Understanding Discourse-Topic Dependent Data Loss in Social Media Archives

Muhammad Taimoor Khan, Johannes Kiesel, Dimitar Dimitrov and Stefan Dietze

GESIS - Leibniz Institute for the Social Sciences

Social media platforms offer direct access to streams of information that reflect and react to real-world events. These streams are becoming a prevalent data source in computational social science to study behavioral patterns and analyze discourse Khan et al. (2025). However, a part of this important data is lost over time due to many factors, such as users deleting their posts or user accounts, which makes all their posts unavailable in the public space. The users can also change the privacy settings of their individual posts, making them change their availability. The platform may also deactivate or suspend the users' accounts for violating its terms and policies. In computational reproducibility literature, data loss is acknowledged as a crucial factor that degrades the data quality in data archives, while also limiting the reproducibility of research findings.

In this research, we investigate the association of data loss patterns in social media archives with time and topical discourse. For this purpose, social media posts are sampled for topics of interest from the archive and compared for their data loss patterns. We present two hypotheses, i.e., i) most of the data loss occurs shortly after the posts are being shared, and ii) the rate of data loss varies across discourses depending on their topic and linguistic characteristics. For the former, we analyze the percentage of unavailable posts from the month in which they were posted. For the latter, the posts are separated into topics using topic-specific keywords, while the unavailable posts are compared across them.

To empirically evaluate our hypothesis, we resort to the Twitter (now X) 1% data archive called TweetsKB[1], having billions of tweets collected over a decade. We extracted 9.4 million tweets on COVID-19 from the time window October 2019 to May 2020 and call them COVID-19-specific discourse. To extract the COVID-19-related posts, we used the COVID-19 keywords. For comparison, we sampled another 9.4M random tweets from the same time frame and called it the general discourse. Each of the two discourses is further separated into three segments based on their sentiment polarity, i.e., positive COVID-19 tweets, neutral COVID-19 tweets, and negative COVID-19 tweets. Similarly, the positive general tweets, neutral general tweets, and negative general tweets. Finally, we also considered polarization-based subtopics for data loss, i.e., non-polarized COVID-19 tweets, polarized COVID-19 tweets, non-polarized general tweets, and polarized general tweets. This makes two higher-level topical comparisons between COVID-19 and general tweets, while also having their subtopic comparisons along sentiment polarity and polarization dimensions.

Our contribution is the formulation and estimation of data loss with time and discourse-topic-dependent data loss. Our first hypothesis is found to be true as the majority of the tweets are deleted within their first month of posting. We also found that topical discourse is a factor in data loss, where COVID-19 discourse, despite being a divisive topic, has lower deletion. It may be associated with the nature of discussions, i.e., mostly around life affecting matters, e.g., religion, job, money, etc. While the general discourse has many more unavailable tweets, which is also aligned with existing literature, associating tweet deletion with loose and informal language, or disclosing personal information, leading to embarrassment. We also found that neutral tweets have a higher probability of deletions compared to negative and positive tweets. This may be due to not creating enough engagement on the platform. This pattern is observed in the polarity-based subtopics of both COVID-19 and general discourses. Surprisingly, we also observed that in both COVID-19 and general discourse, the tweets with higher polarization have a lower probability of being deleted, where the deletion probability drops further for

MethodsNET'25: The 2nd MethodsNET conference, 10-12 September 2025, Louvain-la-Neuve, Belgium
taimoor.khan@gesis.org (M. T. Khan)

[1] https://data.gesis.org/tweetskb/

highly polarized tweets.

Our study has some limitations that require alternative approaches for future in-depth investigation. The platform restricts information about unavailable tweets, as they cannot be rehydrated without any further information on why the tweet is not available. Therefore, it is unknown whether the tweet is deleted by the user, removed by the platform, or unavailable due to other reasons. The polarization categorization of the tweets is based on external resources, i.e., MediaBiasFactCheck[2], based on the URLs shared in the tweet only, without considering the tweet content. The tweets for this study are sampled either on COVID-19 keywords for the COVID-19 discourse or at random for the general discourse, and therefore, the users in the study may have other posts that are not considered in the study.

In this study, we found that data loss in a discourse has a correlation with the discourse topic, where the COVID-19 discourse has fewer deletions than the general discourse. Most of the data is lost shortly after being shared, where the deletions dry out in the following months. The data loss is also associated with the polarity-based subtopics, where the neutral tweets have a higher deletion probability. It may be due to lesser user engagement. Tweet deletion and polarization are negatively correlated, where the polarized subtopics within the general and COVID-19 discourse have fewer deletions that drop further with higher polarization. The study contributes to improving the understanding of data loss in social media archives in relation to the topical discourse that needs further exploration.

References

Khan, M. T., Dimitrov, D., and Dietze, S. (2025). Characterization of tweet deletion patterns in the context of covid-19 discourse and polarization. In *Proceedings of the 36th ACM Conference on Hypertext and Social Media*, pages 43–47.

[2]https://mediabiasfactcheck.com/