2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

# Quantifying the Transience of Social Web Datasets

Mohammed Afaan Ansari*
*Sardar Patel Institute of Technology*
Mumbai, India
afaan.ansari@spit.ac.in

Jiten Sidhpura*
*Sardar Patel Institute of Technology*
Mumbai, India
jiten.sidhpura@spit.ac.in

Vivek Kumar Mandal*
*Sardar Patel Institute of Technology*
Mumbai, India
vivekkumar.mandal@spit.ac.in

Ashiqur R. KhudaBukhsh
*Rochester Institute of Technology*
Rochester, USA
axkvse@rit.edu

***Abstract*—The social web presents a modern-day instrument to analyze a wide range of behavioral research questions. Of these platforms, Twitter has played a key role in social science research for more than a decade. This paper looks into an underexplored aspect – transience of Twitter datasets and makes the following three contributions. First, via a comprehensive investigation of more than 40 Twitter datasets, we identify that many of these datasets suffer from severe retrieval loss. Second, we demonstrate that the retrieval loss across labels is often imbalanced with inappropriate labels (e.g., misinformation, hate speech) suffering from more retrieval loss. Finally, we demonstrate that imbalanced retrieval loss may impact machine learning models differently than balanced retrieval loss.**

***Index Terms*—Quantifying Dataset Transience, Quantifying Retrieval Imbalance, Social Web**

## I. Introduction

The social web presents a powerful instrument for analyzing a wide range of behavioral research questions involving political polarization [1, 2], modern conflicts [3], misinformation [4, 5] and hate speech [6, 7], and response to key policy issues [1]. The social science research community greatly benefits from publicly available datasets that not only answer such key behavioral research questions but also often serve as benchmarks to track methodological advancements.

Although social web data holds an important place in understanding a wide range of social science research questions, the shelf-life of such datasets could be uncertain for several reasons. Platform moderation [8], platform deprecation [9], users voluntarily deleting (or changing privacy settings of) content, state censorship [10], or the more recent platform restrictions to prevent web scraping for foundation models [11] – many are the reasons a social web post may not continue seeing the light of day.

This paper looks at a critical aspect of social web datasets largely unexplored heretofore – the transience of social web datasets. Our paper makes a strong case in starting the dialogue that the very assumption of permanence of train and test data in classical supervised machine learning may need to be revisited due to the transient nature of the modern social web. Specifically, we consider Twitter, one of the most prominent (and erstwhile most accessible to academic researchers) platforms. Via a comprehensive sample of 40 well-known Twitter datasets, we investigate the following research questions.

- `RQ1`: To what extent do social web datasets experience retrieval loss once they are published and shared?
- `RQ2`: For multi-class datasets, is the retrieval loss balanced across classes?
- `RQ3`: Does the performance of supervised solutions get affected if the loss is imbalanced as opposed to a balanced retrieval loss?

Our main contributions are the following:

- Via a comprehensive analysis of 40 Twitter datasets, we quantify retrieval loss in well-known web data sets.
- We demonstrate that retrieval loss is imbalanced and often affects the minority classes more than the majority classes.
- We demonstrate that supervised solutions' performance varies with different retrieval loss assumptions.

## II. Design Considerations

### A. Twitter as the Platform Choice and Why our Work is Still Relevant Regardless of Twitter's Data Access Policy Changes

While mainstream social web platforms such as Reddit, YouTube, and fringe social web platforms such as Gab [12] or Stormfront [13] have contributed to important social web datasets, Twitter has been the de facto primary source for social web datasets for nearly one decade for following reasons. First, the global daily participation of more than

* These authors contributed equally to this work.



*ASONAM '23*, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11…$15.00
https://doi.org/10.1145/3625007.3627596

237 million users makes this platform a virtual public square with vast topical, linguistic, political, and cultural diversity. Second, until recently, Twitter provided free API access to both developers and academics catering to the different needs of academic researchers and developers. Before the free academic Twitter API got deprecated in Feb 2023 [14], the API had free access, generous rate limits, and allowed historical search up to any time horizon. This openness and ease of data access fostered social media research using Twitter data. Finally, Twitter also presented an effective "rehydration mechanism" where only the tweet ids sufficed for subsequent reconstruction. Any subsequent reconstruction of the dataset can be performed by "rehydrating" the tweets. This mechanism protects users' rights to be forgotten as a deleted tweet id cannot be recovered through rehydration.

Twitter, as it was a year ago, was a natural choice as a platform for our investigation. With the current deprecation of free academic Twitter APIs [14] and more draconian policies on user rate daily limits [15], a pertinent question is why is our study still relevant.

From one of the earliest works on election winner prediction through social media data in 2010 [16] to releasing multimodal Tweets on election fraud claims in 2020 [17] – Twitter has been a stable platform for social science research over a long period of time which makes it suitable for studying dataset shelf-life. Second, we argue that shifting access to data will remain a key consideration future ML applications will continually grapple with. In the wake of generative AI and Large Language Models research, big social media platforms beyond Twitter are already restricting free data access to prevent extensive data scraping [11]. User privacy concerns will also be another important reason for platforms tightening their grip on how much information researchers have access to. Hence Twitter's shifting data access policies notwithstanding, our work quantifies and chronicles an important aspect of ML resources going forward.

### B. Twitter Datasets to Quantify Shelf-life

We identify 40 Twitter datasets published at well-known venues that include prominent conferences such as AAAI, ICWSM, EMNLP, ACL, IJNLP, IJCNN, IEEE Big Data, and well-regarded journals such as Machine Learning Journal (MLJ), Transactions of the Association for Computational Linguistics (TACL), and Information Processing and Management Journal. These datasets span a broad range of tasks such as stance mining, sentiment classification, spam or misinformation detection, hate speech detection, and rumor detection. Table I lists our datasets. We collect these datasets during the months of November and December in 2022.

We first note that our selected datasets not only present a wide range of tasks, but they also greatly vary in their size (397,768 $\pm$ 1,353,814). Many of these datasets are labeled with multiple classes for supervised learning with the number of classes ranging from 2 to 1,068.

### C. Computing Retrieval Loss

For each dataset, we consider the reference dataset size as reported in the paper. We tally the reported dataset size with the retrieved dataset size and compute the retrieval loss. For datasets with individual class labels, we compute the retrieval loss for each class as well. We illustrate our retrieval loss calculation with a simple example next.

Consider a dataset $\mathcal{D}$ with class labels $\mathcal{A}$ and $\mathcal{B}$. As reported in the paper, consider $\mathcal{D}$ has overall 100 instances, 80 instances of $\mathcal{A}$, and 20 instances of $\mathcal{B}$. Upon retrieval, we obtain 75 instances of $\mathcal{A}$ and 5 instances of $\mathcal{B}$. The overall retrieval loss is 100 - (75 + 5) = 20%. For class $\mathcal{A}$, the retrieval loss is $\frac{80-75}{80} = 6.25\%$. For class $\mathcal{B}$, the retrieval loss is $\frac{10-5}{10} = 75\%$.

## III. Related Work

While ML literature has considered dynamic settings such as concept drift [18, 19] and paid considerable attention to dealing with missing data [20, 21], and assessing temporal persistence of prediction systems [22, 23], little or no work exists that challenges the assumption of data permanence in a real-world setting and seeks to quantify shelf-life of web data through a comprehensive study like ours.

The unavailability of data as a possible challenge has been discussed in the medical literature [24] and social web transience has been cited as a limitation in studies dating back to 2010 [16] to a study as recent as in 2023 [25]. However, to our knowledge, no comprehensive analysis has quantified and characterized the transience of social web datasets at our scale.

At a philosophical level, our work is a part of the broader conversation on machine learning audits where platforms [26], datasets [25], and models [27] are audited for potential limitations and blind spots.

## IV. Results and Analyses

### A. Overall Retrieval Loss

**RQ 1:** *to what extent do social web datasets experience retrieval loss once they are published and shared?*

**Observation 1:** *considerable retrieval loss is a common phenomenon observed in well-known Twitter datasets.*

Table I reports the retrieval loss observed in each of these datasets. As already mentioned, we report the loss with respect to the reported dataset size in the published papers. We notice that barring a handful of datasets, almost all datasets experience considerable retrieval loss. In four datasets 80% or more tweets can no longer be retrieved and fourteen datasets have 40% or more retrieval loss. Intuitively, retrieval loss is correlated with the publication age of the dataset (correlation coefficient 0.83) – datasets published earlier in general exhibited greater retrieval loss. Our finding has serious implications for ML research that rely on social web datasets as traditional machine learning typically assumes data permanence. Several retrieval losses would make performance comparison and reproducibility extremely challenging.

We observe that the nature of the tweet sources can influence retrieval loss. For instance, both $\mathcal{D}_{BlackLivesMatter}$ [28] and

TABLE I: Twitter Dataset Information. Retrieval loss is computed with respect to the dataset size reported in the published paper. For a given dataset, $RetrievalImbalance_{KL}$ measures imbalance in retrieval loss across different classes. A higher value indicates greater imbalance.

| **Dataset** | **Published size** | **Retrieval loss** | **# of classes** | $RetrievalImbalance_{KL}$ |
|---|---|---|---|---|
| $\mathcal{D}_{BlackLivesMatter}$ [28] | 6,645 | 1.78 | NA | NA |
| $\mathcal{D}_{WhitenessViolence}$ [29] | 64,808 | 42.61 | NA | NA |
| $\mathcal{D}_{HateSpeechTwitter}$ [30] | 79,996 | 42.19 | 4 | 0.0192 |
| $\mathcal{D}_{RumorDetection}$ [31] | 1,793 | 13.55 | - | NA |
| $\mathcal{D}_{HateSpeechAbusiveBehavior}$ [32] | 89,899 | 49.64 | - | NA |
| $\mathcal{D}_{EmpathyAndHope}$ [33] | 2,40,176 | 26.70 | - | NA |
| $\mathcal{D}_{ForecastingWinnersAndLosers}$ [34] | 18,173 | 25.11 | - | NA |
| $\mathcal{D}_{ClimateCOVIDMilitaryMisinformation}$ [35] | 8,75,575 | 3.77 | 3 | 0.077 |
| $\mathcal{D}_{StanceDetectionInFinancialDomain}$ [36] | 49,711 | 41.35 | 4 | 0.0007 |
| $\mathcal{D}_{COVIDCountryImage}$ [37] | 7,939 | 39.11 | 4 | 0.006 |
| $\mathcal{D}_{COVIDCQStance}$ [38] | 14,200 | 37.92 | 3 | 0.059 |
| $\mathcal{D}_{PropagandaTechTwitter}$ [39] | 9,848 | 100.00 | 20 | 0 |
| $\mathcal{D}_{Flood}$ [40] | 2,500 | 11.48 | 4 | 0.4141 |
| $\mathcal{D}_{SentimentClassification}$ [41] | 1,576,090 | 51.22 | 2 | 0.0004 |
| $\mathcal{D}_{VoterFraud2020}$ [42] | 7,367,124 | 50.02 | 5 | 0.0456 |
| $\mathcal{D}_{GeolocationPrediction}$ [43] | 9,900 | 52.79 | 966 | 0.0640 |
| $\mathcal{D}_{Traffic\text{-}relatedTweets}$ [44] | 50,600 | 10.17 | 3 | 0.1425 |
| $\mathcal{D}_{SMILEEmotion}$ [45] | 3,000 | 17.30 | 13 | 0.1116 |
| $\mathcal{D}_{PersonalHealthMention}$[46] | 7,100 | 38.28 | 4 | 0.0113 |
| $\mathcal{D}_{SpamDetection}$ [47] | 14,600 | 19.66 | 2 | 0.1689 |
| $\mathcal{D}_{TrustAndBelieve}$ [48] | 900 | 95.22 | 2 | 0.0002 |
| $\mathcal{D}_{GenderClassification}$ [49] | 9,514 | 40.54 | 2 | 0.0001 |
| $\mathcal{D}_{EmojiExtraction}$ [50] | 4,006,994 | 43.52 | 1088 | 0.0398 |
| $\mathcal{D}_{COVID19Misinformation}$[51] | 500 | 26.00 | 2 | 0.0057 |
| $\mathcal{D}_{DocumentClusteringBenchmark}$ [52] | 2,600 | 13.5 | 15 | 0.0518 |
| $\mathcal{D}_{ContextualizeMourning}$ [53] | 2,500 | 21.64 | 2 | 0.0001 |
| $\mathcal{D}_{CMU\text{-}MisCov19}$ [54] | 4,500 | 28.44 | 16 | 0.1806 |
| $\mathcal{D}_{ChangeInEmotion}$ [55] | 28,836 | 18.75 | 5 | 0.0188 |
| $\mathcal{D}_{CivilUnrest}$ [56] | 7,826 | 35.89 | 2 | 0.0019 |
| $\mathcal{D}_{MultiModalSarcasmDetection}$ [57] | 36,100 | 24.12 | - | NA |
| $\mathcal{D}_{SocialMediaEnglish}$ [58] | 10,398 | 45.74 | 7 | 0.0578 |
| $\mathcal{D}_{OpenDatasetOfScholar}$[59] | 389,722 | 84.17 | 2 | 0.00001 |
| $\mathcal{D}_{AbusiveLanguageContext}$ [60] | 3,400 | 25.79 | 2 | 0.0151 |
| $\mathcal{D}_{Conversation}$ [61] | 9,867 | 0.0 | 7 | NA |
| $\mathcal{D}_{CyberThreatDetection}$ [62] | 8,100 | 97.20 | 2 | 0.0368 |
| $\mathcal{D}_{EarlyCOVIDFakeNews}$ [63] | 3,900 | 21.41 | 2 | 0.4175 |
| $\mathcal{D}_{MeToo}$ [64] | 9,874 | 21.88 | 5 | 0.0734 |
| $\mathcal{D}_{Covid\text{-}HeRA}$ [65] | 90,000 | 17.68 | 5 | 0.0875 |
| $\mathcal{D}_{TopicAndDiscourse}$ [66] | 420,400 | 21.28 | - | NA |
| $\mathcal{D}_{HatefulSymbolOrHatefulPeople}$ [67] | 16,903 | 39.92 | - | NA |

$\mathcal{D}_{CivilUnrest}$ [56] concern civil unrest. However, $\mathcal{D}_{BlackLivesMatter}$ consists of corporate responses to the Black Lives Matter movement where many participating companies are Fortune 100 companies. Whereas $\mathcal{D}_{CivilUnrest}$ are user-generated tweets from Australia relevant to civil unrest events in Australia. We notice that tweets from official corporate handles suffered minimal retrieval loss (1.78%) whereas tweets from regular Twitter users suffered considerable retrieval loss (35.89%).

The nature of the dataset also influenced retrieval loss. Datasets with potentially inappropriate/dangerous content experienced substantial retrieval loss. For instance, $\mathcal{D}_{PropagandaTechTwitter}$ [39] and $\mathcal{D}_{CyberThreatDetection}$ [62] exhibit 100% and 97.2% retrieval loss, respectively.

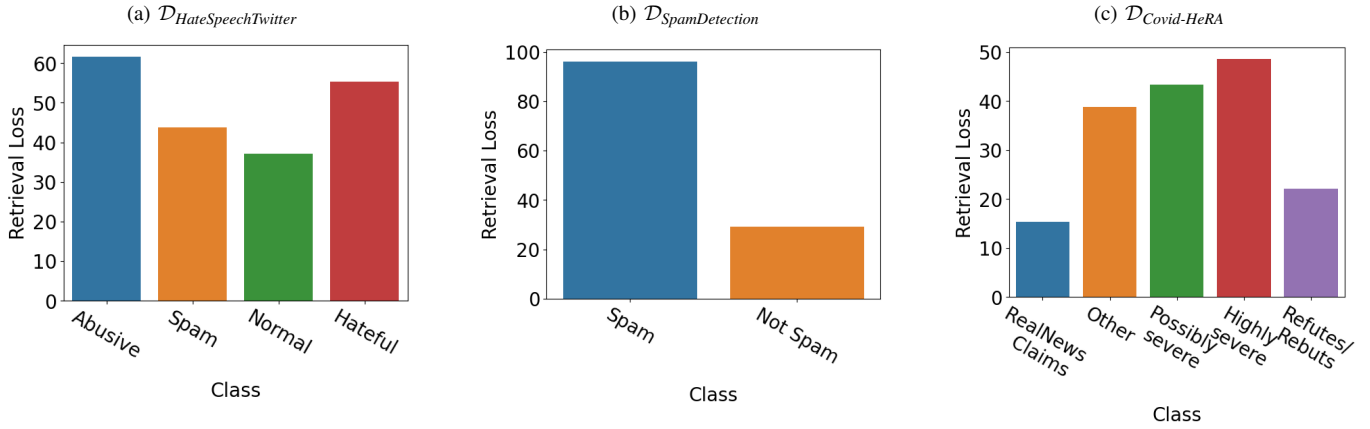
### *B. Retrieval Loss Across Classes*

**RQ 2:** *For multi-class datasets, is the retrieval loss balanced across classes?*

**Observation 2:** *Retrieval loss often varies across class labels with inappropriate classes (e.g., misinformation, hate speech, spam) experiencing larger retrieval loss.*

Figure 1 presents retrieval loss at the granularity of individual classes for three datasets: $\mathcal{D}_{HateSpeechTwitter}$; $\mathcal{D}_{SpamDetection}$; and $\mathcal{D}_{Covid\text{-}HeRA}$. $\mathcal{D}_{Covid\text{-}HeRA}$ is a dataset that categorized social media posts based on health risk assessment. The *possibly severe* and *highly severe* are the top two highest-risk categories of social media posts which if followed may have a severe health-related impact. We notice that in all three datasets, inappropriate classes (e.g., abusive, spam, hateful, severe, and highly severe) exhibited greater retrieval loss than non-

Fig. 1: Retrieval loss across different classes.




inappropriate classes. Sometimes, the retrieval loss between the appropriate and non-inappropriate class is stark. It is possible that platform moderation may affect the inappropriate classes disproportionately triggering larger retrieval loss. For instance, in $\mathcal{D}_{SpamDetection}$, the *spam* class loses 96.93% data whereas the *not spam* class loses 26.75%.

*1) Quantifying Retrieval Imbalance:* As we note the imbalance in retrieval loss across classes, we next formalize a quantifiable measure of retrieval imbalance. Consider a dataset $\mathcal{D}$ has $k$ classes denoted by $\mathcal{C}_1$, $\mathcal{C}_2$, ..., $\mathcal{C}_k$. Consider the retrieval loss for each of these classes are $l_1, l_2, \ldots, l_k$, respectively. We first construct the probability vector $[\frac{l_1}{\Sigma_{i=1}^{k} l_i}, \frac{l_2}{\Sigma_{i=1}^{k} l_i}, \ldots, \frac{l_k}{\Sigma_{i=1}^{k} l_i}]$. For $\mathcal{D}$, $RetrievalImbalance_{KL}(\mathcal{D}) = \text{KL}([\frac{l_1}{\Sigma_{i=1}^{k} l_i}, \frac{l_2}{\Sigma_{i=1}^{k} l_i}, \ldots, \frac{l_k}{\Sigma_{i=1}^{k} l_i}], [\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}])$, i.e., the KL divergence of the retrieval loss probability vector with discrete uniform distribution. A higher value indicates a greater imbalance in the retrieval loss. As indicated in Table I, of the three datasets shown in Figure 1, $\mathcal{D}_{SpamDetection}$ shows the highest $RetrievalImbalance_{KL}$ value.

### C. Impact on Supervised Solutions

We observe that retrieval loss can disproportionately impact minority classes. Class imbalance is a well-studied machine learning challenge [68] and extreme class imbalance often calls for specialized solutions. We now present an exploratory study on the impact of imbalance in retrieval loss on supervised solutions. We consider two retrieval loss settings: balanced and imbalanced. In a balanced setting, all classes lose data at the same rate. In an imbalanced setting, individual classes lose data according to the retrieval loss rate of that class.

Our two retrieval loss settings will be easy to explain through an example. Consider we have a dataset with two classes $\mathcal{A}$ and $\mathcal{B}$ with individual retrieval loss 20% and 80% respectively. We are told that the dataset has lost 100 instances. Under the balanced retrieval loss assumption, we assume both classes lost 50 instances apiece. Under the imbalances loss assumption, we would assume $\mathcal{A}$ lost 20 instances where $\mathcal{B}$ lost 80.

For a given dataset, we first compute the overall retrieval loss $l$ and assume that the dataset loses $l\%$ of the currently available instances at every step. At each step, for both balanced and imbalanced retrieval loss settings, we keep the number of overall lost instances fixed for a fair comparison. However, we apportion the lost instances differently for the two retrieval loss settings as described in the previous paragraph.

### D. Model Training

We consider three datasets to train our models: $\mathcal{D}_{CivilUnrest}$; $\mathcal{D}_{CyberThreatDetection}$; and $\mathcal{D}_{COVIDCQStance}$. $\mathcal{D}_{CivilUnrest}$ [56] is a dataset used to predict civil unrest events in Australian cities from social media data. $\mathcal{D}_{CyberThreatDetection}$ [62] is a dataset to detect cyber threats relevant to IT infrastructure. $\mathcal{D}_{COVIDCQStance}$ [38] is a stance dataset developed with the objective to identify and understand Twitter users' stances on whether chloroquine and hydroxychloroquine could be used as a cure for the coronavirus.

We use the CatBoost [69] model to understand the impact of retrieval loss. CatBoost is a gradient-boosting tree-based model that handles the categorical and text features without the need of pre-processing them. The algorithm is used widely for search, self-driving, weather prediction, and many other applications.

We train our models on the datasets we have fetched by applying both the balanced as well as the imbalanced retrieval loss. For balanced data loss, we compute the overall percentage of loss of tweets with the help of the total number of tweets and retrieved tweets and use it for all the classes of the dataset. Whereas in the case of imbalanced retrieval loss, we decide the data loss percentage value for each specific class. We keep initially 10% of the dataset aside for evaluation purposes and report the F1-score accordingly.

### E. Experiment Setup and Evaluation

We ran our experiments extensively on Google Colab with our environment configuration having 12.7GB of RAM, 107.7GB of disk space, Python version 3.10.12, scikit-learn version 1.2.2, and catboost version 1.2 respectively. We use F1-Score due to the presence of class imbalance in our datasets.

### F. Model Performance

Table II summarizes our modeling experiment. For $\mathcal{D}_{CivilUnrest}$ and $\mathcal{D}_{CyberThreatDetection}$, we do not notice any significant performance difference between the balanced and imbalanced retrieval loss settings. However, for $\mathcal{D}_{COVIDCQStance}$, we observe that the two settings result in markedly different performances. Hence, the primary takeaway is the retrieval loss setting may affect performance under certain conditions. It merits a deeper investigation to understand how different factors (e.g., the difficulty level of a particular class, the extent of retrieval loss imbalance) contribute to the performance gap between the two settings.

## V. Conclusions and Discussion

In this paper, we investigate a problem barely explored in information science – transience of social web datasets. Via a comprehensive analysis of 40 well-known Twitter datasets published in well-regarded computer science venues, we conclude that

1) The shelf-life of social web datasets is shortlived. Several datasets exhibit considerable retrieval loss some to the extent that training models on them would be infeasible.
2) The minority class, which often signals inappropriate content (e.g., hate speech, spam, misinformation, cyber threat), suffers from larger retrieval loss than the majority class.
3) The performance of supervised solutions may get impacted by the retrieval loss assumptions.

Our work raises the following important points to ponder upon.

- ***Robust evaluation framework for social web datasets.*** The fleeting nature of the social web datasets indicates performance evaluation on social web datasets requires more robustness checks. One possible direction could be reporting average performance over a broad range of retrieval loss scenarios. Many of the cutting-edge learning algorithms are data-hungry. Comparing performance under this new lens can lead to finding more robust algorithms suitable for real-world settings.
- ***Curating robust social web datasets.*** Similar to robust evaluation, during the curation step, if we curate our datasets in a transience-aware way, lending more redundancy to minority classes that potentially signal inappropriate content, the dataset will remain useful for a longer period of time.
- ***Sequential Learning Settings and Learning Theory.*** Our paper raises an important point that data available now may not be data available later on. Sequential learning settings such as active learning [70] and their variants that considered other relaxed assumptions [71] can be further reimagined under this new assumption. We hope this research will draw the attention of ML theorists to lend theoretical grounding to this setting.
- ***Social web preservation.*** While the Internet Archive presents a reliable snapshot of web content, social media content preservation is an uphill task given its gigantic growth. An important discussion point for AI ethicists could be how to separate content from the user in a privacy-preserving way. Consider a user deleting a tweet that was present in a dataset. At this point, is it ethical for the dataset publisher to remove all identifiable information about the user from the tweet but release the content (e.g., the tweet text)? If yes, then a viable path for the dataset publisher could be updating deleted tweets with non-identifiable content information.
- ***Beyond Twitter.*** Finally, as we already mentioned, with the deprecation of free academic Twitter API, analyzing the shelf-life of Twitter datasets will hardly inform Twitter-specific dataset practices in the future. However, dataset permanence will remain a relevant issue no matter which social media platform fills up the scientific research void.

## VI. Ethical Statement

In our research, we place high regard for the rights of users to be forgotten, ensuring their privacy and data protection. Our research only reports aggregate analysis on labels; no user-specific study – aggregate or on specific individual accounts is conducted. It is important to emphasize that our objective is not to release new datasets but rather to conduct rigorous audits of previously published and extensively cited datasets. Through these audits, we aim to contribute to a comprehensive understanding of reproducibility challenges, promoting transparency, and upholding research integrity in the field.

## References


[1] D. Demszky, N. Garg, R. Voigt, J. Zou, J. Shapiro, M. Gentzkow, and D. Jurafsky, "Analyzing polarization in social media: Method and application to tweets on 21 mass shootings," in *NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 2970–3005.

[2] A. R. KhudaBukhsh, R. Sarkar, M. S. Kamlet, and T. Mitchell, "We don't speak the same language: Interpreting polarization through machine translation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 893–14 901.

[3] S. Palakodety, A. R. KhudaBukhsh, and J. G. Carbonell, "Hope speech detection: A computational analysis of the voice of peace," in *ECAI 2020 - 24th European Conference on Artificial Intelligence*, ser. Frontiers in Artificial Intelligence and Applications, vol. 325. IOS Press, 2020, pp. 1881–1889.

[4] R. Kouzy, J. Abi Jaoude, A. Kraitem, M. B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E. W. Akl, and K. Baddour, "Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter," *Cureus*, vol. 12, no. 3, 2020.

[5] P. Juneja, M. M. Bhuiyan, and T. Mitra, "Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on youtube," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023*. ACM, 2023, pp. 545:1–545:22.

[6] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016*. The Association for Computational Linguistics, 2016, pp. 88–93.

[7] P. Saha, K. Garimella, N. K. Kalyan, S. K. Pandey, P. M. Meher, B. Mathew, and A. Mukherjee, "On the rise of fear speech in online social media," *Proceedings of the National Academy of Sciences*, vol. 120, no. 11, p. e2212270120, 2023.

TABLE II: Modelling Result

| Dataset | Retrival Loss Type | Loss (value) | Step | F1 Score |
|---|---|---|---|---|
| $\mathcal{D}_{CivilUnrest}$ | Symmetric | 0.3589 | 1 | False: 0.9587<br>True: 0.7664 |
| $\mathcal{D}_{CivilUnrest}$ | Asymmetric | False: 0.3671<br>True: 0.3241 | 1 | False: 0.9604<br>True: 0.7736 |
| $\mathcal{D}_{CivilUnrest}$ | Symmetric | 0.4878 | 2 | False: 0.9498<br>True: 0.7368 |
| $\mathcal{D}_{CivilUnrest}$ | Asymmetric | False: 0.5019<br>True: 0.4290 | 2 | False: 0.9571<br>True: 0.7547 |
| $\mathcal{D}_{CyberThreatDetection}$ | Symmetric | 0.1096 | 1 | yes: 0.8719<br>no: 0.9118 |
| $\mathcal{D}_{CyberThreatDetection}$ | Asymmetric | yes: 0.1435<br>no: 0.0831 | 1 | yes: 0.8747<br>no: 0.9105 |
| $\mathcal{D}_{CyberThreatDetection}$ | Symmetric | 0.1216 | 2 | yes: 0.8703<br>no: 0.9094 |
| $\mathcal{D}_{CyberThreatDetection}$ | Asymmetric | yes: 0.1640<br>no: 0.0900 | 2 | yes: 0.8681<br>no: 0.9104 |
| $\mathcal{D}_{COVIDCQStance}$ | Symmetric | 0.3792 | 1 | Neutral: 0.3800<br>Against: 0.7680<br>Favor: 0.6593 |
| $\mathcal{D}_{COVIDCQStance}$ | Asymmetric | Neutral: 0.3115<br>Against: 0.2239<br>Favor: 0.5139 | 1 | Neutral: 0.4681<br>Against: 0.7884<br>Favor: 0.6025 |
| $\mathcal{D}_{COVIDCQStance}$ | Symmetric | 0.5231 | 2 | Neutral: 0.3261<br>Against: 0.7545<br>Favor: 0.6452 |
| $\mathcal{D}_{COVIDCQStance}$ | Asymmetric | Neutral: 0.4085<br>Against: 0.2740<br>Favor: 0.7783 | 2 | Neutral: 0.3960<br>Against: 0.6941<br>Favor: 0.2585 |


[8] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proc. ACM Hum. Comput. Interact.*, vol. 1, no. CSCW, pp. 31:1–31:22, 2017.

[9] E. Huet, "Google Finally Shuts Down Orkut, Its First Social Network," https://www.forbes.com/sites/ellenhuet/2014/06/30/google-kills-orkut/?sh=11aad44e634b, 2014, forbes.

[10] D. Bamman, B. O'Connor, and N. A. Smith, "Censorship and deletion practices in chinese social media," *First Monday*, vol. 17, no. 3, 2012.

[11] K. Hines, "Reddit Follows Twitter's Lead, Announces Paid Access To Data API ," https://www.searchenginejournal.com/reddit-paid-api/485172/#close, 2023, search Engine Journal.

[12] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn, "What is Gab: A bastion of free speech or an alt-right echo chamber," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1007–1014.

[13] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[14] H. Ledford, "Researchers scramble as Twitter plans to end free data access," https://www.nature.com/articles/d41586-023-00460-z, 2023, nature.

[15] J. Taylor and D. Milmo, "How Twitter's new drastic changes will affect what users can view on the site," https://www.theguardian.com/technology/2023/jul/03/how-twitter-new-changes-will-affect-users-rate-limited-limit-exceeded-restrictions, 2023, the Guardian.

[16] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of the international AAAI conference on web and social media*, vol. 4, no. 1, 2010, pp. 122–129.

[17] A. Abilov, Y. Hua, H. Matatov, O. Amir, and M. Naaman, "Voterfraud2020: a multi-modal dataset of election fraud claims on Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 901–912.

[18] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.

[19] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[20] K. Lakshminarayan, S. A. Harp, R. P. Goldman, T. Samad *et al.*, "Imputation of missing data using machine learning techniques." in *KDD*, vol. 96, 1996.

[21] P. Raja and K. Thangavel, "Missing value imputation using unsupervised machine learning techniques," *Soft Computing*, vol. 24, no. 6, pp. 4361–4392, 2020.

[22] R. Alkhalifa, E. Kochkina, and A. Zubiaga, "Building for tomorrow: Assessing the temporal persistence of text classifiers," *Information Processing & Management*, vol. 60, no. 2, p. 103200, 2023.

[23] S. Liu and A. Ritter, "Do CoNLL-2003 named entity taggers still work well in 2023?" in *ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 8254–8271.

[24] L. M. Stevens, B. J. Mortazavi, R. C. Deo, L. Curtis, and D. P. Kao, "Recommendations for reporting machine learning analyses in clinical research," *Circulation: Cardiovascular Quality and Outcomes*, vol. 13, no. 10, p. e006556, 2020.

[25] "Auditing and robustifying covid-19 misinformation datasets via anticontent sampling," vol. 37, pp. 15 260–15 268, Jun. 2023.

[26] M. Das, A. Dash, S. D. Jaiswal, B. Mathew, P. Saha, and A. Mukherjee, "Platform governance: Past, present and future," *GetMobile Mob. Comput. Commun.*, vol. 26, no. 1, pp. 14–20, 2022.

[27] S. Park, S. Kim, and Y.-s. Lim, "Fairness audit of machine learning models with confidential computing," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3488–3499.

[28] K. McElwee., "Fortune 100 response to 2020 blm protests." [Online]. Available: https://catalog.docnow.io/datasets/20210129-fortune-100-response-to-2020-blm-protests/

[29] B. Jules., "Jessica krug aka jess la bombalera." [Online]. Available: https://catalog.docnow.io/datasets/20201029-jessica-krug-aka-jess-la-bombalera/

[30] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of Twitter abusive behavior," in *Proceedings of the international AAAI conference on web and social media*, vol. 12, no. 1, 2018.

[31] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks." Association for Computational Linguistics, 2018.

[32] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[33] C. H. Yoo, S. Palakodety, R. Sarkar, and A. KhudaBukhsh, "Empathy and hope: Resource transfer to model inter-country social media dynamics," in *Proceedings of the 1st Workshop on NLP for Positive Impact*. Online: Association for Computational Linguistics, Aug. 2021, pp. 125–134.

[34] S. Swamy, A. Ritter, and M. de Marneffe, ""i have a feeling trump will win.....": Forecasting Winners and Losers from User Predictions on Twitter," in *EMNLP 2017*, M. Palmer, R. Hwa, and S. Riedel, Eds. Association for Computational Linguistics, 2017, pp. 1583–1592.

[35] G. Biamby, G. Luo, T. Darrell, and A. Rohrbach, "Twitter-comms: Detecting climate, covid, and military multimodal misinformation," in *NAACL 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds. Association for Computational Linguistics, 2022, pp. 1530–1549.

[36] C. Conforti, J. Berndt, M. T. Pilehvar, C. Giannitsarou, F. Toxvaerd, and N. Collier, "Will-they-won't-they: A very large dataset for stance detection on twitter," in *ACL 2020*. Association for Computational Linguistics, 2020, pp. 1715–1724.

[37] H. Chen, Z. Zhu, F. Qi, Y. Ye, Z. Liu, M. Sun, and J. Jin, "Country image in covid-19 pandemic: A case study of china," *IEEE Transactions on Big Data*, vol. 7, no. 1, pp. 81–92, 2020.

[38] E. C. Mutlu, T. Oghaz, J. Jasser, E. Tutunculer, A. Rajabi, A. Tayebi, O. Ozmen, and I. Garibay, "A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19," *Data in brief*, vol. 33, p. 106401, 2020.

[39] R. Chang, C. Lai, K. Chang, and C. Lin, "Dataset of propaganda techniques of the state-sponsored information operation of the people's republic of china," *CoRR*, vol. abs/2106.07544, 2021. [Online]. Available: https://arxiv.org/abs/2106.07544

[40] B. Barz, K. Schröter, A.-C. Kra, and J. Denzler, "Finding relevant flood images on twitter using content-based filters," in *Pattern Recognition. ICPR International Workshops and Challenges*. Springer, 2021, pp. 5–14.

[41] T. Sahni, C. Chandak, N. R. Chedeti, and M. Singh, "Efficient twitter sentiment classification using subjective distant supervision," in *2017 9th International Conference on Communication Systems and Networks (COMSNETS)*. IEEE, 2017, pp. 548–553.

[42] A. Abilov, Y. Hua, H. Matatov, O. Amir, and M. Naaman, "Voterfraud2020: a multi-modal dataset of election fraud claims on Twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, 2021, pp. 901–912.

[43] J. H. Lau, L. Chi, K. Tran, and T. Cohn, "End-to-end network for twitter geolocation prediction and hashing," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017*, 2017, pp. 744–753.

[44] S. Dabiri, "Tweets with traffic-related labels for developing a twitter-based traffic information system," 2018.

[45] B. Wang, M. Liakata, A. Zubiaga, R. Procter, and E. Jensen, "Smile: Twitter emotion classification using domain adaptation," in *CEUR Workshop Proceedings*, vol. 1619. Sun SITE Central Europe, 2016, pp. 15–21.

[46] P. Khan, S. Siddiqui, I. Razzak, A. Dengel, and S. Ahmed, "Improving health mentioning classification of tweets using contrastive adversarial training," 03 2022.

[47] K. Kawintiranon, L. Singh, and C. Budak, "Traditional and context-specific spam detection in low resource settings," *Machine Learning*, vol. 111, no. 7, pp. 2515–2536, 2022.

[48] T. Khan and A. Michalas, "Trust and believe-should we? evaluating the trustworthiness of Twitter users," in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2020, pp. 1791–1800.

[49] P. Vashisth and K. Meehan, "Gender classification using Twitter text data," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.

[50] M. Kejriwal, "Emoji Extractions from Geotagged Twitter Data." 2020.

[51] R. Pranesh, M. Farokhenajd, A. Shekhar, and G. Vargas-Solar, "CMTA: COVID-19 misinformation multilingual analysis on Twitter," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*. Online: Association for Computational Linguistics, Aug. 2021, pp. 270–283. [Online]. Available: https://aclanthology.org/2021.acl-srw.28

[52] S. A. Curiskis, B. L. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit," vol. 57, 2020, p. 102034.

[53] X. Xu, R. F. Manrique, and B. P. Nunes, "Rip emojis and words to contextualize mourning on twitter," *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 2021.

[54] S. A. Memon and K. M. Carley, "Cmu-miscov19: A novel twitter dataset for characterizing covid-19 misinformation," 2020.

[55] D. R. A. B. Apriantoni, Apriantoni; Purwitasari, "Indonesian tweets dataset for identifying emotion changes among twitter users following the onset of the covid-19."

[56] L. Mitchell, "Civil unrest event-relevant twitter classifier training data," 2018.

[57] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[58] S. L. Blodgett, J. Wei, and B. T. O'Connor, "A dataset and classifier for recognizing social media english," in *NUT@EMNLP*, 2017.

[59] P. Mongeon, T. D. Bowman, and R. Costas, "An open dataset of scholars on twitter," *ArXiv*, vol. abs/2208.11065, 2022.

[60] S. Menini, A. P. Aprosio, and S. Tonelli, "Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection," *ArXiv*, vol. abs/2103.14916, 2021.

[61] "Twitter conversations dataset." [Online]. Available: https://paperswithcode.com/dataset/twitter-conversations-dataset

[62] N. Dionísio, F. Alves, P. M. Ferreira, and A. N. Bessani, "Cyberthreat detection from twitter using deep neural networks," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019.

[63] R. Bansal, W. S. Paka, Nidhi, S. Sengupta, and T. Chakraborty, "Combining exogenous and endogenous signals with a semi-supervised co-attention network for early detection of covid-19 fake tweets," *ArXiv*, vol. abs/2104.05321, 2021.

[64] A. K. Gautam, P. Mathur, R. Gosangi, D. Mahata, R. Sawhney, and R. R. Shah, "#metooma: Multi-aspect annotations of tweets related to the metoo movement," in *International Conference on Web and Social Media*, 2019.

[65] A. Dharawat, I. Lourentzou, A. Morales, and C. Zhai, "Drink bleach or do what now? covid-hera: A study of risk-informed health decision making in the presence of covid-19 misinformation," in *International Conference on Web and Social Media*, 2020.

[66] J. Zeng, J. Li, Y. He, C. Gao, M. R. Lyu, and I. King, "What you say and how you say it: Joint modeling of topics and discourse in microblog conversations," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 267–281, 2019.

[67] Z. Talat and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *North American Chapter of the Association for Computational Linguistics*, 2016.

[68] S. M. Abd Elrahman and A. Abraham, "A review of class imbalance problem," *Journal of Network and Innovative Computing*, vol. 1, no. 2013, pp. 332–340, 2013.

[69] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems: NeurIPS 2018*, 2018, pp. 6639–6649.
[70] B. Settles, "Active learning literature survey," 2009.
[71] P. Donmez and J. G. Carbonell, "Proactive learning: cost-sensitive active learning with multiple imperfect oracles," in *Proceedings of the 17th ACM conference on Information and knowledge management*, 2008, pp. 619–628.