GENERAL SOCIAL MEDIA ARCHIVE
METHODOLOGY REPORT

February 2023 Release

Introduction to the General Social Media Archive

The General Social Media Archive (GSMA) is a public-use data source of social media data curated by NORC at the University of Chicago (NORC) Social Data Collaboratory (SDC), in collaboration with the General Social Survey (GSS). The primary purpose of the GSMA is to provide researchers with social media data to contextualize and complement other public opinion data. The GSMA contains variables created by the GSMA team using native Twitter metadata (e.g., overall tweet volume) and additional derived features (e.g., tweet counts from government organizations, daily averaged sentiment scores) aggregated by day or by the U.S. state from which content was posted. Twitter data was collected from two sources, further described in the following sections: Twitter's 1% sampled stream, and Twitter's GNIP Historical PowerTrack. All information provided in this report is reflective of the February 2023 release based on information gathered during a specific period, detailed for each data source later in this document. An accompanying data dictionary provides variable-level details and is available with the data sets. Please contact the SDC team at SDCcore@norc.org with any questions or requests.

GSMA Social Data Explorer (SDE) Twitter 1%

NORC used data from the Twitter 1% sampled stream collected from January 1, 2019, through February 28, 2022. This data is available through the Social Data Explorer (SDE), an internal tool to NORC that allows researchers to easily search the Twitter 1% data by keywords, view summary statistics, and create reports. The GSMA SDE Twitter 1% extracted six prominent public opinion topics via keyword query, including "marijuana", "abortion", "gay marriage", "gun control", "taxation", and "climate change." Tweets were aggregated by day and are restricted to English-only tweets. Because location data is not readily available for the 1% stream, this data is not constrained by geography.

Metrics

Topics. The six topic areas are denoted in the variable name prefixes: abortion (displayed as the prefix "abortion"), climate change (prefix "climatechange"), gay marriage (prefix "gaymarriage"), gun control (prefix "guncontrol"), marijuana (prefix "mj"), and taxation (prefix "taxation").

Volume. Volume is defined as the number of posts over time, which are counted and aggregated at the day level here. This is denoted by the variable name suffix "vol".

Sentiment. Sentiment measurement comes from VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. VADER provides a positive and negative sentiment score for each tweet. NORC calculated a daily **positive to negative ratio** by dividing the sum of all positive VADER scores for a day by the sum of all negative VADER scores for a day. NORC also calculated an average **compound sentiment score** where the compound score is the daily sum of all positive, negative, and neutral scores which have been normalized between –1 (most extreme negative) and +1 (most extreme positive). These are denoted by the variable name suffixes "vader_pos_to_neg" and "vader_comp_avg", respectively.

Positive to negative ratios with a value of infinity ('Inf') are a direct result of the VADER's negative sum being zero (i.e., any number divided by zero is undefined). For example, if no negatively coded tweets were observed on a given topic on a given day, the sum of the negative VADER scores would be zero. Dividing the sum of positive VADER scores by zero will result in a calculation error. Positive to negative ratios with a value of infinity are assigned a reserve code, or special missing value: .i in SAS and Stata and 888888 in the CSV format. Users should exclude the 888888 reserve code cases before conducting analyses, especially ones related to distributions or summary statistics.

Missing Data

Missing post volume data is present in the GSMA SDE Twitter 1% due to system outages in the real-time data collection methods. The missing volume data spans a total of 16 days, a missing volume rate of approximately 1.4%, on average. NORC chose not to impute missing volume values, because social media traffic can be highly volatile.

Missing VADER sentiment data are also present in the GSMA SDE Twitter 1% largely as a result of missing values for volume. If a daily volume observation is missing, no sentiment information will be available as no tweets are present on which a sentiment measure can be based. If volume is present but a sentiment measure sum is missing, this is likely because VADER identified no sentiment of the desired class for that day (either positive or negative). This is most common when daily volume for a given topic is low as the amount of text on which VADER can measure sentiment is low.

GSMA PowerTrack All Daily

The complete set of publicly-available Twitter data for the topic of marijuana was purchased and collected from Twitter's GNIP Historical PowerTrack (GNIP) for the time range of August 1, 2016, through February 28, 2022. Tweets are aggregated by day and are restricted to English-only tweets originating from the United States or an unspecified location.

To ensure marijuana-related Twitter content was explicitly relevant to the topic of marijuana, NORC developed a Logistic Regression Classifier to classify tweets acquired from GNIP into relevant marijuana content. The classifier was trained using a human coded sample of 1,980 tweets and tested against a human coded sample of 298 tweets. The classifier archived a precision of 0.94, recall of 0.96, and F1 score of 0.95 on the test set.

Metrics

Volume and sentiment metrics are consistent with the GSMA SDE Twitter 1% data. When volume is present, but a sentiment measure sum is missing, this is likely because VADER identified no sentiment of

the desired class for that day (either positive or negative). This is most common when daily volume for a given topic is low as the amount of text on which VADER measures sentiment is low.

Engagement. Engagement is defined as the total number of retweets, replies, or quotes a given tweet received **within 30 days** of the original tweet. It is different from reach, which is a measurement of exposure. This engagement statistic attempt to measure **active participation**. This metric is denoted by the variable name suffix "30d_engage".

Legalization. NORC developed a regular expression (regex) filter to flag tweets for content related to legalization. Critically, this captures any discussion pertaining to the legalization of marijuana and is functionally a subset of overall marijuana-related Twitter content. Moreover, legalization content captures both favorable and unfavorable attitudes toward legalization of marijuana. A sample of 200 tweets was manually coded by humans to determine the regex quality. The regex has an average precision of 0.92, recall of 0.90, and F1 score of 0.90. Marijuana legalization is denoted by the variable name prefix "mjlegal".

Figure Type. To identify accounts as Political Figure, Government Agency, News-Media, or Activist, NORC curated and compiled accounts from numerous Twitter lists (thematic groups of Twitter accounts created by Twitter users) containing Twitter handles of verified accounts into each of these categories. These are associated with the variable name prefixes "political", "govtorg", "newsmedia", and "mjactivist".

Commercial and Bot. NORC developed a regular expression (regex) that examines account profile details to identify overtly commercial accounts. A sample of 200 accounts were manually coded by humans to determine the regex quality. The regex has an average precision of 0.95, recall of 0.94, and F1 score of 0.94. Bot accounts are identified using Botometer Pro API. These are denoted using the variable name prefixes "mjcomm", "comm_not_bot", "bot_not_comm", and "comm_and_bot".

GSMA PowerTrack State Daily

The GSMA PowerTrack State Daily data set is a subset of the GSMA PowerTrack All Daily aggregated to the U.S. State level based on available geolocation information. NORC obtained user location from GNIP Profile geo enrichment which derives user profile location data from longitude and latitude coordinates, where possible, as well as related location metadata. While a large percentage of tweets contain a geolocation, not all tweets contain location information. Approximately 60.2% of tweets obtained from the GNIP data for this marijuana curated data set contain geolocation information. Using this location info when available, NORC can map tweets to a U.S. State via Federal Information Processing Standard (FIPS) code. NORC then aggregated the subset of mappable tweets and calculated daily metrics at the state level.

Metrics

Volume and sentiment metrics are consistent with the GSMA SDE Twitter 1% data. Missing data should be expected and more prevalent in this dataset as the unit of aggregation has increased in granularity.

Engagement and legalization metrics are consistent with the GSMA PowerTrack All Daily data.