tweets (second lines) stay below 2.0 for our data. For further conclusions, it might be interesting to know where these data centers are located. When we look at the Tweet volume over the course of the UTC day for data center 10: ‿‿‿‿‿‿, data center 11: ‿‿‿‿‿‿, and data center 13: ‿‿‿‿‿‿, we can assume that individual data centers serve geographically different parts of the world, with data center number 11 most likely serving the United States of America and countries in similar time zones. Interestingly, this would confirm Morstatter et al.'s (2013) observation that Tweets from North America could be over-represented and Tweets from Asia underrepresented in Twitter's Streaming APIs.

Load. A reason for data delay that leads to missing Tweets in our day_0 dataset could be an overburdened worker/server. If this is the case, then Tweets with higher sequence numbers should have a higher chance to be delayed. Or, the other way round, Tweets that are delayed should have on average higher sequence numbers. However, a t-test ($p = 0.342$) clearly leads to the rejection of the assumption that the differences in mean are in any way significant.

Time. We also took a closer look at the timestamps of the day_0 missing Tweets and made a very astonishing observation. While the mean of the millisecond information of all Tweets is, at 502.1, very close to the expected value, the mean millisecond value of the missing Tweets is 40.9—in other words, the majority of Tweets that are missing arrived during the early part of the observed seconds. Figure 8 shows the extent of this effect. The black area marks the proportion of Tweets per millisecond that are part of day_0—for the first couple of milliseconds, 80% of Tweets are missing.

Looking at Fig. 8 might reveal some interesting insights into Twitter's system architecture. If we want to use the
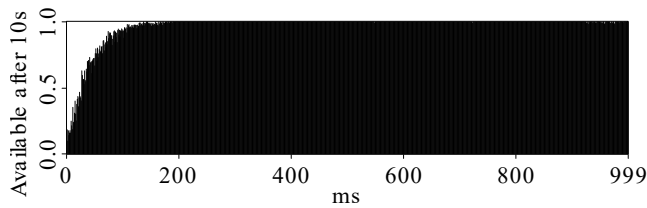
Figure 8: Proportion of Tweets per millisecond of timestamp available after 10 seconds.

Academic API to collect Tweets too close to the current time, we will get an error message from the API that "the time window of data collection must end a minimum of 10 seconds prior to the request time." Consequently, there is a good chance that this API limitation comes from a limitation of the Tweet archival architecture and that the Tweets take some seconds to arrive completely and get indexed at the server/database which handles the Academic API requests—the phenomenon seen in Fig. 8 could be the result of a stack process in which Tweets get pushed to an API server every second following a last-in first-out procedure. If our observation actually results from such a technical side-effect, we would recommend Twitter to increase the time blocking window in the API to 15 or 20 seconds.

Discussion

In the final section, we will discuss some implications for applying the Academic API for research studies. Based on our observations and insight, we also offer some recommendations for working with the Academic API in scientific research projects and for subsequent publishing.

New data—new research questions. For many researchers, having the restriction to collect only 10 million Tweets per month (the limit at the time of publication of this article) will be perceived as a clear limitation for their data collection efforts. In comparison, we can collect at least the ten-fold number of Tweets with the 1% Sample API. In contrast, the Academic API allows the study of very specific research questions, because we can submit complex queries for a specific time period back to the first Tweets in 2006. This will open Twitter research to new audiences with elevated needs for data quality, especially in the social sciences. At the same time, we will need a growing awareness about what we are actually analyzing when we collect historic Tweets. Twitter's policy of removing all public data from accounts that have been closed by the user or banned by the platform will impact certain research questions. For instance, all these Tweets from Donald Trump that attracted so much global attention over the last years are not on Twitter anymore because the account was suspended from Twitter. And with the removal of the Tweets, all Retweets of these Tweets have been removed as well.

Another aspect that needs to be considered are Twitter's efforts to reducing toxicity on the platform, e.g. Twitter's collaboration with jigsaw to deploy a *Harassment Manager*—a "technology to help women journalists document

and manage online abuse[4]". If you want to use Twitter data in the future to study online abuse, you might end up studying the effectiveness of Twitter's efforts to identify and remove this content instead.

Data decay. We have shown in this article that, while the Academic API delivers very good samples, there is a significant drop of available Tweets over time. Unfortunately, this introduces a temporal bias into our data collection, i.e. time periods that are further in the past are under-represented in the dataset. Collecting parts of the data over time could be a possible approach to reduce temporal biasing. In any case, it is clear that this side effect of non-live data collection is hard to overcome and researchers need to be aware of the possible impact on their analyses.

Twitter meta data. When historic Tweets are collected, it is important to realize that most meta data that come with the Tweet or the account that created the Tweet are in the state of the data collection time and not the Tweet creation time. As Zubiaga (2018) has stated, "The representivity of the metadata, however, keeps fading over time, both because the dataset shrinks and because certain metadata, such as the users' number of followers, keeps changing". In other words, the creation timestamp, the Tweet text, the Tweet ID, and the account ID of the sender stay constant independently from the time of data collection. Other data like the profile information of an account, the information about followers and followees, and even the account name (which can be changed on Twitter) will be from the moment of data collection. This renders questions related to follower networks impossible after a certain time period. On the plus side of getting a sample of the current state of Twitter is that we will find how often every Tweet has been retweeted, replied to, or liked in the meta-data of the Tweet without collecting the Retweets.

Open research. It has become a common practice for Twitter research to share the IDs of the Tweets that were used in the analysis of a publication. This practice is also supported by Twitter's terms of use and allows reproducibility to a certain extent. When using the Academic API, researchers should also document and publish (either in the article or in supplementary material) the following information: Date(s) of data collection, the exact search query, and the utilized API and data endpoint.

Content compliance. Twitter's developer agreement includes a section called "Content compliance" that says: "If you store Twitter Content offline, you must keep it up to date with the current state of that content on Twitter. Specifically, you must delete or modify any content you have if it is deleted or modified on Twitter." This is very explicit and if you are planning to keep to that agreement, collecting Tweets with the Academic API and then immediately analyzing them is by far the easiest way of doing so.

Data minimization. With the API v2 data endpoint, most data fields need to be requested explicitly via call parameters. Intuitively, academic researchers tend to request all possible data fields to maximize the collected dataset. How-

---

[4]https://medium.com/jigsaw/technology-to-help-women-journalists-document-and-manage-online-abuse-5edcac127872

ever, it might be worthwhile reconsidering one's data collection strategy with the API v2 endpoint. For instance, when conforming to the EU General Data Protection Regulation (GDPR), personal data shall be "limited to what is necessary in relation to the purposes for which they are processed ('data minimization')" [GDPR, Article 5 1.(c)]. Even though Twitter data is often not interpreted as being subject to ethics committees, the Academic API gives us the opportunity to explicitly *not* collect user profile information—especially when this data is irrelevant for our analysis. Furthermore, collecting less data will make the collection process faster.

Towards a more collaborative future? Twitter has become an important part of political discussions and global dynamics, and scholars in academia use Twitter data for their research. And other scholars cite this research and build on it. By providing data to an enormous number of researchers, Twitter also has to take on responsibility for the quality of the data. This responsibility can be shared with some level of transparency. We are aware that hiding details about algorithms as well as about the software and hardware architecture is a security measure of platform providers. However, we assume that many flaws and curiosities that researchers have found over the last years—e.g. possible regional biases in Twitter's streaming data (Morstatter et al. 2013)—could be easily interpreted if we had some level of general information about the system architecture. For instance, at this point, we are confident that Twitter's API architecture is separated from their *regular* system architecture. Consequently, the idea that Tweets from certain areas in the world may be delayed for real-time data collection would be a logical conclusion. Furthermore, some technical artifacts discussed in this or previous articles (Pfeffer, Mayer, and Morstatter 2018) show an indication of not being intended by the system architects. Being more transparent about their systems could lead to researchers identifying bugs and unintended features earlier—which would in return help the platform providers to improve their services.

Our future work will continue helping researchers in assessing the quality of these data sources and will provide further hints for Twitter about how to improve their services. With the release of the Academic API, the improved data endpoint v2, and a more active engagement with academic scholars, Twitter is on the right track to becoming a more trustworthy partner and a more reliable data source for academia.

Ethics Statement

Our research was not subject to the Ethics Commission at the Technical University of Munich, where the experiments were conducted. Only public Twitter data were used and the publicly available APIs were used. We were not interested in Personal Identifiable Information (PII) in this study and none of the analyses was performed on the individual account level. For those experiments in which we have sent Tweets, we only created Tweets with a single random 42-character string to minimize the risk that our Tweets will manipulate the data collection of other researchers.

With this study, we have violated Twitter's *Terms of Service*, in particular, the following section: "You may not do

any of the following [...] (ii) probe, scan, or test the vulnerability of any system or network ..." However, we strongly believe that our work, as well as similar work, is important to better understand the data that are used by thousands of researchers and practitioners. More transparency about the data will lead to higher quality and more trustworthy research, which in turn, will also be an added value for Twitter.

Acknowledgements

Author Contribution

References

Alkulaib, L.; Alhamadani, A.; Ji, T.; and Lu, C.-T. 2020. Collect Ethically: Reduce Bias in Twitter Datasets. In Lossio-Ventura, J. A.; Condori-Fernandez, N.; and Valverde-Rebaza, J. C., eds., *Information Management and Big Data*, 106–114. Springer International Publishing. ISBN 978-3-030-46140-9.

Aryal, R. R.; and Bhattarai, A. 2021. Sentiment Analysis on Covid-19 Vaccination Tweets using Naïve Bayes and LSTM. *Advances in Engineering and Technology: An International Journal*, 1(1): 57–70. Section: Articles.

Biswas, M.; Alam, T.; Househ, M.; and Shah, Z. 2022. Public Sentiment Towards Vaccination After COVID-19 Outbreak in the Arab World. In *Studies in health technology and informatics*, volume 289. ISBN 978-1-64368-250-1. Journal Abbreviation: Studies in health technology and informatics.

Bustos, V.; Comer, C.; Manstein, S.; Laikhter, E.; Shiah, E.; Xun, H.; Lee, B.; and Lin, S. 2022. Twitter Voices: Twitter Users' Sentiments and Emotions About COVID-19 Vaccination within the United States. *European Journal of Environment and Public Health*, 6: em0096.

Chen, K.; Duan, Z.; and Yang, S. 2021. Twitter as research data: Tools, costs, skill sets, and lessons learned. *Politics and the Life Sciences*, 1–17. Edition: 2021/08/09 Publisher: Cambridge University Press.

Chen, R.; Muralidharan, K.; and Samelson-Jones, B. J. 2022. Digital haemophilia: Insights into the use of social media for haemophilia care, research and advocacy. *Haemophilia*, 28(2): 247–253. Publisher: John Wiley & Sons, Ltd.

Emre Kiciman; Nicole B. Ellison; Bernie Hogan; Paul Resnick; and Ian Soboroff. 2013. Proceedings of the Sev-

enth International AAAI Conference on Weblogs and Social Media.

Eriksson Krutrök, M.; and Lindgren, S. 2022. Social media amplification loops and false alarms: Towards a Sociotechnical understanding of misinformation during emergencies. *The Communication Review*, 1–15. Publisher: Routledge.

Jungherr, A. 2015. Twitter use in election campaigns: A systematic literature review. *Journal of Information Technology & Politics*, 13(1): 72–91.

Karami, A.; Lundy, M.; Webb, F.; and Dwivedi, Y. K. 2020. Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, 8: 67698–67717.

Kergl, D.; Roedler, R.; and Seeber, S. 2014. On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 357–364.

Kim, Y.; Nordgren, R.; and Emery, S. 2020. The Story of Goldilocks and Three Twitter's APIs: A Pilot Study on Twitter Data Sources and Disclosure. *International Journal of Environmental Research and Public Health*, 17: 864.

Lamba, H.; Malik, M. M.; and Pfeffer, J. 2015. A Tempest in a Teacup? Analyzing Firestorms on Twitter. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Lasser, J. 2022. Twitter API v2 sampled Stream Tweet IDs. http://osf.io/dqx39. Accessed: 2022-09-03.

Marcec, R.; and Likic, R. 2021. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal*, postgradmedj–2021–140685.

Metzler, H.; Rimé, B.; Pellert, M.; Niederkrotenthaler, T.; Di Natale, A.; and Garcia, D. 2021. Collective Emotions during the COVID-19 Outbreak. *PsyArXiv*.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Seventh International AAAI Conference on Weblogs and Social Media*, 400–408.

Muric, G.; Wu, Y.; and Ferrara, E. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health Surveill*, 7(11): e30642.

Paradowski, R. 2021. Women's Strike in Poland as a Communication Phenomenon on Twitter. In *Wokół strajków kobiet*, 239–268. Dziwak, Ewelina and Gheorghe, Kamila. ISBN 978-83-67074-16-2.

Pellert, M.; Lasser, J.; Metzler, H.; and Garcia, D. 2020. Dashboard of sentiment in Austrian social media during COVID-19. *Frontiers in big Data*, 32.

Pfeffer, J.; Mayer, K.; and Morstatter, F. 2018. Tampering with Twitter's Sample API. *EPJ Data Science*, 7(50).

Pradyumn, M.; Kapoor, A.; and Tabrizi, N. 2018. Big Data Analytics on Twitter. In *Big Data – BigData 2018*, 326–333. Springer International Publishing.

Ruths, D.; and Pfeffer, J. 2014. Social Media for Large Studies of Behavior. *Science*, 346(6213): 1063–1064.

Salvatore, C.; Biffignandi, S.; and Bianchi, A. 2021. Social Media and Twitter Data Quality for New Social Indicators. *Social Indicators Research*, 156(2): 601–630.

Sinnenberg, L.; Buttenheim, A. M.; Padrez, K.; Mancheno, C.; Ungar, L.; and Merchant, R. M. 2017. Twitter as a Tool for Health Research: A Systematic Review. *American Journal of Public Health*, 107(1): e1–e8.

Thelwall, M.; and Kousha, K. 2021. Researchers' attitudes towards the h-index on Twitter 2007–2020: criticism and acceptance. *Scientometrics*, 126(6): 5361–5368.

Timoneda, J. 2018. Where in the world is my tweet: Detecting irregular removal patterns on Twitter. *PLOS ONE*, 13(9).

Wallaschek, S.; Kaushik, K.; Verbalyte, M.; Sojka, A.; Sorci, G.; Trenz, H.-J.; and Eigmüller, M. 2022. Same Same but Different? Gender Politics and (Trans-)National Value Contestation in Europe on Twitter. *Politics and Governance*, 10: 146–160.

Wang, Y.; Callan, J.; and Zheng, B. 2015. Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API. *ACM Trans. Web*, 9(13).

Weber, D.; Nasim, M.; Mitchell, L.; and Falzon, L. 2021. Exploring the effect of streamed social media data variations on social network analysis. *Social Network Analysis and Mining*, 11(1): 62.

Williams, S. A.; Terras, M. M.; and Warwick, C. 2013. What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, 69(3): 384–410.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2020. Variation across Scales: Measurement Fidelity under Twitter Data Sampling. *Proceedings of the International AAAI Conference on Web and Social Media*, 14: 715–725.

Yu, J.; and Muñoz-Justicia, J. 2020. A Bibliometric Overview of Twitter-Related Studies Indexed in Web of Science. *Future Internet*, 12(5): 91.

Zimmer, M.; and Proferes, N. J. 2014. A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3): 250–261.

Zubiaga, A. 2018. A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, 69(8): 974–984. Publisher: John Wiley & Sons, Ltd.