Facing the volatility of tweets in altmetric research

Fang, Z.; Dudek, J.; Costas, R.

Received: 11 July 2021 | Revised: 5 December 2021 | Accepted: 23 January 2022

DOI: 10.1002/asi.24624

BRIEF COMMUNICATION

JASIST WILEY

Facing the volatility of tweets in altmetric research

Zhichao Fang[1] | Jonathan Dudek[1] | Rodrigo Costas[1,2]

[1]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands

[2]DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy, Stellenbosch University, Stellenbosch, South Africa

Correspondence
Zhichao Fang, Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands.
Email: z.fang@cwts.leidenuniv.nl

Funding information
China Scholarship Council, Grant/Award Number: 201706060201; The South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP)

Abstract

The data re-collection for tweets from data snapshots is a common methodological step in Twitter-based research. Understanding better the volatility of tweets over time is important for validating the reliability of metrics based on Twitter data. We tracked a set of 37,918 original scholarly tweets mentioning COVID-19-related research daily for 56 days and captured the reasons for the changes in their availability over time. Results show that the proportion of unavailable tweets increased from 1.6 to 2.6% in the time window observed. Of the 1,323 tweets that became unavailable at some point in the period observed, 30.5% became available again afterwards. "Revived" tweets resulted mainly from the unprotecting, reactivating, or unsuspending of users' accounts. Our findings highlight the importance of noting this dynamic nature of Twitter data in altmetric research and testify to the challenges that this poses for the retrieval, processing, and interpretation of Twitter data about scientific papers.

1 | INTRODUCTION

Since its launch in 2006, Twitter has provided data for researchers to understand its use in public conversations, which is spurred further by the more recent developments around the Twitter Application Programming Interface (API). The data policy launched by Twitter as of 2020 enables even more fine-grained and larger-scale data collection of tweets, especially for academic research purposes (Cairns & Shetty, 2020).

The abundance of data available from Twitter results in numerous data snapshots of tweets created and analyzed in different contexts, such as those created by Altmetric.com or Crossref Event Data which record tweets mentioning scientific papers (specifically referred to as *scholarly tweets* in this study), or other more general Twitter data snapshots such as those containing tweets on COVID-19 (Chen et al., 2020). Because of the Twitter restriction on content redistribution,[1] what researchers can obtain directly from Twitter data snapshots is often limited to tweet IDs or Twitter user IDs, making it necessary to perform data re-collection via the Twitter API to get more detailed and up-to-date information about the tweets.

An important feature of Twitter data snapshots is that they capture Twitter activities at a given point in time. However, the tweets recorded in a snapshot may have changed between the time of the creation of the snapshot and the time of the update of the tweets. Tweets may disappear because of tweet deletion, or the protection or suspension of user accounts; and user accounts can also be restored, thus making their tweets available again in the platform. These are aspects that may go relatively unknown to the scientometric research community, since in contrast to scientometric data (e.g., citations, publications) that are

stable and perdurable,[2] Twitter data is of a far more volatile nature.

The variable availability of tweets is uncertain as long as they are sealed in snapshots, until the up-to-date state of the tweets is rechecked. There have been some previous studies aiming at rechecking the already retrieved tweets to inspect how many of them had become unavailable after a while. For instance, Xu et al. (2013) rechecked a set of 0.3 million bullying-related tweets, finding that after 2 weeks around 4% of all the tweets became unavailable. In 2016, Zubiaga (2018) rechecked the completeness of 30 Twitter datasets associated with different real-world events posted between 2012 and 2016, finding that, on the whole, 18.6% of the tweets were unavailable during the data rechecking. Bastos (2021) reported that 33% of the nearly 3 million tweets about the Brexit referendum posted in 2016 were no longer available 3 years after the vote. In addition, by rechecking the 42.5 million scholarly tweets recorded by Altmetric.com until October 2017, Fang et al. (2020) found that 13% of the tweets had become unavailable by September 2019.

Previous literature mainly analyzed the transition of tweets from availability to unavailability after a certain period. In this study we contribute to this discussion by longitudinally tracking the changing availability of tweets over time. We adopt a more dynamic perspective and aim to unravel the processes behind and the reasons for tweets becoming unavailable and available at different points in time.

2 | DATA

Our dataset stemmed from a set of scholarly tweets recorded by Altmetric.com, which referred to COVID-19-related scientific papers recorded in the CORD-19 dataset and the file of COVID-19-related research provided by Dimensions.[3] The pandemic has generated a large sample of tweets, being relevant for studying the social attention towards COVID-19-related topics in online environments. Due to the many controversial topics surrounding COVID-19 on Twitter, there may exist more frequent changes in the availability state of related tweets for observation purpose. Therefore, we selected all of the 37,918 original scholarly tweets posted between January 1, 2021 and January 20, 2021 as our dataset. Their availability state was rechecked on a daily basis from March 6, 2021 to April 30, 2021 (a 56-day observation window), using the Twitter API. For all unavailable tweets identified, the error codes returned by the Twitter API were collected to study the specific reasons for the unavailability.

3 | RESULTS

3.1 | Longitudinal unavailability rate of tweets

A total of 1,323 out of the 37,918 original tweets (3.5%) became unavailable at least once. Table 1 lists the four main unavailability reasons and the number of unavailable tweets.[4] Tweet deletion (error code 144) is the most common reason of unavailability, followed by account protection (error code 179), account suspension (error code 63), and account deactivation (error code 34).

Figure 1 shows the daily unavailability rate (i.e., proportion of tweets that are unavailable) of tweets during the observation period. The number of unavailable tweets presents an uptrend over time, with the unavailability rate increasing from 1.6% at the beginning to 2.6% at the end. However, the uptrend is not continuous, exhibiting some occasional ups and downs. For example, on March 6 there were 596 unavailable tweets, while in the next day the number decreased to 593, implying that the number of "revived" tweets was by three higher than the number of tweets becoming unavailable in that day.

3.2 | The volatility of the availability state of tweets

We classified the 1,323 observed unavailable tweets into two types:

- Type one: tweets that became unavailable at some point and remained unavailable till the end of the period (919 tweets, 69.5%).
- Type two: tweets that became unavailable at some point and then turned back to being available at least once during the period (404 tweets, 30.5%).

Figure 2 shows the error codes for the unavailable tweets by type. All unavailable tweets caused by tweet deletion remained unavailable, as it is not possible to restore them to publicly available state.[5] Most unavailable tweets from suspended accounts also did not experience further switching of state, although some of them became available again, probably due to the unsuspension of Twitter accounts.[6] About 61.1% of type two tweets came from accounts that changed their Twitter users' protection behavior. Unavailable tweets caused by account protection can become easily available again once users choose to stop protecting their accounts. Finally, about 30.4% of unavailable tweets became available again as a result of users reactivating their Twitter accounts shortly after their accounts had been deactivated.[7]

TABLE 1 Number of unavailable tweets and their unavailability reasons

Error code	Description	Error message displayed on Twitter	Number of tweets
144	"The requested tweet ID is not found (if it existed, it was probably deleted)"	"Hmm... This page doesn't exist. Try searching for something else"	419
179	"Thrown when a tweet cannot be viewed by the authenticating user, usually due to the tweet's author having protected their tweets"	"You're unable to view this tweet because this account owner limits who can view their tweets"	394
63	"The user account has been suspended and information cannot be retrieved"	"This tweet is from a suspended account"	353
34	"The specified resource was not found"	"This tweet is from an account that no longer exists"	287

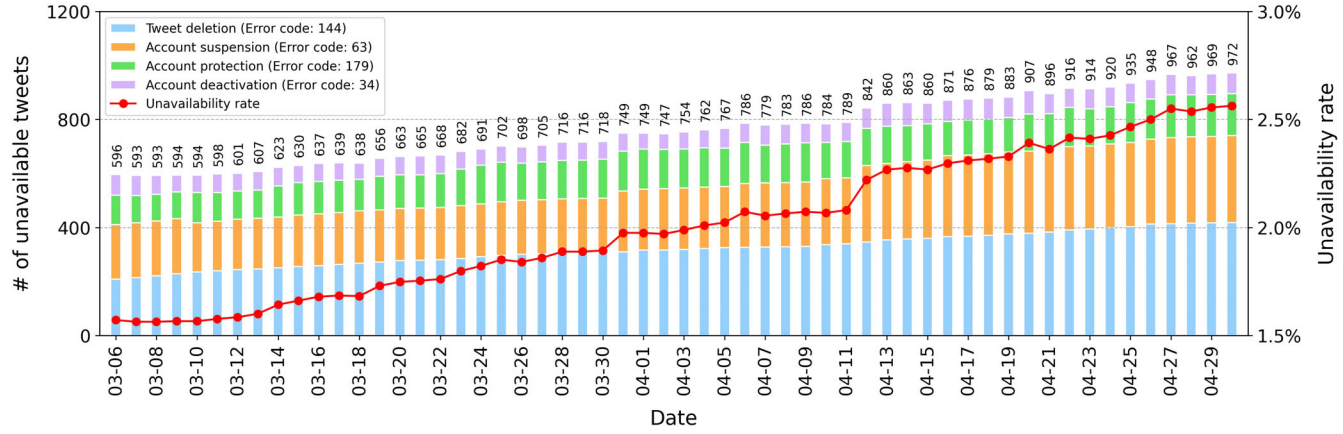

FIGURE 1 Daily number of unavailable tweets over the observation period

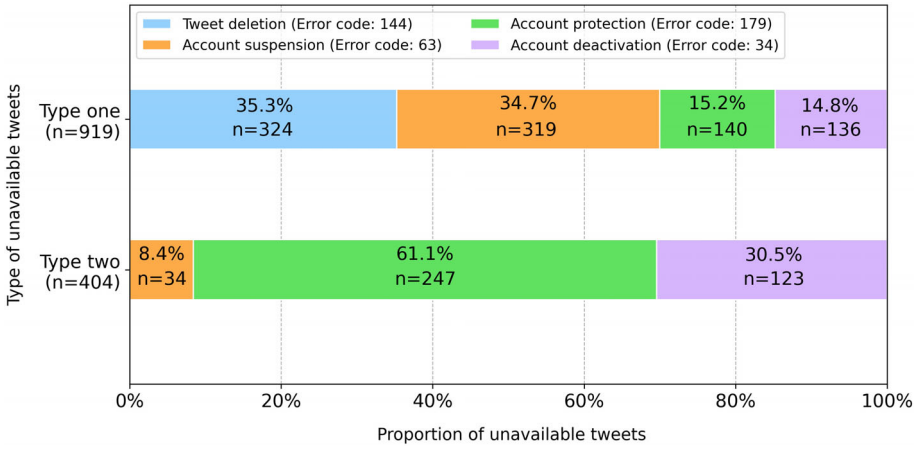

FIGURE 2 Proportion of unavailable tweets by type of unavailable tweets and cause of unavailability

4 | CONCLUSION

In this study we discuss the role of the volatility of tweets in the study of scholarly tweets. From a conceptual point of view, the volatility of tweets poses a relevant challenge for the stability of metrics based on Twitter data in the context of *altmetrics* (Haustein, 2016). About 3.5% of the tracked scholarly tweets became unavailable at some point in a period of 56 days. This is a small percentage, but it may be a nontrivial amount when working with older and larger sets of tweets, or tweets related to controversial topics or accounts whose deletion or suspension[8] would lead to a cascade of related Twitter data becoming unavailable (Bastos, 2021; Fang et al., 2020).

About 30.5% of all unavailable tweets became available again at a later stage. Changes happening at the user

account level are the reason for the *revival* of tweets. Such large share of revived tweets suggests that different data rechecking rounds over the same set of tweets may provide different results. Screening for the unavailability reasons may become a relevant methodological step to estimate the potential effect that these revived tweets may have on the analyses.

As an exploratory study, this paper is limited to a specific research area (i.e., COVID-19). The situation of the unavailability and revival of tweets related to other topics may present some distinguishing patterns, which in itself is an interesting area for future research. Our main point in this study is that as a common phenomenon intrinsic to the Twitter ecosystem, the volatility of Twitter data hints to the importance of considering this challenge in future Twitter research in general, and in altmetrics in particular, in which some of the tweets under analysis may no longer be available (or unavailable) at the time of the study.

ACKNOWLEDGMENTS

Zhichao Fang is financially supported by the China Scholarship Council (201706060201). Rodrigo Costas is partially funded by the South African DST-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). The authors thank Altmetric.com for providing the tweet IDs for research purposes, and also thank the three anonymous reviewers for their valuable comments.

ORCID

Zhichao Fang https://orcid.org/0000-0002-3802-2227
Jonathan Dudek https://orcid.org/0000-0003-2031-4616
Rodrigo Costas https://orcid.org/0000-0002-7465-6462

ENDNOTES

[1] See more information about the developer policy released by Twitter at: https://developer.twitter.com/en/developer-terms/agreement-and-policy

[2] From a fundamental point of view, publications (even retracted ones) and their citations stay forever as part of the public scientific record. Databases and indexing platforms may fail in properly identifying them, bringing some degree of volatility in scientometric data, but this is more a technical form of volatility compared to Twitter.

[3] CORD-19 (COVID-19 Open Research Dataset, version: January 18, 2021): https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html; the file of COVID-19-related research made available by Dimensions (version: December 2, 2020) can be found here: https://www.dimensions.ai/news/dimensions-is-facilitating-access-to-covid-19-research/

[4] There are 126 tweets with more than one unavailability reason during the observation period. A full-counting was applied in calculating the number of unavailable tweets for each reason.

[5] Twitter states: "once a tweet has been deleted, the tweet contents, associated metadata, and all analytical information about that tweet is no longer publicly available on Twitter" (https://help.twitter.com/en/using-twitter/delete-tweets).

[6] See detailed rules about (un)suspension of Twitter accounts at: https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts

[7] After an account is deactivated, the user has 30 days to reactivate it. Otherwise the data associated with the deactivated account will not be restorable. See more information about account deactivation at: https://help.twitter.com/en/managing-your-account/how-to-deactivate-twitter-account

[8] The Twitter account of the former U.S. president Donald Trump—suspended on January 6, 2021—had over 88 million followers at the time of its suspension, leading to the disappearance of all the tweets by the account and the related engagement events (e.g., retweets and likes). https://en.wikipedia.org/wiki/Social_media_use_by_Donald_Trump

REFERENCES

Bastos, M. (2021). This account doesn't exist: Tweet decay and the politics of deletion in the Brexit debate. *American Behavioral Scientist*, *65*(5), 757–773. https://doi.org/10.1177/0002764221989772

Cairns, I., & Shetty, P. (2020, July 16). *Introducing a new and improved Twitter API*. Retrieved from https://blog.twitter.com/developer/en_us/topics/tools/2020/introducing_new_twitter_api.html

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2), e19273. https://doi.org/10.2196/19273

Fang, Z., Dudek, J., & Costas, R. (2020). The stability of Twitter metrics: A study on unavailable Twitter mentions of scientific publications. *Journal of the Association for Information Science and Technology*, *71*(12), 1455–1469. https://doi.org/10.1002/asi.24344

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*, *108*(1), 413–423. https://doi.org/10.1007/s11192-016-1910-9

Xu, J. M., Burchfiel, B., Zhu, X., & Bellmore, A. (2013). An examination of regret in bullying tweets. In *Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 697–702). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/N13-1082/

Zubiaga, A. (2018). A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology*, *69*(8), 974–984. https://doi.org/10.1002/asi.24026

How to cite this article: Fang, Z., Dudek, J., & Costas, R. (2022). Facing the volatility of tweets in altmetric research. *Journal of the Association for Information Science and Technology*, *73*(8), 1192–1195. https://doi.org/10.1002/asi.24624