The Utility of Big Data for Evaluating Public Opinion

Michael Robbins
RAND Corporation

Collaborators:
Max Griswold, Michael Pollard

August 6, 2024

- Social media data sources like Twitter (now X) provide a wealth of information that could be used to evaluate public opinion in real time.
  - Large language models (e.g., ChatGPT) have proven capable of quantifying sentiment of nuanced text

- Users of ~~Twitter~~ (in particular the most vocal ones) are not representative of the general population

- Is it possible to apply weights to ~~Twitter~~ users to correct for non-representativeness?

~~Twitter~~ Survey

We used a targeted ad campaign to collect survey data from a sample of ~~Twitter~~ users.

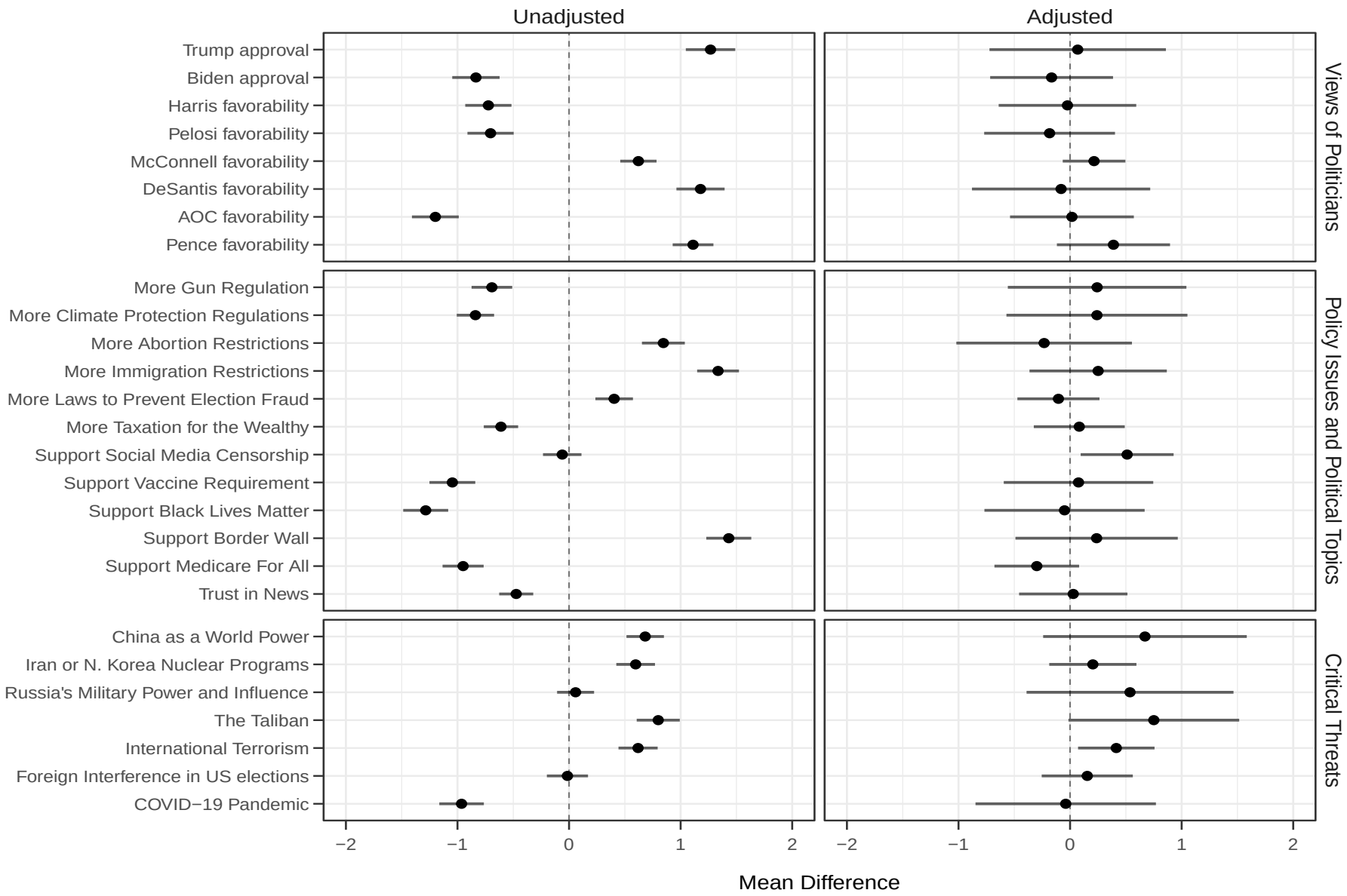
- Sample frame includes users who were determined to be in the US based upon location in user profile
- Respondents are among the more vocal users
- Collected demographics and a wide range of political opinions
- Data collected from March to July 2022
  - Elon Musk took over ~~Twitter~~ in October 2022
  - The demographic composition of ~~Twitter~~ may have changed since.

~~Twitter~~ Sample Characteristics

	Probability Sample	~~Twitter~~ Sample
Sample Size	1,972	822
Age	48	37.7
18-29	19.9%	35.6%
30-39	17.2%	27.1%
40-49	16.0%	14.9%
50-59	16.5%	10.4%
60+	30.4%	11.9%
Gender (Male)	46.1%	59.9%
Race (White)	62.2%	55.0%
Income	$73,000	$60,000
College degree	35.1%	47.9%
Voted for Biden	34.9%	64.3%

- Application of propensity score weighting methods for blending probability and convenience samples (e.g., Robbins *et al.*, 2021, *JSSM*) shows that it is possible to correct for bias in the ~~Twitter~~ sample.
  - This assumes that demographics and related information is observed for the users.
  - Variables used include: Age, gender, race, income, *political ideology*, vaccination status, education, time spent using technology.

Blending ~~Twitter~~ and probability samples



- Application of propensity score weighting methods for blending probability and convenience samples (e.g., Robbins *et al.*, 2021, *JSSM*) shows that it is possible to correct for bias in the ~~Twitter~~ sample.
  - This assumes that demographics and related information is observed for the users.
  - Variables used include: Age, gender, race, income, *political ideology*, vaccination status, education, time spent using technology.
- Such characteristics are not known for ~~Twitter~~ users.
- We do have:
  - User profile
  - Tweets
- Can we use this information to develop proxies for the demographic characteristics of interest?

~~Twitter~~ Sample

Data collected on ~~Twitter~~ survey respondents:

- User profile:
  - Name, description, location
  - Accounts followed (1514 per user)
- Tweets:
  - Collected all tweets and retweets posted by the survey respondents between 9/20/2020 and 7/20/2021
  - 1.44 million total tweets and 1.46 million total retweets
    - 1768 tweets and 1816 retweets per user

Demographics can sometimes be inferred from the user profile.

What information can we glean from the tweets themselves?

Big data

For each ~~Twitter~~ respondent, tabulate:

- **Words**: The frequency with which a respondent uses each of 57,563 words that appear in the tweets.
- **Emojis**: The frequency with which a respondent uses each of 4,733 possible emojis.
- **Followings**: Does the respondent following a specific account? (Out of 11,769 accounts that are followed by at least 10 of the users.)
- **Re-tweets**: How often did the respondent retweet a specific account? (Out of 30,576 accounts.)
- **Tags**: How often did the respondent tag a specific account? (Out of 20,040 accounts.)
- **Hastags**: The frequency with which a respondent uses each of 3,230 hastags that appear in the tweets
- **Readability**: The # of sentences, syllables, misspellings, etc.

Words

Table: Correlation with age and use of specific words.

Positive (Older)			Negative (Younger)		
Word	Corr.	Count	Word	Corr.	Count
administration	0.442	247	rn	-0.314	231
pence	0.435	207	idk	-0.308	276
constitution	0.428	194	lmao	-0.307	256
insurrection	0.427	165	bruh	-0.306	177
gop	0.426	207	fr	-0.285	131
sedition	0.422	100	lmaooo	-0.271	73
oath	0.422	103	yall	-0.267	147
giuliani	0.418	96	pls	-0.264	152
fauci	0.416	142	bro	-0.258	366
impeachment	0.415	184	tf	-0.257	197
voters	0.411	254	tho	-0.255	295
mcconnell	0.411	191	tryna	-0.253	93

Emojis

Your choice of emojis can literally convey who you are.

- For example, gender- and/or race-specific versions of emojis exist.
  - 
  - 
  - 
  - 

Emojis

Table: Correlation with gender and use of specific emojis.

Positive (Female)			Negative (Male)		
Emoji	Corr.	Count	Emoji	Corr.	Count
💜	0.341	136	🤦‍♂️	-0.230	91
💕	0.294	132	🤦	-0.216	58
🥰	0.280	239	🤦🏻‍♂️	-0.139	23
💔	0.273	158	🤦🏻	-0.106	15
🎂	0.258	123	🤦🏾‍♂️	-0.094	16
🤦‍♀️	0.245	56	🤦🏽	-0.093	13
😘	0.244	158	🤦🏾	-0.093	11
🤗	0.244	132	📲	-0.084	13
💙	0.240	174	🐐	-0.078	84
😍	0.236	302	💪	-0.068	10
😊	0.231	229	🤦🏽‍♂️	-0.068	18
🤦	0.229	44	📈	-0.067	33

Accounts followed

Table: Correlation with party ID and specific accounts followed.

Positive (Democrat)			Negative (Republican)		
Account	Corr.	Count	Account	Corr.	Count
barackobama	0.415	328	charliekirk11	-0.396	32
kamalaharris	0.413	222	realjameswoods	-0.391	36
joebiden	0.401	274	realcandaceo	-0.380	35
vp	0.385	207	thebabylonbee	-0.370	43
potus	0.381	247	libsoftiktok	-0.364	34
danrather	0.365	125	dbongino	-0.362	30
hillaryclinton	0.361	162	donaldjtrumpjr	-0.353	51
whitehouse	0.354	176	tuckercarlson	-0.349	46
ewarren	0.352	145	mrandyngo	-0.348	28
maddow	0.342	146	catturd2	-0.346	27
senwarren	0.333	131	ryanafournier	-0.345	29
yamiche	0.331	101	jackposobiec	-0.342	27

Prediction

The greater goal here is to build a model for relevant demographic characteristics based on user profile and tweets

- In all, we have 124,789 predictors.
- Use lasso ($\alpha = 1$), elastic net ($0 < \alpha < 1$), or ridge regression ($\alpha = 0$) with k-folds cross validation to model demographic characteristics on these predictors.

Variable	Best α	Best λ	MSE	# Pred.	R^2
Education	0.05	1.002	2.182	1052	0.813
Gender	0.05	0.109	0.854	2987	0.962
Age	0.10	6.552	91.326	610	0.786
Race (white)	0.05	0.176	0.924	2434	0.912
Race (Black)	0.50	0.060	0.577	160	0.633
Income	0.00	65.316	16.979	109302	0.948
Religion (Christian)	0.05	0.259	1.211	2302	0.806
Party ID	0.25	0.463	1.748	179	0.595

Next Steps

Can pretrained large language models and transformer methods be fine-tuned using survey responses to produce better proxies for demographics?

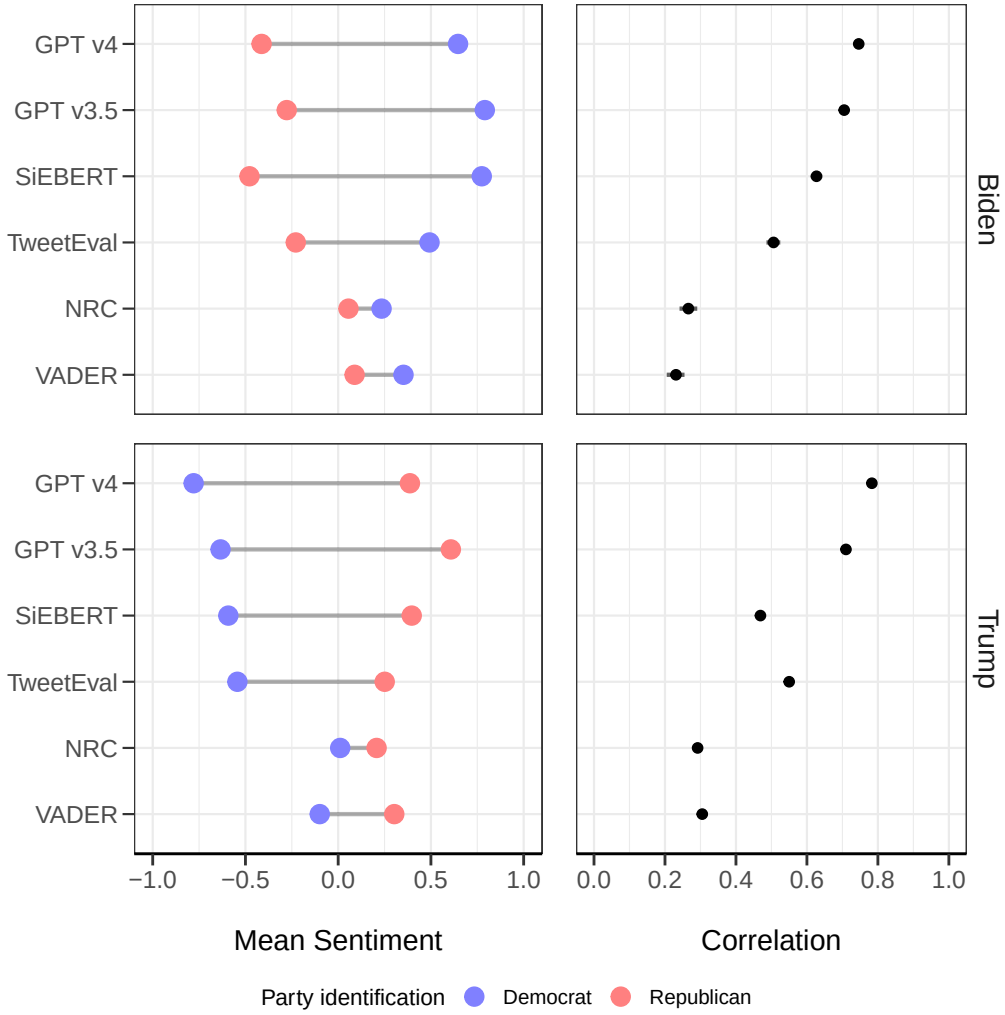
- BERT
- RoBERTa
- SieBERT
- TweetEval
- GPT

GPT has proven very successful at quanitfying sentiment of tweets.

- "Provide a value between [-1, 1] which determines the degree to which the author of the following text likes or dislikes {person}..."

Next Steps

Applied LLMs to 20,000+ tweets from politicians about Trump or Biden:



Next Steps

Compile a universe of $\sim 40,000$ highly vocal ~~Twitter~~ users.

- Outcome: Sentiment expressed towards Trump and/or Biden

Weighting:

- Proxies for demographic characteristics are not substituted for the characteristics themselves in a weighting scheme. Instead...
- Treating the ~~Twitter~~ survey as a convenience sample, blend with the probability sample using known methods.
- Find benchmarks for the proxy variables using weights from this blending.
- Calibrate the ~~Twitter~~ universe to match these benchmarks.
- Analogous propensity score-like procedures can be developed.

Next Steps

Differences between public opinion polls and ~~Twitter~~ sentiment scores are obvious.

- TBD: Can weighting correct for such biases?