THE GENERAL SOCIAL MEDIA ARCHIVE: COMPARING SOCIAL MEDIA DATA WITH THE GENERAL SOCIAL SURVEY ON SIX PRESSING SOCIAL ISSUES

GSS Project Report No. 34

Brian M. Wells, Simon Page, Hy Tran, Andrew Norris

February 15, 2023

INTRODUCTION

The rise of social media over the last fifteen years has provided a new window and data source for measuring and understanding public opinion over time. This abundant data source offers opportunities to collect data on the opinions, attitudes, and behaviors of the public unlike any traditional social science method before.(1) As a form of data collection, social media data can be far less costly to collect than a traditionally designed and fielded survey. Passive capture of social media information can allow for quicker turnaround and more frequent collection.(2) While the majority of social media posts on sites like Facebook and Twitter are generally public, the insufficient and largely unknown coverage of the population along with the strong presence of major media, political, and commercial accounts makes social media data unrepresentative of the population as a whole. However, social media can provide early signals about factors that shape public opinion,(3) help researchers measure and characterize the communication surrounding major events in real time,(4) and examine the use of social media for agenda setting and opinion shaping.(5) Therefore, social media can complement traditional survey data by contextualizing survey findings, often more quickly than surveys, to better understand how a topic is discussed concurrently with survey data.

In the spirit of this exploration, NORC's Social Data Collaboratory, in conjunction with the General Social Survey (GSS) team, have released the General Social Media Archive, a new public-use data source of social media data to contextualize and complement other public opinion data. This report provides an introductory look into the General Social Media Archive and conducts an exploratory examination of its data. We will introduce the archive, and then compare measures of positive to negative sentiment from the archive against similar measures from the GSS on six pressing social issues: marijuana, abortion, gay marriage, gun control, taxation, and climate change. We conclude by examining the benefits of this new data source, discuss initial limitations when pairing its data alongside large-scale surveys, and pose further questions that could be answered with this new resource.

THE GENERAL SOCIAL MEDIA ARCHIVE

The General Social Media Archive (GSMA), curated by NORC's Social Data Collaboratory, currently consists of three data sets: the Social Data Explorer (SDE)[i] Twitter 1%, the PowerTrack All Daily, and the PowerTrack State Daily.

The SDE Twitter 1% primarily consists of data from Twitter's Streaming API.[(6)] The Streaming API delivers a random sample of 1% publicly available tweets in real-time allowing users to identify and track trends and monitor general sentiment. This data set provides researchers with two major metrics: relative overall volume of messaging about a topic, and average sentiment of messages about that topic. The SDE Twitter 1% enables researchers to estimate the number of relevant posts over time (volume), and in relation to major events; and identify the sentiment polarity (positive or negative) of each post. Tweet sentiment is derived from VADER, a parsimonious rule-based model for sentiment analysis of social media text.[(7)] Currently, this data set includes metrics related to six search terms: "marijuana", "abortion", "gay marriage", "gun control", "taxation", and "climate change." The NORC team identified these six as some of the most prominent public opinion topics over the last several years. The GSMA SDE Twitter 1% currently includes tweets from January 2019 through February 2022 aggregated by day. For purposes of the archive, we restrict eligible cases to English only tweets originating from the US or an unspecified location.

The PowerTrack All Daily and PowerTrack State Daily primarily consist of data from Twitter's Historical PowerTrack API aggregated by day.[(8)] The Historical PowerTrack delivers an exhaustive set of publicly available tweets covering a specified historical period that are available at the time of query. We restricted PowerTrack data to English-only tweets originating from the US or an undefined location.[ii] Both the GSMA PowerTrack All Daily and GSMA PowerTrack State Daily data sets are currently focused exclusively on marijuana-relevant tweets. The NORC team enhanced the GSMA PowerTrack data to include a richer set of metrics than what is available in the GSMA SDE Twitter 1%. In addition to volume, including information on tweet source (metadata on type of account, e.g., commercial, bot, news); engagement (total number of retweets, replies, and quote tweets); geography (e.g., state level for the United States); and volume, engagement, and sentiment specific to the topic of marijuana legalization, a subset of the overall marijuana-related content contained in the GSMA PowerTrack data sets. We again derive sentiment for these data sets using VADER. The GSMA PowerTrack All Daily provides all the above information, excluding geography, for these marijuana-related tweets. The data currently include relevant tweets from August 2016 through February 2022 aggregated by day. The GSMA PowerTrack State Daily summarizes overall volume, legalization volume, and sentiment at a state level (restricted to tweets that have a US location) and covers the same time period aggregated by day.

The GSMA data is available on the GSMA webpage. For more details about the GSMA data, please refer to the GSMA Methodology Report.

[i] The Social Data Explorer is an internal NORC data tool used by the Social Data Collaboratory for analysis of social media data.

[ii] This means that English language tweets from an unspecified location may be included in the archive.

METHODS

To illustrate the potential of the GSMA, we focus our exploration on the SDE Twitter 1% and all six search terms included therein. We summarize the volume and sentiment from daily estimates to monthly estimates for this analysis.

We compare the GSMA SDE Twitter 1% data with data from the General Social Survey (GSS). The GSS is a nationally representative survey designed to study social change in the United States about every two years. The GSS has been conducted by NORC since its inception in 1972 and its design stresses replication and comparability over time. The GSS covers a wide variety of topics including politics, social life, economic life, lifestyles, science, religion, and civil liberties.

We identified seven variables that corresponded most closely with six search terms from the GSMA SDE Twitter 1%: favoring marijuana legalization, favoring abortion for any reason, favoring gay marriage, favoring gun permits, favoring the government raising taxes on the rich or giving income assistance to the poor, and two climate change measures regarding the attitudes toward spending on alternative energy sources and protecting the environment. To be consistent with the timeframe of the GSMA SDE Twitter 1% data, we use data from the 2018 and 2021 GSS. We include all valid response options, including “don’t know” or “can’t choose” responses. We account for the GSS survey design by using the survey weight and design variables.

When comparing both data sources, we focus on the positive to negative sentiment ratio. To produce a comparable sentiment measure to the GSMA SDE Twitter 1% data, we calculated the ratio of positive to negative GSS responses (e.g., favor over oppose, affirm over deny) for each of our variables. For each positive to negative ratio of sentiment, we include two versions: a clean favor over oppose ratio and a favor over other (oppose, neither, and/or don’t know) ratio. One is the threshold value for this positive to negative sentiment ratio: anything over one is considered positive, and anything under one should be considered negative, on average.

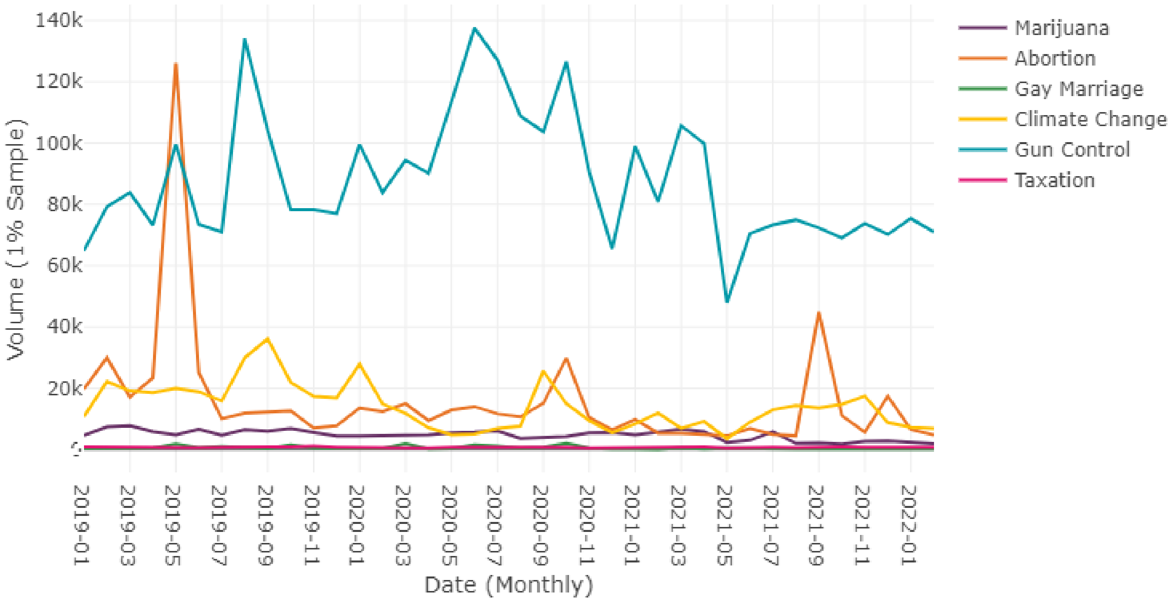
RESULTS

We first examine the relevant measures from the GSMA SDE Twitter 1% on each of the six topics. We then compare the GSMA SDE Twitter 1% and the GSS against each other focusing on positive to negative sentiment.

GSMA SDE Twitter 1%

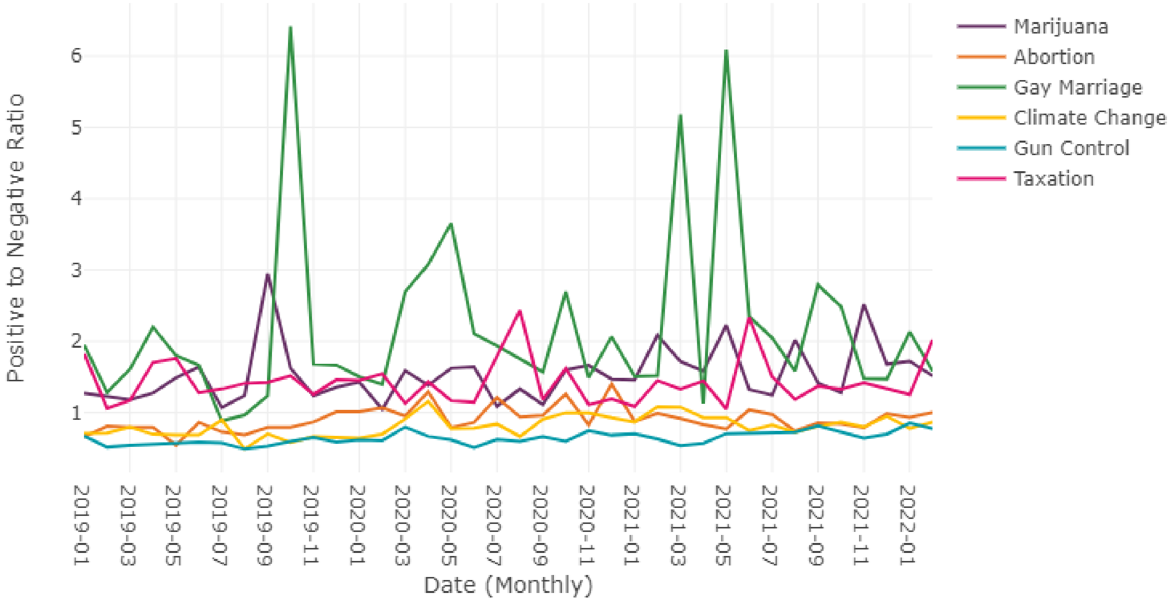
Among the six public opinion topics included in the GSMA SDE Twitter 1%, gun control stands out as the most talked about topic over the three-year period. There are multiple surges in the gun control conversation: August 2019 corresponding with the back-to-back mass shootings in El Paso, Texas and Dayton, Ohio; Summer 2020 corresponding with the tense protests surrounding the deaths of Breonna Taylor and George Floyd; and March 2021 corresponding with the mass shooting in Boulder, Colorado (see Figure 1). The topic of abortion saw multiple

Figure 1. Volume of tweets on six social topics



Source: General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022.

Figure 2. Sentiment of tweets on six social topics



Source: General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022.

spikes corresponding with the surge of abortion bans signed into law across the country. Climate change sat as the second most tweeted about topic in late 2019, though the volume of climate change tweets has diminished during the COVID-19 pandemic. We see the largest spike around September 2019 corresponding with the Global Week for Future, or the international climate strikes and protests, with a similar spike in September 2020. Marijuana, taxation, and gay marriage each individually have less than ten thousand tweets in the 1% Twitter sample, translating to less than a million tweets a month with no remarkable spikes over the three-year period.

Despite the low volume of tweets regarding gay marriage, we see that there is overwhelmingly positive sentiment in these tweets, with ratios reaching over 6.0 multiple times over the three-year period (see Figure 2). Marijuana and taxation tweets are generally positive. Tweets on gun control consistently exhibit negative sentiment while tweets on both climate change and abortion are generally negative with occasional exceptions.

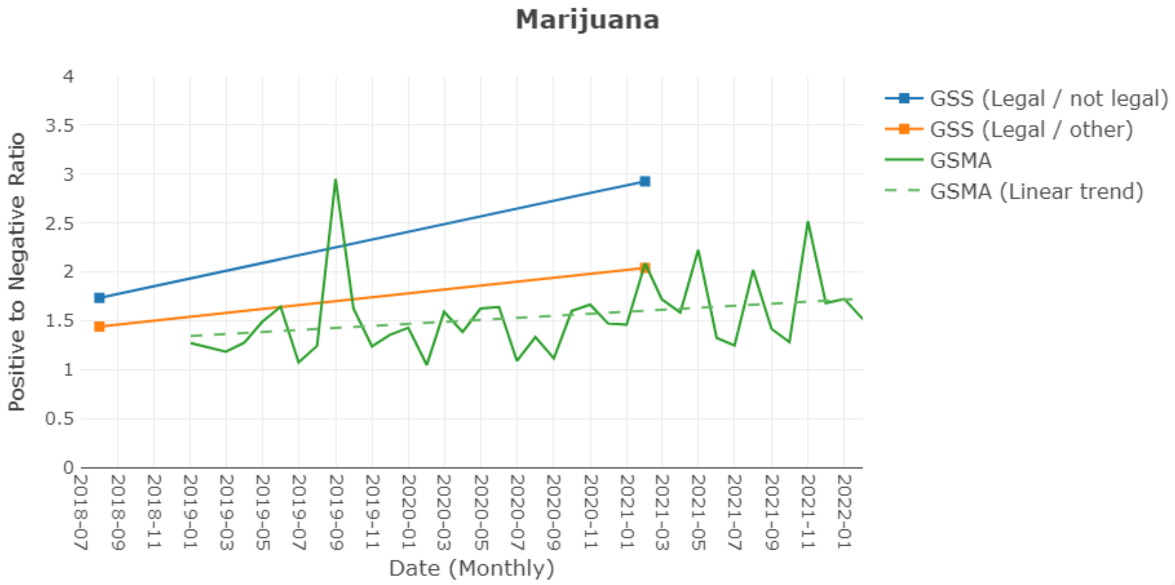
Comparison to the General Social Survey

To better show changing sentiment over time, we include a trendline for the GSMA SDE Twitter 1% data based on a simple linear regression model. We display GSS data in the following graphs at the midpoint of data collection for each year: August 2018 for 2018 GSS and February 2021 for 2021 GSS.

Beginning with sentiment towards marijuana, we saw an increasingly positive sentiment from both the GSMA SDE Twitter 1% and the GSS (see Figure 3). The legal / not legal ratio from the GSS data suggested a steeper rise in positive sentiment over this time period compared to the legal / other ratio, with the latter more consistent with the trendline from the Twitter data.

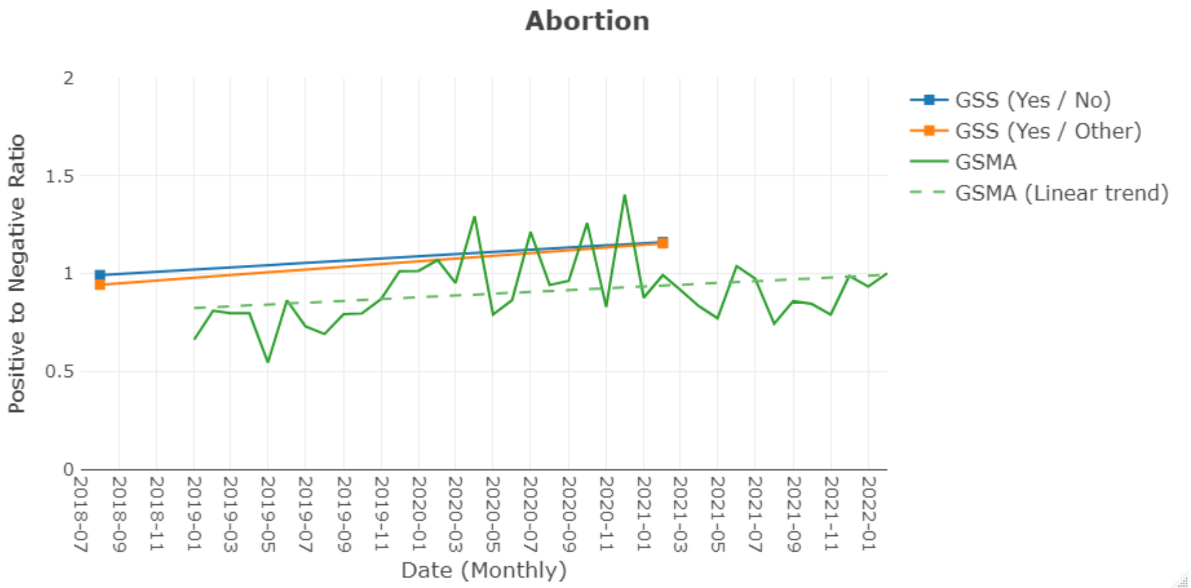
Similarly, when examining the sentiment towards abortion, we observed a rising positive sentiment (see Figure 4). The positive to negative ratio on abortion from the GSS data crossed from negative to positive sentiment between 2018 and 2021. The sentiment on Twitter over the 3-year window was generally negative, but the trendline puts it transitioning to a positive sentiment after the examined timeframe.

Figure 3. Comparison of positive to negative sentiment on marijuana



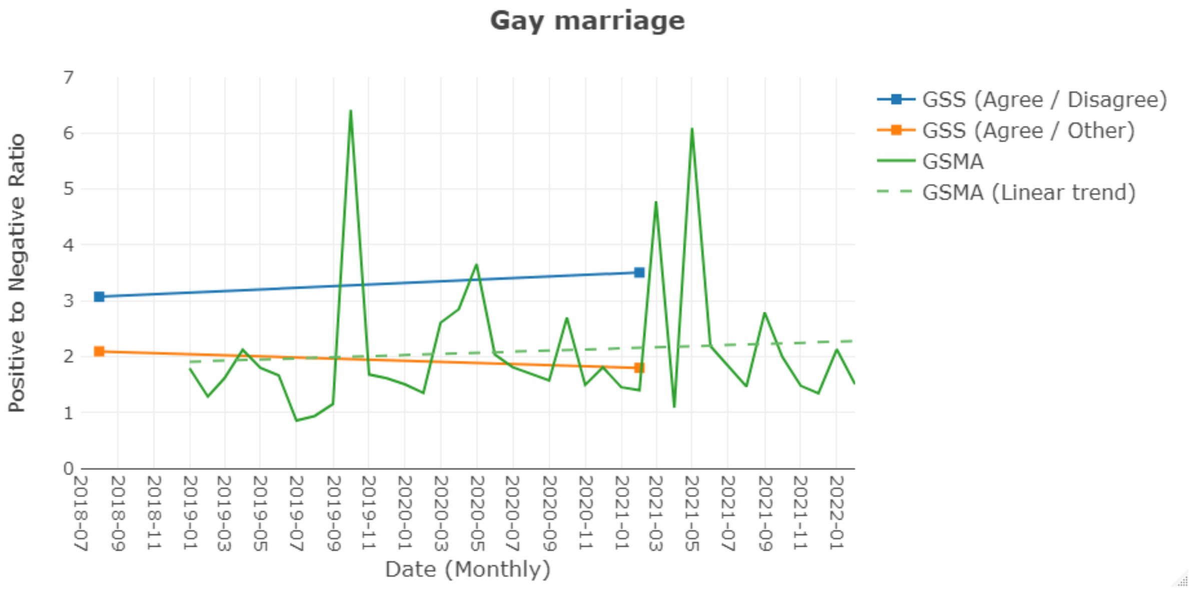
Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS question examines marijuana legalization while the GSMA SDE search term captures marijuana generally.

Figure 4. Comparison of positive to negative sentiment on abortion



Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS question examines abortion for any reason while the GSMA SDE search term captures abortion generally.

Figure 5. Comparison of positive to negative sentiment on gay marriage



Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS question examines agreement that same-sex couples have the right to marry one another while the GSMA SDE search term captures gay marriage generally. For GSS, “Agree” combines 1 “Strongly agree” and 2 “Agree” and “Disagree” combines 4 “Disagree” and 5 “Strongly disagree.”

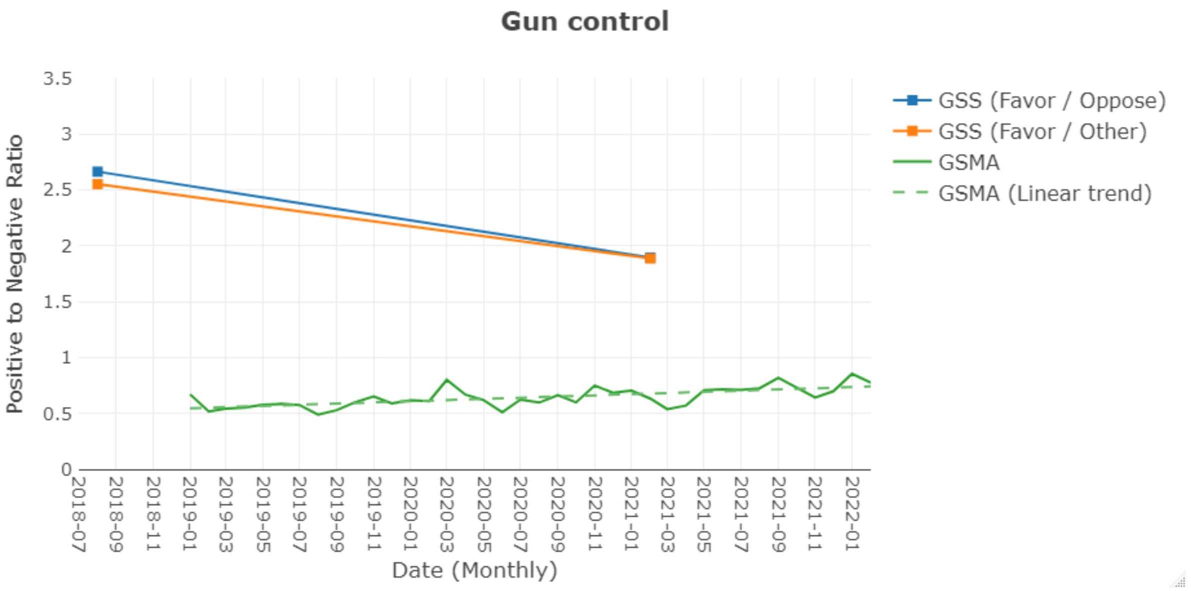
Sentiment towards gay marriage was overall positive in both the GSMA SDE Twitter 1% and the GSS (see Figure 5). The GSS sentiment ratio excluding neutral categories was nearly parallel with the SDE ratio, though the GSS ratio was much larger. The alternative GSS ratio including the neutral categories was more similar to the SDE ratio, despite a declining positive to negative ratio.

Gun control provided us the first major disconnect between the SDE and GSS data (see Figure 6). While favor for gun control was overwhelmingly positive (though declining) in GSS, the GSMA SDE Twitter 1% data showed overwhelmingly negative sentiment, though there was an upward trend across the three-year period. This stark contrast may suggest that negative sentiment in tweets regarding gun control is not specifically referring to negative sentiment toward gun control laws, but, rather, a lack of gun control.

The sentiment towards taxation was stagnant in the GSMA SDE data over the three-year period whereas the sentiment towards government interventions to decrease the income disparities had growing positive sentiment in the GSS data (see Figure 7). The overly specific, and double-barreled, question in GSS may not pair best with the generic nature of the Twitter search term.

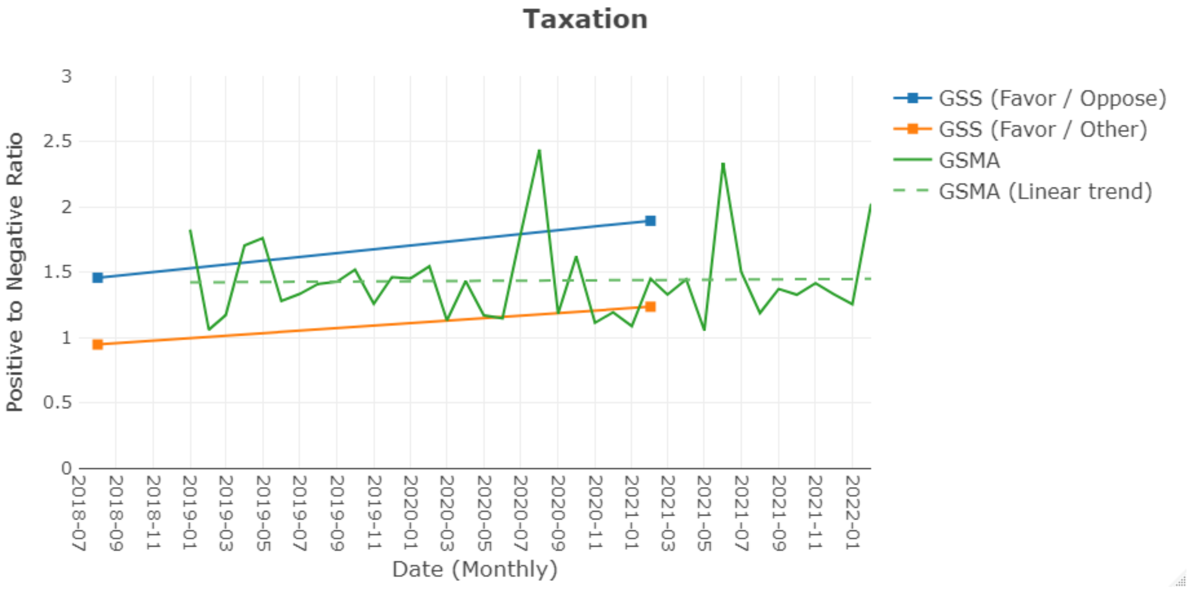
Both the GMSA SDE Twitter 1% and the GSS data denoted negative sentiment towards climate change (see Figure 8). The negative sentiment of climate change tweets was decreasing over time while the negative sentiment was staying steady or increasing, depending on the measure of climate change, based on the GSS data. Our results here echo potential effects we noted for both gun control and taxation. People may be expressing their negative sentiment towards climate change and its general impacts and not to any specific policies or approaches. Thus, this generic search term paired with two specific questions may be more of a non-equivalent comparison.

Figure 6. Comparison of positive to negative sentiment on gun control



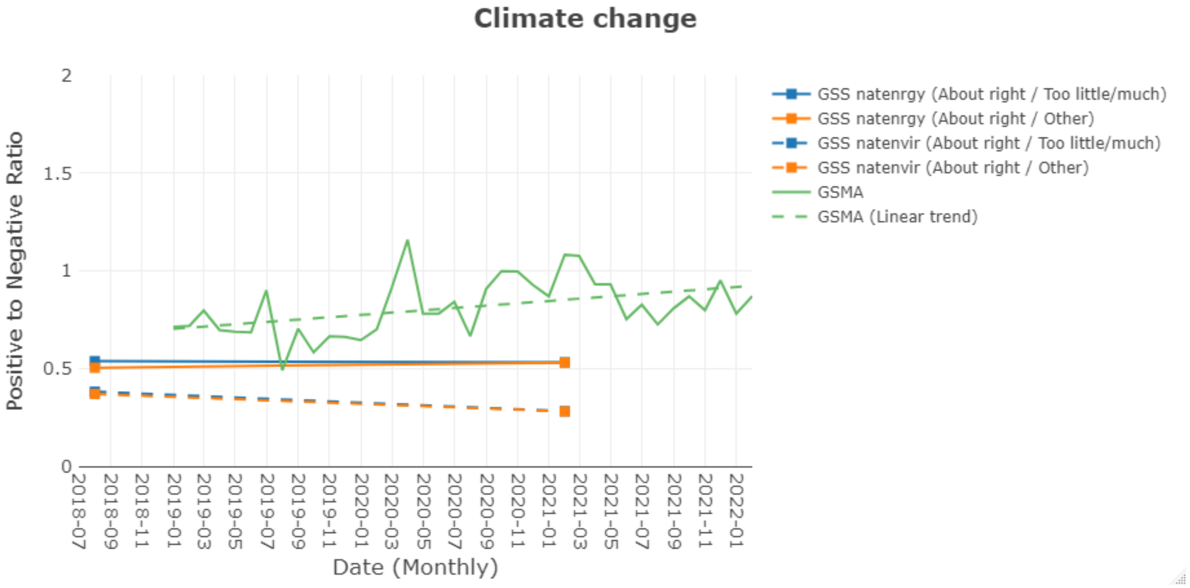
Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS question examines favoring of gun control while the GSMA SDE search term captures gun control generally.

Figure 7. Comparison of positive to negative sentiment on taxation



Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS question examines whether respondents favor raises taxes on the rich or giving income assistance to the poor while the GSMA SDE search term captures taxation generally.

Figure 8. Comparison of positive to negative sentiment on climate change



Source: General Social Survey, 2018-2021, and General Social Media Archive Social Data Explorer Twitter 1%, Jan. 2019 – Feb. 2022. Note. The GSS questions examine attitudes toward spending on developing alternative energy resources (natenrgy) and protecting the environment (natenvir), respectively, while the GSMA SDE search term captures climate change generally.

DISCUSSION

The General Social Media Archive is one of the first public-use data sets of social media information available to researchers. The inclusion of volume and sentiment in the GSMA SDE Twitter 1% data allows for a variety of different explorations. This report focused on comparing sentiment between social media data from the GSMA SDE Twitter 1% and survey data from the General Social Survey. Our initial analyses found that sentiment towards marijuana, abortion, and gay marriage were generally consistent between the survey and social media data, though the level of positive to negative sentiment differed to a degree. Our results on gun control, taxation, and climate change did not align as well with GSS data. Our chosen Twitter search terms and their associated GSS survey questions may not have strongly correlated, because the survey questions are too specific to the broad, multidimensional topic areas searched on Twitter.

Our initial comparison of the GSMA SDE and the GSS data is not without limitations. With only two years of GSS data available for the corresponding GSMA SDE timeframe, it is difficult to measure comparable trends over time. This analysis could be expanded following the release of the 2022 GSS, expected in the Spring of 2023. Comparing GSMA with other data sources with more frequent collection and reporting may also result in more meaningful comparisons.

In addition, the use of positive to negative sentiment for GSS in particular was exploratory. Alternate groupings of GSS responses for gay marriage, taxation, and climate change may also have impacted findings. The decision to include neutral categories and “don’t know” responses did create some differences for select variables. Regardless of the way we calculated the sentiment ratio for GSS, the positive to negative ratio was often larger than those from the GSMA SDE (i.e., more positive when both were positive, more negative when both were negative).

Sentiment, as measured by VADER, leaves room for improvement when comparing to a survey like GSS as it fails to consider some contexts of a tweet. While a survey question should have a clear attitudinal direction (e.g., “yes” means one thing, “no” means the opposite), social media posts do not follow such a clear pattern. For example, consider the following two statements: “I’m a proud to have the right to bear arms. Say no to gun control.” and “I’m happy to see my state vote to increase gun control restrictions on semi-automatics.” VADER would rate both statements overall as having “positive” sentiment because language surrounding a mention of the term “gun control” is positive, despite being on opposite sides of the issue of gun control (against and for, respectively). This may explain, for example, a portion of the strong negative sentiment to gun control if people are expressing disappointment in the lack of gun control. This lack of alignment to the social issues’ differing points of view may have resulted in large misclassifications resulting in overestimating or deflating positive to negative ratios. The detection of overall tweet sentiment is a helpful contribution, but additional work is needed to properly align sentiment with social context.

Finally, these data are even more critical now as Twitter experiences dramatic structural changes in leadership and philosophy. Following the finalization of Twitter’s acquisition in late October 2022, Twitter’s future is uncertain.(9) Advertisers and users have been leaving the platform en masse in the wake of the purchase, many over concerns surrounding the company’s changes in policy and leadership.[10][11] Future releases of GSMA data that will include Twitter data pre and post-acquisition will be a valuable asset for understanding the changing media environment.

With historically two hundred billion tweets coming from Twitter per year,[12] the GSMA SDE Twitter 1% data provides a large (averaging nearly two billion tweets per year), but workable cross-section of social media posts for the GSMA. We hope to release more social media data through the archive on a variety of new topics in the future to allow for greater utility of the GSMA SDE Twitter 1% data and expand the types of social topics researchers can explore. In addition, the GSMA includes two more data sets using the richer Twitter PowerTrack API data which are not explored in this report and provide a host of additional information helpful in understanding and breaking down specific social topics, including by tweet source and geography. NORC is releasing these data to the public in the hopes that other research teams will be inspired to develop their own research into how opinions are changing in the US, contextualizing and complementing various public opinion surveys with the openly expressed opinions of those on social media.

REFERENCES

(1) Link, Michael. “AAPOR 2025 and the opportunities in the decade before us.” Presidential address given at the 2015 American Association for Public Opinion Research (AAPOR) Annual Conference, Hollywood, FL, May 15, 2015. https://www.aapor.org/About-Us/History/Presidential-Addresses/2015-Presidential-Address.aspx.

(2) Murphy, Joe, Michael W. Link, Jennifer H. Childs, Casey L. Tesfaye, Elizabeth Dean, Michael Stern, Josh Pasek, Jon Cohen, Mario Callegaro, and Paul Harwood. “Social Media in Public Opinion Research: Executive Summary of the AAPOR Task Force on Emerging Technologies in Public Opinion Research.” *Public Opinion Quarterly* 78, no. 4 (2014): 788-794. https://doi.org/10.1093/poq/nfu053

(3) Czaplicki, Lauren, Ganna Kostygina, Yoonsang Kim, Siobhan N. Perks, Glen Szczypka, Sherry L. Emery, Donna Vallone, and Elizabeth C. Hair. "Characterising JUUL-related posts on Instagram." *Tobacco Control* 29, no. 6 (2020): 612-617.

(4) Colditz, Jason B., Kar-Hai Chu, Sherry L. Emery, Chandler R. Larkin, A. Everette James, Joel Welling, and Brian A. Primack. "Toward real-time infoveillance of Twitter health messages." *American journal of public health* 108, no. 8 (2018): 1009-1014.

(5) Kostygina, Ganna, Hy Tran, Barbara Schillo, Nathan A. Silver, and Sherry L. Emery. "Industry response to strengthened regulations: amount and themes of flavoured electronic cigarette promotion by product vendors and manufacturers on Instagram." *Tobacco Control* 31, no. Suppl 3 (2022): s249-s254.

(6) Twitter. “Volume streams.” Accessed December 6, 2022, https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction

(7) Hutto, C. and Eric Gilbert. 2014. “VADER: A parsimonious rule-based model for sentiment analysis of social media text.” *Proceedings of the International AAAI*

Conference on Web and Social Media 8 (1):216-25. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

(8) Twitter. “Historical PowerTrack API.” Accessed December 6, 2022, https://developer.twitter.com/en/docs/twitter-api/enterprise/historical-powertrack-api/overview

(9) Conger, Kate, and Lauren Hirsch. “Elon Musk Completes $44 Billion Deal to Own Twitter.” *The New York Times,* October 27, 2022. https://www.nytimes.com/2022/10/27/technology/elon-musk-twitter-deal-complete.html

(10) Hutton, Christopher. “Timeline: Twitter’s wild first two weeks under Elon Musk.” *Washington Examiner,* November 14, 2022. https://www.washingtonexaminer.com/policy/technology/timeline-twitters-wild-first-two-weeks-under-elon-musk

(11) Winder, Davey. “Twitter Users Warned Not to Delete Their Accounts – Here’s Why.” *Forbes,* November 27, 2022. https://www.forbes.com/sites/daveywinder/2022/11/27/twitter-users-warned-not-to-delete-their-accounts-heres-why/?sh=607eb07c70f5

(12) Ruby, Daniel. *“*Twitter Statistics: Facts and Figures After Elon Musk Takeover (2022).” Demand Sage. Accessed December 6, 2022. https://www.demandsage.com/twitter-statistics/

APPENDIX

General Social Survey variables

Topic	Variable	Question	Responses	Positive to negative ratio	Analysis notes
Abortion	abany	Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason?	1 Yes 2 No .d Don't know	(1) Yes / No (2) Yes / (No + Don't know)	
Climate change	natenrgy	Are we spending too much, too little, or about the right amount on developing alternative energy sources?	1 Too little 2 About right 3 Too much .d Don't know	(1) About right / (Too little + Too much) (2) About right / (Too little + Too much + Don't know)	
Climate change	natenvir	Are we spending too much, too little, or about the right amount on improving and protecting the environment?	1 Too little 2 About right 3 Too much .d Don't know	(1) About right / (Too little + Too much) (2) About right / (Too little + Too much + Don't know)	
Gay marriage	marhomo	Do you agree or disagree? Homosexual couples should have the right to marry one another.	1 Strongly agree 2 Agree 3 Neither agree or disagree 4 Disagree 5 Strongly disagree .d Don't know	(1) (Strongly agree + Agree) / (Disagree + Strongly disagree) (2) (Strongly agree + Agree) / (Neither agree or disagree + Disagree + Strongly disagree + Don't know)	GSS has asked two slight variations of this question over the years with "marhomo" (1988, 2004, 2021) stating "Do you agree or disagree? Homosexual couples *should* have the right to marry one another." (emphasis added to denote the subtle difference) and "marhomo1" (2006-2018) stating "Do you agree or disagree with the following statement? Homosexual couples have the right to

					marry one another.” For purposes of our analysis, we use the corresponding variable based on the year it was available: “marhomo1” for 2018 and “marhomo” for 2021.
Gun control	gunlaw	Would you favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun?	1 Favor 2 Oppose .d Don't know	(1) Favor / Oppose (2) Favor / (Oppose + Don't know)	
Marijuana	grass	Do you think the use of marijuana should be made legal or not?	1 Should be legal 2 Should not be legal .d Don’t know	(1) Should be legal / Should not be legal (2) Should be legal / (Should not be legal + Don’t know)	With the transition to web data collection in 2021, the variable “grass” does not currently exist. GSS conducted a question design experiment either providing the “Don’t know” response option explicitly to respondents (“grassv”) or requiring respondents to volunteer such a response (“grassnv”). For purposes of our analysis, we use a comparable version of “grass” for 2021 by combining “grassv” and “grassnv,” resulting in a similar distribution for “Don’t know” to previous years.

Taxation	eqwlth	Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think that the government should not concern itself with reducing this income difference between the rich and the poor. Think of a score of 1 as meaning that the government ought to reduce the income differences between rich and poor, and a score of 7 meaning that the government should not concern itself with reducing income differences. What score between 1 and 7 comes closest to the way you feel?	1 The government should reduce income differences 2 3 4 5 6 7 The government should not concern itself with reducing income differences	(1) (1+2+3) / (5+6+7) (2) (1+2+3) / (4+5+6+7 + Don't know)	