Facebook URLs-Training Codebook

Christina DeGregorio, Gary King, Solomon Messing, Zagreb Mukerjee, Chaya Nayak,
Nathaniel Persily, Bogdan State, Arjun Wilkins

18 June 2019

This codebook results from a collaboration between Facebook and Social Science One, originally prepared for Social Science One grantees. It describes the training URLs dataset, including its scope, structure, and fields.

Citation

DeGregorio, Christina; King, Gary; Messing, Solomon; Mukerjee, Zagreb; Nayak, Chaya; Persily, Nathaniel; State, Bogdan; Wilkins, Arjun, 2019, "Facebook URL Shares-Training."

Status

These data may only be accessed through a special privacy-preserving computational infrastructure built by Facebook. No data may leave the system.

Data Access

To obtain access to these data, see SocialScience.one. No other means of access is allowed.

Purpose and Summary

We are in the process of releasing a “URLs-light data set, which constitutes a prelude to the “Full URLs” data set, in which we hope to make demographic aggregates available in a privacy-preserving way as well.

The light data set will need to be analyzed in Facebook's analysis environment, which protects user privacy while making it possible to conduct analysis for publication. To help researchers acclimate to data analysis under the constraints imposed by this system, we are releasing this "training" data set.

The key constraint imposed by this system is a "privacy budget," which quantifies and limits the total amount of formally defined private information that can be accessed through the system. So that researchers do not exceed this "privacy budget" in the process of exploring the URLs light data, we are releasing this modified "training" version of the URLs-light data set, in which fields containing (aggregated) user engagement statistics have been completely simulated from random number generators. Fields like these that contain aggregated user data will be accessible only via differentially private queries and subject to this privacy budget. The underlying raw data are not accessible.

These training data use the same URLs shared in each country and URL-level descriptors as in the URLs-light dataset. We also add noise to the total number of shares and compute a *noisy* ranking of the top 2000 most shared URLs in each country.

Just as in URLs-light data, this data set describes web page addresses (URLs) that have been shared on Facebook starting January 1, 2017 through to and including February 19, 2019. URLs are included if shared by more than on average 100 (+ Laplacian noise (mean = 0, scale = 5) to minimize information leakage) unique accounts with public privacy settings. We have also post-processed the URLs (as detailed here) to remove potentially private and/or sensitive data.

The unit of analysis for the URLs-light data is the URL. The data set is about 7 gigabytes compressed, comprising approximately 32 million URLs, and about 544 million cell values.

All aggregates attached to the URL are simulated as described below.

Just as in the URLs light data, user accounts deleted prior to this date do not appear in the training dataset. Both datasets includes content that has been taken down due to Community Standards violations, meaning that some URLs that do not work from inside of Facebook may work outside. Other URLs, which worked when the dataset was created, may no longer resolve to a webpage that still exists. Finally, we have tried to remove URLs and all associated aggregated engagement statistics that link to known child exploitative imagery from our dataset. We have also tried to remove URLs, the 'Title' and 'Blurb' in our dataset for known non-consensual intimate imagery, suicide and self-harm, although the associated engagement statistics with these links remain in the dataset.

Infrastructure. Facebook has provided approved researchers access to a system on which they conduct all analyses in a manner that protects user privacy and ensures the integrity of research results.

Aside from certain fields (URL, Domain, Title and Main blurb), data are accessible to researchers only via "differentially private" statistical results. The specific differential privacy algorithms applied, their inferential consequences, and how to avoid or document the statistical biases and properly represent uncertainty will be generated, documented, and shared with researchers prior to data access beginning.

The interface requires the use of a special software library, similar to SparkML, but which only allows analysts to execute queries only in a differentially private way. Differential privacy also means that each research team will have a "privacy budget," which limits the amount of information (i.e., a combination of the granularity and total number of unique queries run) that analysts can extract. Facebook and Social Science One will work with researchers so they understand the consequences of their analyses and to give them the budget they need to conduct their research.

Recommended capabilities. Research teams should have experience working with data sets that do not fit into memory. Researchers will need the capability to query SparkML and understand what kinds of queries are expensive and likely to exhaust the resources of our computing cluster. They will also need to write R and/or Python analysis code that does not exhaust system RAM (e.g., on a modern server with around 64GB RAM). Having at least one individual with experience with SQL/HQL, Python, and Linux is highly recommended.

Field Name	**Data Type**	**Description**	**Is this field differentially private?**	**Is this field Synthetic ?**	**Data Generation Description If Synthetic**
Url_rid	text	a unique URL id created specifically for this data set.	No	No	NA
Clean_url	text	The webpage URL after processing. This is the full URL, not just the domain (e.g., https://www.nytimes.com/2018/07/09/world/asia/thailand-cave-rescue-live-updates.html). URLs that are no longer reachable will persist in the data. Our processing attempts to consolidate different web addresses that point to the same URL and to remove potentially private and/or sensitive data.	No	No	NA

	text	Additional detail available in the previous codebook: https://socialscience.one/files/partnershipone/files/facebook_urls-light_codebook_v1.0.pdf			
Parent_domain	text	parent domain name from the URL (eg. foxnews.com).	No	No	NA
Full_domain	text	full domain name from the URL (eg. www.foxnews.com, video.foxnews.com, nation.foxnews.com, insider.foxnews.com).	No	No	NA
first_post_time	timestamp	The date/time when URL was first posted by a user on Facebook, truncated to 10 minute increments. NOTE: this is rounded to day in the training data set. The exact format is YYYY-MM-DD HH:MM:SS, for example: 2015-12-02 18:10:00	No	No, note rounded to day	NA
first_post_time_unix	unix timestamp	first_post_time translated into unix time---the number of seconds since 1970-01-01 00:00:00, for example: 1449079800. NOTE: this is rounded to day in the training data set	No	No, note rounded to day	NA
Share_title	text	Provided by the author of the URL's content pulled from og:title field in original html if possible).	No	No	NA
Share_main_blurb	text	Provided by the author of the URL's content (pulled from og:description field in original html if possible).	No	No	NA
tpfc_rating	text	If URL was sent to third-party fact-checkers (3pfc), did they rate it (NULL if not) and if so, how did they rate it? Category values include: 'True', 'False', 'Prank Generator' , 'False Headline or Mixture', 'Opinion', 'Satire', 'Not Eligible, 'Not Rated.' Definitions, and a list of fact checkers, are available here: https://www.facebook.com/help/publisher/182222309230722. More information on how news that may be false is selected can be found here:	No	No	NA

		https://www.facebook.com/help/1952307158131536. Only available for some stories, and only available in Argentina, Brazil, Cameroon, Canada, Colombia, Denmark, France, Germany, India, Indonesia, Ireland, Italy, Kenya, Mexico, Middle East and North Africa, Netherlands, Nigeria, Norway, Pakistan, Philippines, Senegal, South Africa, Sweden, Turkey, UK, US. When more than one rating is given to a story, we use Facebook's precedence rules, which can also be found in the more descriptive codebook document: https://doi.org/10.7910/DVN/ZT0WZY			
tpfc_first_fact_check	timestamp	The date-time that article was first fact-checked, if at all. If the article has not been fact checked, this will be NULL. NOTE: this is rounded to day in the training data set. The exact format is YYYY-MM-DD HH:MM:SS, for example: 2015-12-02 18:10:00.	No	No, note rounded to day	NA
tpfc_first_fact_check	unix timestamp	tpfc_first_fact_check translated into unix time---the number of seconds since 1970-01-01 00:00:00, for example: 1449079800. NOTE: this is rounded to day in the training data set	No	No, note rounded to day	NA
total_public_shares	integer	total number of unique accounts that shared the URL with public privacy settings.	Yes	Synthetic	Exp(Exp(U(4, 5))) w/ 80% prob; Exp(Exp(U(4, 15))) w/ 20% prob
total_spam _usr_feedback**	integer	the total number of unique users who reported posts containing the URL as spam.	Yes	Synthetic	Exp(Exp(U(1,5))) w/ 7% prob, 0 otherwise
total_false_news_usr_feedback**	integer	the total number of unique users who reported posts containing the URL as false news.	Yes	Synthetic	Exp(Exp(U(1,5))) w/ 7% prob, 0 otherwise

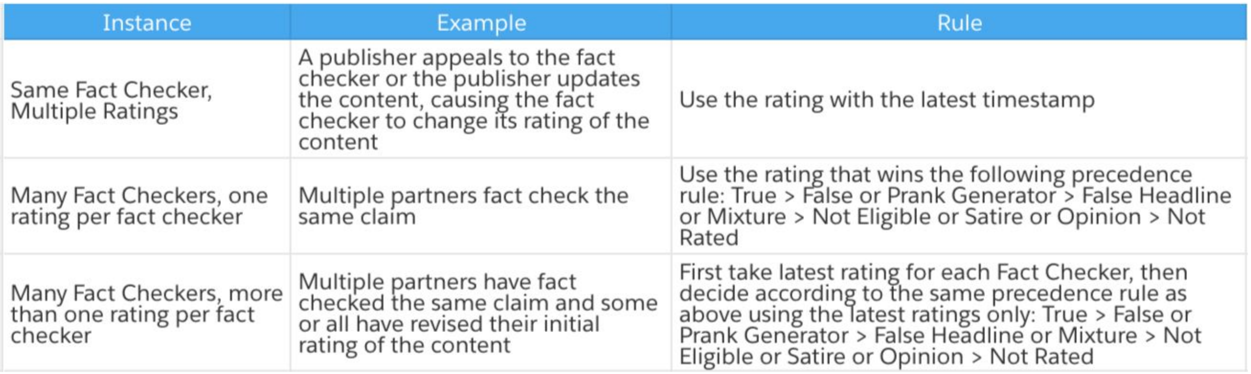
total_hate_speech_usr_feedback**	integer	the total number of unique users who reported posts containing the URL as hate speech.	Yes	Synthetic	Exp(Exp(U(1,5))) w/ 9% prob, 0 otherwise
Prop_share_without_clicks	double	number of users who shared a post but did not click on the link, divided by the total number of unique users who shared the post containing the URL. (Many users share articles without first clicking through to the actual content. Hence, this number may help identify articles that users are sharing without reading, or URLs used in organized campaigns to spread content.)	Yes	Synthetic	.75 -.02*ln(1-rand()) +.05*ln(1-rand()) with ceiling 1 & floor 0
Public_shares_top_country	text	Top user country among users who publicly shared the URL, provided as an ISO 3166-1 alpha-2 letter code	No	No	NA

Third Party Fact Checking (TPFC) Ratings and Precedence Rules

Based on a single false fact-check, Facebook reduces the distribution of a specific piece of false content. Facebook also uses similarity detection methods to identify duplicates of debunked stories and reduce their distribution as well. Facebook uses this as a signal to reduce the overall distribution of Pages and web sites that repeatedly share things found to be false by fact-checkers. Facebook gets signals about false content that can feed back into our machine learning model, helping us more effectively detect potentially false items in the future.

Occasionally, multiple fact-checkers apply different ratings to the same piece of content. In these cases, the more definitive rating takes precedence, e.g. ‘False’ or ‘True’ trumps ‘Mixture’. In rare cases where the two most definitive ratings, ‘True’ and ‘False’, are applied to the same piece of content, ‘True’ takes precedence since we refrain from demoting content rated ‘True’ by a fact-checking partner. Our tpfc_rating incorporates the below precedence rules when decisioning how we handle multiple fact checker ratings for the same URL. It is very rare for multiple fact-checkers to rate the same URL.

For third-party fact-checked content, a fact-checker in a country other than the top public shares country may have rated content if it circulated broadly within their country. For a complete list of our third party fact checkers, please visit this website and this one.



Instance	Example	Rule
Same Fact Checker, Multiple Ratings	A publisher appeals to the fact checker or the publisher updates the content, causing the fact checker to change its rating of the content	Use the rating with the latest timestamp
Many Fact Checkers, one rating per fact checker	Multiple partners fact check the same claim	Use the rating that wins the following precedence rule: True > False or Prank Generator > False Headline or Mixture > Not Eligible or Satire or Opinion > Not Rated
Many Fact Checkers, more than one rating per fact checker	Multiple partners have fact checked the same claim and some or all have revised their initial rating of the content	First take latest rating for each Fact Checker, then decide according to the same precedence rule as above using the latest ratings only: True > False or Prank Generator > False Headline or Mixture > Not Eligible or Satire or Opinion > Not Rated

**User feedback fields: these fields constitute information provided by users, which may not be indicative of actual violations of our Community Standards, and like any survey question or coding exercise, may not be a measure of the concept the researcher intends. For example, for the variable “total_hate_speech_usr_feedback”, users may share URLs to endorse or oppose the content. Endorsements of hate speech violate Facebook's community standards policy, while opposing it does not. Users may also flag content as hate speech because they disagree with it (if they perceive the difference or can distinguish if they do), rather than to actually indicate hate speech, resulting in ambiguous or false positive reports of hate speech, if taken literally. Similar issues apply to other fields.

For “total_spam_usr_feedback”, in contrast to URLs found to contain hate speech (which Facebook deliberately does not block due to the subtleties above), URLs containing content that violates spam policies are blocked from the platform for future sharing.

The data were originally collected or derived from operational information or data sources or otherwise -- and not for research purposes. Features of the dataset may be inaccurate, incomplete, or collected in ways that are not compatible with research goals. Researchers need to adapt their method, research designs, and quantities of interest to the data at hand. Please let us know if you see anything we might be able to adjust in generic ways for everyone.

To learn how Facebook defines and measures key issues, refer to the Community Standards Enforcement Report. The numbers cited in this report are not comparable to the data presented in this RFP as they reference different underlying data. For additional info: Community Standards|Enforcement Report Guide.

Reference

Gary King. 1995. “Replication, Replication.” *PS: Political Science and Politics*, 28, Pp. 444-452. http://j.mp/2oSOXJL