9^{th} International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Comparing Methods of Creating a Random Sample of Twitter Users from a Country

Keywords: Twitter, Sampling, Data Quality, Nationally Representative, Twitter API

Extended Abstract

Twitter data has been widely used by researchers across various social and computer sciences disciplines. One of the key challenges in working with Twitter data is to obtain a random sample of users from a country. The goal is usually to get a platform or population representative sample of users. However, there are at least two major challenges in obtaining random sample of users from a country: 1) several methods has been proposed in the literature and it is not clear which one is the best, and 2) the extent to which these samples are representative of the population or even platform is questionable. Since Twitter has not been transparent about how data sampling is performed, early research on sampling from Twitter were focused on reverse-engineering of its sampling [1], and demonstrating its results being unrepresentative [1] and bias [2].

There are at least four popular methods to devise a random sample of Twitter users for a specific country. First, Twitter provides 1% of all tweets worldwide in real-time through its free stream API. One can collect this stream for a time period, filter for language and country of interest, obtain a list of users who posted tweets and sample from them (e.g. [3]). Second and third, it is possible to use Twitter's search-tweet API and query for a specific language or country, and after ingesting tweets for a period of time, filter for a country or language respectively (e.g. [4]). Fourth, Twitter API allows to query based on a country name or its bounding box coordinates and researchers used it to get a random sample of a country's Twitter users (e.g. [5]). The extent to which these four methods produce similar results and which one is more representative of population is mostly unexplored.

In this paper, we use these four methods to collect data on US and Switzerland and compare the methods with respect to their results' size and representativeness of topics, events, and population. We use Pew Research Center's 'Social Media Use in 2021' report and US Census data to obtain ground-truth data of American users, and the CensusHub [1] data for Swiss users. We use the M3 model [6] to infer age and gender of Twitter users. The model operates on 32 languages and only requires users' profile information such as profile picture, screen name, and biography. As for location, we follow previous research and use self-reported location of users or their close friends to estimate the location of users at the state or province level [4]. In all four methods, we first take a random sample of 30K users from the corpus and remove bots and non-individual accounts (e.g. celebrities, news anchors, journalists, organizations, politicians, etc.), and then randomly sample 5K users from the remaining users.

We compare the results of each nationally-random sample creation of Twitter users method according to three categories of criteria including 1) tweet-level, 2) account-level, and 3) population-level criteria. Table 1 lists all criteria. Tweet- and account-level criteria are self-explanatory. The population-level criteria, reports five different mean absolute percentage errors (MAPE) for the prediction task of estimating the population of a country using the nationally-representative sample of Twitter users obtained from the model proposed in [6].

[1] https://ec.europa.eu/CensusHub2/

Preliminary results for US show that the 'language query' and '1% stream' methods produce much less tweets compared to 'bounding-box' and 'country query' methods (Table 2). The results further show that the 'country query' and 'bounding box' methods do not use back-end users' information to determine their actual location and instead use their public information to retrieve tweets for a specific country or bounding-box queries. Moreover, the bounding-box method locates a majority of tweets at the centroid of a given country. This probably happens for those users which their self-reported location only include their country, and therefore, Twitter locates their coordinates at the centroid of that country.

In terms of the distribution of tweets, tweets per day, likes, account creation dates, number of followers, and number of friends, the results show no major difference between the four Twitter sampling methods (Fig 1). Same pattern holds true for the distribution of bots (Fig 2) and the distributions of age and gender across the four sampling methods (Fig 3). In terms of the location, although the 5K user samples from all four methods generated at least one user in each US state (Fig 4), they all failed to yield enough users in all states to cover various combinations of age, gender, and location, which is required to compute the inclusion probabilities of users and create a representative sample of the population. Therefore, we are conducting a new round of data collection to be able to increase the sample size from 5K to 10K.

Nonetheless, we were able to compute inclusion probabilities at the region level (i.e. Northeast, Southwest, West, Southeast, and Midwest). Figure 5 illustrates various mean absolute percentage error (MAPE) for the task of inferring the US population using representative Twitter samples created from each of the four sampling methods (this is a validation task proposed in [6]). The results of the leave-one-region-out evaluation (Fig 5) show a clear benefit to use 1) the bounding-box sampling method, and 2) all inferred demographics (as opposed to only one) for creating the best nationally-representative sample of Twitter users.

Our main conclusions, based on the preliminary results from US, are that 1) the four Twitter sampling methods are pretty similar in terms of tweet- and user-level evaluation metrics, 2) with a 5K random sample of users, none of the four sampling methods provide enough data to cover all demographic in all US states, however, all of them are suitable for computing the inclusion probabilities of all demographics at the US region level, and 3) in the prediction task of inferring US population from Twitter data, the bounding-box sampling method obtains the highest accuracy.

References

[1] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 400–408, 2013.

[2] Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 505–514, 2014.

[3] Chengcheng Shao, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. Anatomy of an online misinformation network. *Plos one*, 13(4):e0196087, 2018.

[4] Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901, 2019.

[5] Christopher Barrie and Alexandra Siegel. Kingdom of trolls? influence operations in the saudi twittersphere. *Journal of Quantitative Description: Digital Media*, 1:1173–1199, 2021.

[6] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference*, pages 2056–2067, 2019.

Table 1: List of evaluation criteria for comparing various methods of creating a national-sample of Twitter users.

Category	Criteria	Description
Tweet-Level	Number of tweets	Total number of collected tweets.
	Average tweet per account	Average number of tweets per account.
	Relevant language	Share of tweets in country-specific languages.
	Distribution of tweets per day	Distribution of number of tweets in each day.
Account-Level	Number of accounts	Number of unique accounts.
	Number of likes	For each account, distributions of last tweets likes number that crawled in crawling time period.
	Account creation date	Distributions of account creation time.
	Numbers of followers	Distributions of the numbers of followers.
	Number of friends	Distributions of the numbers of friends.
	Bot score	Distribution of bot similarity score of accounts.
	Number of users per city	Geographical distribution of accounts in each city.
Population-Level	MAPE (Mean Absolute Percentage Error) where N $\sim$ M	Base model that uses only the total population count from the census (N) and Twitter (M).
	MAPE where N $\sim \sum_g$ M(g)	Assumes homogeneity and uses gender marginal counts only (i.e., the total counts of males and females not broken down by ages).
	MAPE where N $\sim \sum_a$ M(a)	Assumes homogeneity and uses age marginal counts only.
	MAPE where N $\sim \sum_{a,g}$ M(a, g)	Assumes homogeneity and uses the joint histograms inferred from Twitter but only the total population values from the census.
	MAPE where log N(a, g) $\sim$ log M(a, g) + a + g	Assumes heterogeneity and uses the joint histograms inferred from Twitter and the joint histograms from the census (i.e. uses all demographics of users).

Table 2: Tweet-level evaluation results.

	BB	*Loc*	*Lang*	*1%*
Number of tweets	18,181,424	18,804,550	4,508,702	174,084
Number of accounts	728,028	738,595	425,041	94,250
Average tweets per accounts	24.974	25.46	10.608	1.847
Relevant language	0.823	0.808	1	0.807
30K sample				
Number of tweets	801,320	737,608	322,296	55,290
Average tweets per accounts	23.828	24.587	10.743	1.843

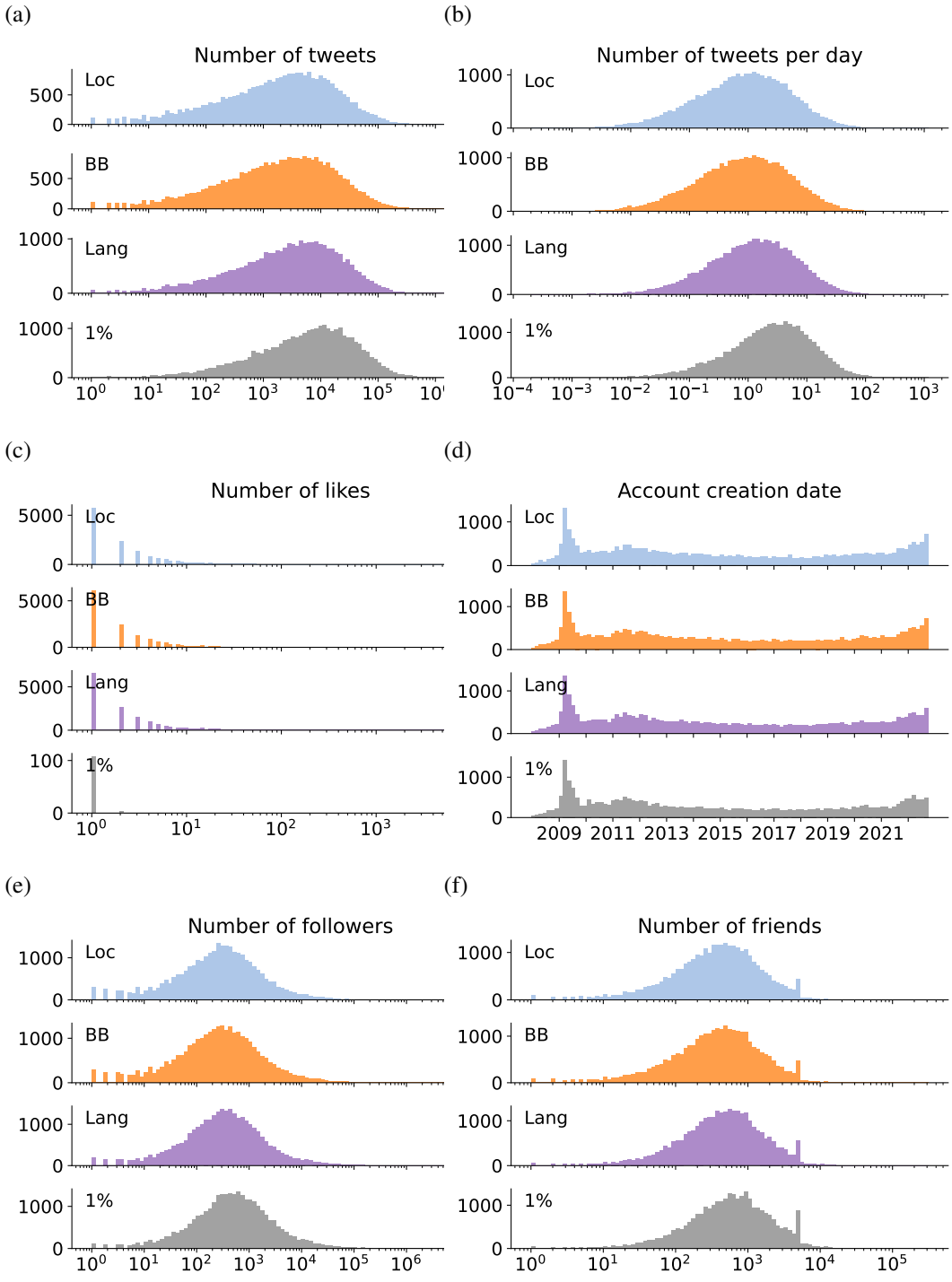

Figure 1: Distributions of (a) number of tweets; (b) average number of tweets per day; (c) number of likes; (d) account creation date; (e) number of followers; and (f) number of friends for different groups.

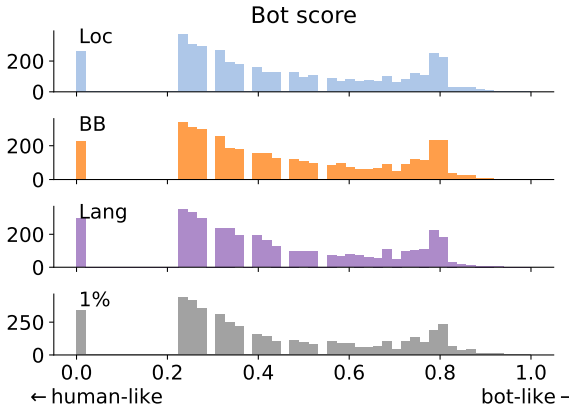

Figure 2: Bot score distributions of different account groups. We annotate the percentage of accounts having bot-score above 0.8.

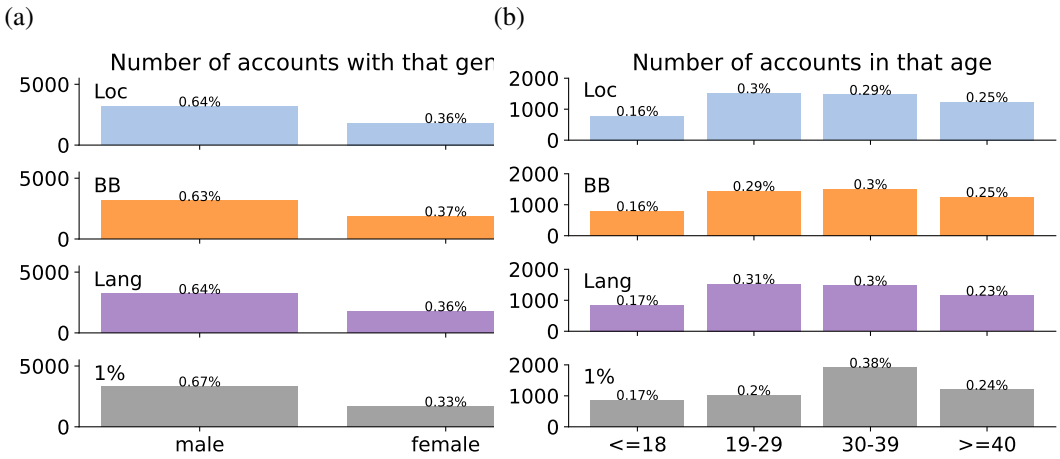

Figure 3: Number of accounts in different (a) genders; (b) aged for each groups. also, the percentage of each bar has shown

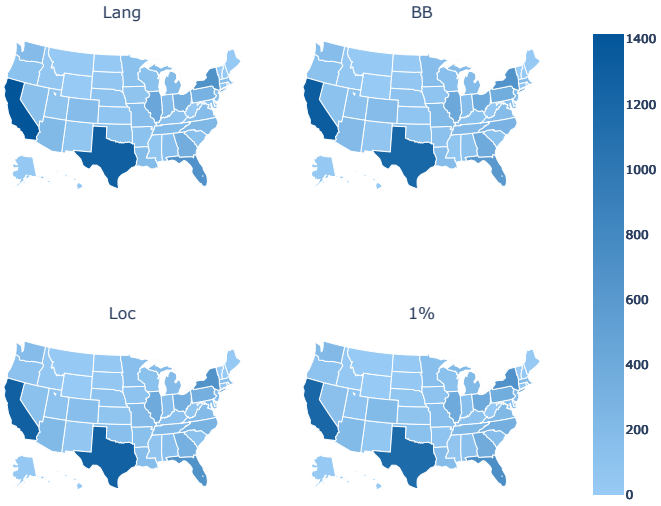

Figure 4: Number of accounts located in each state

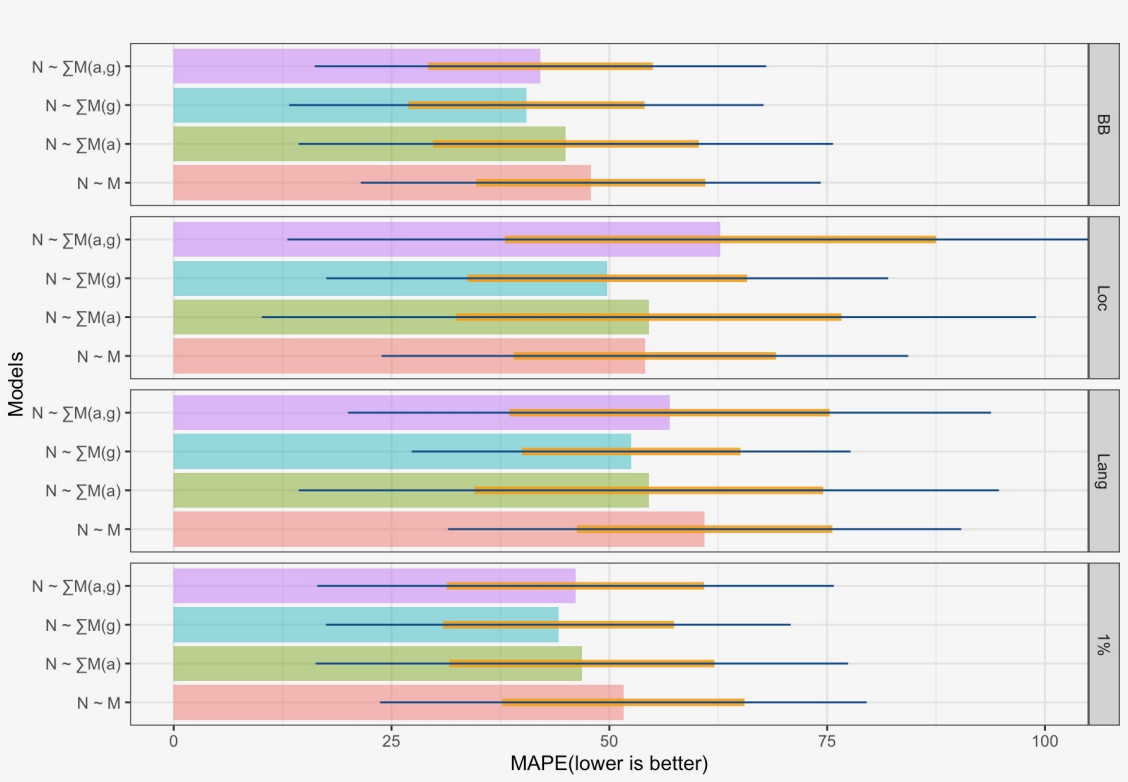

Figure 5: Performance on leave-one-Region-out population inference across different debiasing models. Bars show mean MAPE (N) with 90% confidence intervals (blue line) and standard error (orange line). The above results test predictions when one US region is left out and its population is predicted. The results show a clear benefit to use 1) the bounding-box sampling method, and 2) all inferred demographics (as opposed to only one) for creating the best nationally-representative sample of Twitter users. In boundig-box and language query sampling methods, even when joint distribution are not available by the census, inferring joint distributions from social media data with the model N $\sim \sum_{a,g}$ M(a, g) provides better results in accuracy in population prediction tasks compared to the baseline without debiasing.