Less is more?
How demographic sample weights can improve public opinion estimates based on Twitter data.

Pablo Barberá
School of International Relations
University of Southern California
`@p_barbera`

CELLER REPORTS SMITH STRONG IN MIDWEST

Representative Tells of Train Polls on His Trip Back From Houston.

Representative Emanuel Celler gave optimistic reports on the sentiment for Governor Smith in the Middle West, on his return to the city yesterday from the Democratic Convention at Houston.

"I passed through the States of Colorado, Nebraska and Iowa," said Mr. Celler. "I took straw votes on the observation trains going from Houston to Denver and from Denver east, and in these polls found a healthy and growing sentiment for Smith.

"For example, on the train from Denver to Omaha there were fourteen persons on the observation car. They came from the States of Nebraska, Illinois, Iowa and Colorado. One man was from Pennsylvania, and I was the only New Yorker. The votes stood 1 blank, 3 for Hoover and 10 for Smith. I did not vote.

"The poll on the train going from Chicago to New York was two to one in favor of Smith. From my observation in Denver and other sections in Colorado I am of the opinion that Smith will have more than an even break there and will carry the State of Colorado, including Denver, Pueblo, Boulder, &c. This vote will be more than ample to carry the State.

New York Times, July 8, 1928

CELLER REPOR
STRONG IN

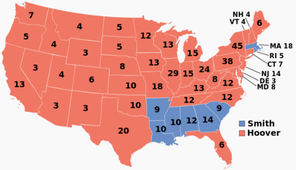
Representative Tel
Polls on His Trip
Houston

Representative Er
gave optimistic repor
ment for Governor Sn
dle West, on his retu
yesterday from the D
vention at Houston.
"I passed through
Colorado, Nebraska a
Mr. Celler. "I took
the observation trai
Houston to Denver an
east, and in these
a healthy and growin
Smith.

United States presidential election, 1928

1924 ←	November 6, 1928	→ 1932
	531 electoral votes of the Electoral College	
	266 electoral votes needed to win	


Nominee	**Herbert Hoover**	Al Smith
Party	Republican	Democratic
Home state	California	New York
Running mate	**Charles Curtis**	Joseph T. Robinson
Electoral vote	444	87
States carried	40	8
Popular vote	21,427,123	15,015,464
Percentage	58.2%	40.8%



le, on the train from
naha there were four-
on the observation car.
rom the States of Ne-
s, Iowa and Colorado.
s from Pennsylvania,
only New Yorker. The
blank, 3 for Hoover and
I did not vote.
n the train going from
w York was two to one
nith. From my obser-
ver and other sections
am of the opinion that
ve more than an even
nd will carry the State
including Denver,
er, &c. This vote will
n ample to carry the

imes, July 8, 1928

The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

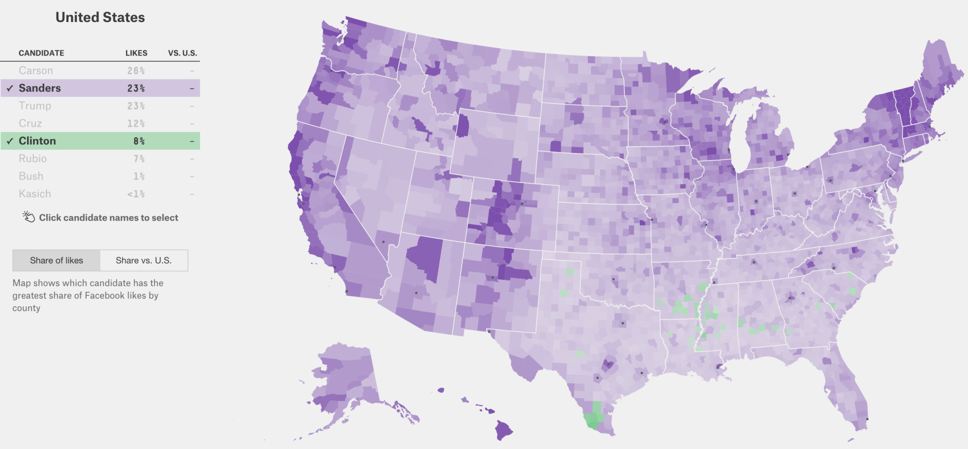
Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens

Facebook likes to presidential candidates, by county



Source: FiveThirtyEight and Facebook

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

"*The predictive power of Twitter regarding elections has been vastly overstated* " (Gayo-Avello, 2012)

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

"*The predictive power of Twitter regarding elections has been vastly overstated*" (Gayo-Avello, 2012)

Three main challenges:

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

"*The predictive power of Twitter regarding elections has been vastly overstated*" (Gayo-Avello, 2012)

Three main challenges:

- **Sampling bias**: not all sociodemographic groups are equally active on Twitter

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

"*The predictive power of Twitter regarding elections has been vastly overstated* " (Gayo-Avello, 2012)

Three main challenges:

- **Sampling bias**: not all sociodemographic groups are equally active on Twitter
- **Non-response bias**: some groups are more likely to post about politics

Measuring public opinion with Twitter data?

Many highly-cited studies claim metrics based on social media data can help us predict electoral results (Tumasjan et al, 2010; O'Connor et al, 2010; DiGrazia et al, 2013; Wang et al, 2012).

"*The predictive power of Twitter regarding elections has been vastly overstated* " (Gayo-Avello, 2012)

Three main challenges:

- **Sampling bias**: not all sociodemographic groups are equally active on Twitter
- **Non-response bias**: some groups are more likely to post about politics
- **Unprompted responses**: difficult to interpret and categorize

Measuring public opinion with Twitter data?

Importing survey research methods:

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)
- **Non-response bias**:

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)
- **Non-response bias**:
  - Tracking polls with panels of respondents

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)
- **Non-response bias**:
  - Tracking polls with panels of respondents

Problems in application of these methods to Twitter:

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)
- **Non-response bias**:
  - Tracking polls with panels of respondents

Problems in application of these methods to Twitter:

- *Privacy paradox* (Golder and Macy, 2014): lack of key individual-level information about Twitter users

Measuring public opinion with Twitter data?

Importing survey research methods:

- **Sampling bias**:
  - Probability sampling and post-stratification using weights
  - With proper adjustment, probability sampling not required (Wang et al, 2014)
- **Non-response bias**:
  - Tracking polls with panels of respondents

Problems in application of these methods to Twitter:

- *Privacy paradox* (Golder and Macy, 2014): lack of key individual-level information about Twitter users
- Sampling at tweet level, not user level

This project

- Method to estimate sociodemographic traits:

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.
- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.
- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) Stable sample alleviates concerns about non-response bias (and spam, bots)

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) Stable sample alleviates concerns about non-response bias (and spam, bots)
  c) 3 applications:

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) Stable sample alleviates concerns about non-response bias (and spam, bots)
  c) 3 applications:
     1. Measurement of issue salience across groups

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) Stable sample alleviates concerns about non-response bias (and spam, bots)
  c) 3 applications:
     1. Measurement of issue salience across groups
     2. Early indicator of changes in candidate approval, using post-stratification to recover representativeness

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) Stable sample alleviates concerns about non-response bias (and spam, bots)
  c) 3 applications:
     1. Measurement of issue salience across groups
     2. Early indicator of changes in candidate approval, using post-stratification to recover representativeness
     3. Who spreads misinformation on Twitter?

Estimating sociodemographic traits of Twitter users

Supervised machine learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known

Estimating sociodemographic traits of Twitter users

Supervised machine learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class

Estimating sociodemographic traits of Twitter users

Supervised machine learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Estimating sociodemographic traits of Twitter users

Supervised machine learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Step 1: training dataset

Geographic location for Twitter users:

Step 1: training dataset

Geographic location for Twitter users:

- 1.2 billion geolocated tweets ($\sim$8TB) from July 2013 to June 2014 $\rightarrow$ 250M in the U.S. (4.4M unique users)

Step 1: training dataset

Geographic location for Twitter users:

- 1.2 billion geolocated tweets ($\sim$8TB) from July 2013 to June 2014 $\rightarrow$ 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

Step 1: training dataset

Geographic location for Twitter users:

- 1.2 billion geolocated tweets (∼8TB) from July 2013 to June 2014 → 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

Voting registration records:

FIRST	LAST	VOTERID	COUNTY	PARTY	2012	GENDER ...	
angela	myers	610901468	franklin	REP	X	F	...
ryan	petrik	610901998	franklin	DEM	X	M	...
...							

	RESIDENTIAL ADDRESS	ZIP	RACE	...
...	123 Main St, Columbus Oh	08001	W	...
...	77 Canal St, Columbus Oh	08009	W	...

Step 1: training dataset

Geographic location for Twitter users:

- 1.2 billion geolocated tweets ($\sim$8TB) from July 2013 to June 2014 $\rightarrow$ 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

Voting registration records:

FIRST	LAST	VOTERID	COUNTY	PARTY	2012	GENDER ...	
angela	myers	610901468	franklin	REP	X	F	...
ryan	petrik	610901998	franklin	DEM	X	M	...
...							

	RESIDENTIAL ADDRESS	ZIP	RACE	...
...	123 Main St, Columbus Oh	08001	W	...
...	77 Canal St, Columbus Oh	08009	W	...

Matching process:

Step 1: training dataset

Geographic location for Twitter users:

- 1.2 billion geolocated tweets ($\sim$8TB) from July 2013 to June 2014 $\rightarrow$ 250M in the U.S. (4.4M unique users)
- Use shape files to identify county and zipcode in U.S.

Voting registration records:

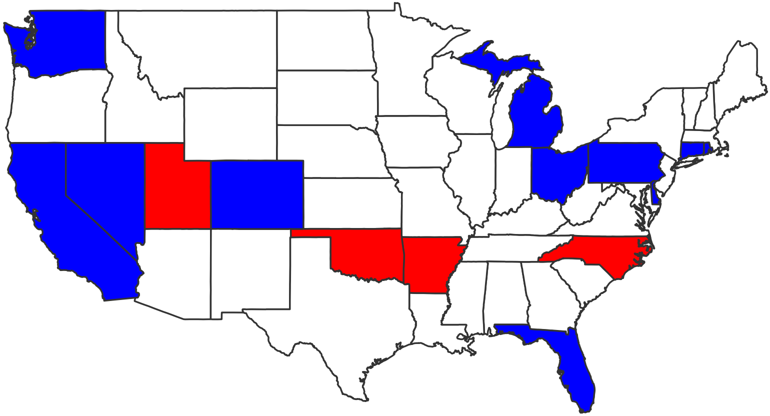
FIRST	LAST	VOTERID	COUNTY	PARTY	2012	GENDER ...	
angela	myers	610901468	franklin	REP	X	F	...
ryan	petrik	610901998	franklin	DEM	X	M	...
...							

	RESIDENTIAL ADDRESS	ZIP	RACE	...
...	123 Main St, Columbus Oh	08001	W	...
...	77 Canal St, Columbus Oh	08009	W	...

Matching process:

- Perfect, unique matches of first/last name at county level
- If duplicated, match at zipcode level.

Matching Twitter Accounts with Offline Voting Records



Python code: `github.com/pablobarbera/voter-files`

15 states, 77M registered voters (35-50% of U.S. total)

Matched Twitter accounts: 250,000 (12.3% match rate)

Estimating sociodemographic traits of Twitter users

Supervised learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Estimating sociodemographic traits of Twitter users

Supervised learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

- Words used by 1+% of users in description (K=25,500)

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

- Words used by 1+% of users in description (K=25,500)

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

- Words used by 1+% of users in description (K=25,500)
- Words used by 1+% of users in tweets (K=34,092)

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

- Words used by 1+% of users in description (K=25,500)
- Words used by 1+% of users in tweets (K=34,092)
- Emoji characters used by 1+% of users (K=627):

  🙄 😅 🎓 🙏 🏀 ⚡ 🐘 🔥

Step 2: machine learning classification

Feature selection:

Networks

- Set of *verified* accounts followed by users
- With 10K+ followers, `lang` is `en` or `es`
- K=61,659 accounts

Text

- Words used by 1+% of users in description (K=25,500)
- Words used by 1+% of users in tweets (K=34,092)
- Emoji characters used by 1+% of users (K=627):

🙄 😅 🎓 🙏 🏀 ⚡ 🐘 🔥

Classifier: gradient boosting (ensemble of decision trees) in `XGBoost` (Chen & Guestrin, 2016). Optimized with 5-fold cross-validation.

Estimating sociodemographic traits of Twitter users

Supervised learning:

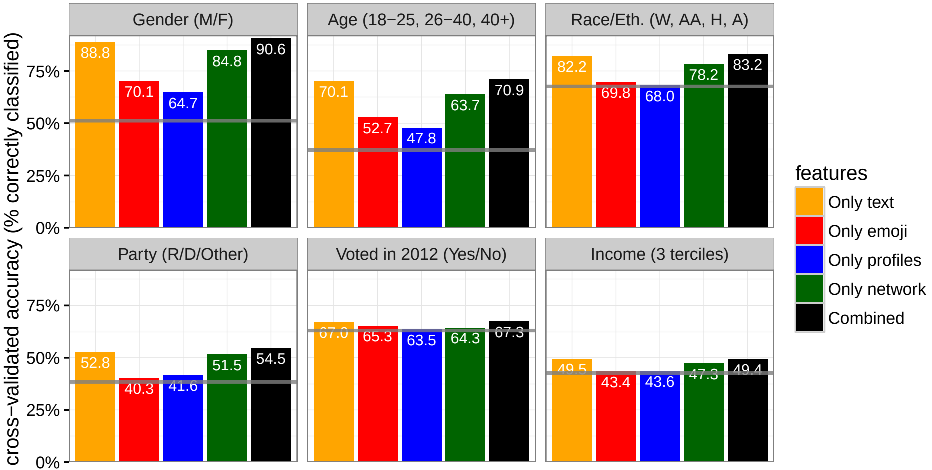
1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Estimating sociodemographic traits of Twitter users

Supervised learning:

1. Collect large dataset of Twitter users whose sociodemographic characteristics are known
2. Train classifier: identify features (emoji, words, profile description, and accounts followed) that are highly predictive of each class
3. Validate: cross-validated accuracy, face validity

Step 3: Validation



Computed on 20% random holdout sample

Step 3: Validation

What are the features with highest predictive power for each category (gender, age, income, race...)?

Validation: age

18-25 👌, 💁, 😑, 😍, 😭, 😏, 😂, 🔫, 🙋, 😒, 🎓. . .

class, college, semester, life, campus, best, literally, like, haha, finals, classes, okay, professor. . .

@SportsCenter, @wizkhalifa, @MileyCyrus, @danieltosh, @instagram, @EmWatson. . .

P: university, major, ❤️, college, student, 16, future, fsu, class, ✨, ☀️, state, ucf, snapchat. . .

26-40 👰, 👶, 👯, 📷, 💪, 😕, ✈️, 😉, 💩, 😆, 🚲. . .

excited, work, amazing, bar, awesome, wedding, #tbt, pretty, #nofilter, ppl, bday, time, lil, #love. . .

@danieltosh, @ConanOBrien, @jtimberlake, @StephenAtHome, @chelseahandler. . .

P: nerd, alum, designer, enthusiast, beer, sports, mommy, lover, gamer, engineer, husband. . .

≥ 40 🎂, 😀, 🏁, 😇, 🐾, ➡️, ⚾, 💝, 🌹, 🌟, 🙏, ⭐, 🎸. . .

daughter, son, good, kids, congratulations, obama, happy, hope, beautiful, grandson, amen. . .

@jimmyfallon, @cnnbrk, @TheEllenShow, @NBCTheVoice, @SteveMartinToGo, @Oprah. . .

P: retired, mom, author, grandmother, dad, kids, mother, conservative, father, children, estate. . .

Validation: party ID

Dem. 🏾, 👀, 😩, 🌈, →, 🏽, 🍸, ✨, 🍷, 💋, 🌹, 💯, 🐧, 💀, 👏, 💃, 💦, 🎬, 🇲🇽, 💅, . . .

philly, barackobama, la, sf, pittsburgh, women, nytimes, philadelphia, smh, president, gop, black, hillaryclinton, gay, republicans . . .

@BarackObama, @rihanna, @maddow, @billclinton, @khloekardashian, @billmaher, @Oprah, @KevinHart4real, @algore, @MichelleObama . . .

PROFILE: philly, activist, writer, liberal, pittsburgh, producer, los, philadelphia, sf, politics, democrat, advocate, angeles, actress, professor, . . .

Rep. 🐊, 🇺🇸, 🏁, ⛳, ⚾, 😳, ❌, ➡️, 🏈, 🐘, ♡, ☀️, ❄️, 👸, ⚡, 🔴, ⭐, ⚡, 💛, ☕, . . .

foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tulsa, byu, seanhannity . . .

@FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan . . .

PROFILE: conservative, jesus, wife, christian, florida, pastor, follower, husband, oklahoma, church, christ, god, married, fsu, grace. . .

This project

- Method to estimate sociodemographic traits:
  a) age, gender, party affiliation, race/ethnicity, past turnout, and income
  b) for any Twitter user in the U.S.

- Tracking a panel of Twitter users in the U.S.
  a) Sociodemographic traits are predicted
  b) 3 applications:
     1. Measurement of issue salience across groups
     2. Early indicator of changes in candidate approval, using post-stratification to recover representativeness
     3. Who spreads misinformation on Twitter?

Measuring Public Opinion with Twitter Data

Building a panel of U.S. Twitter users:

- Random sample of N=500,000 in the U.S.

Measuring Public Opinion with Twitter Data

Building a panel of U.S. Twitter users:

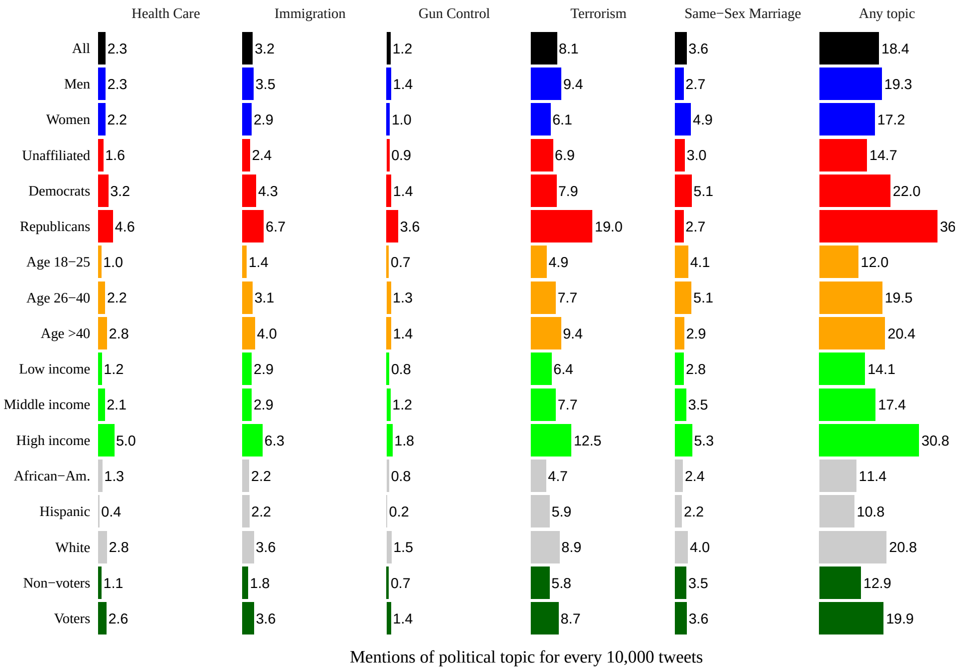
- Random sample of N=500,000 in the U.S.
- Collect all tweets and friends from API
  - $\sim$ 400 million tweets since 01/01/2015
  - $\sim$ 90 million friends
  - ...and counting

Measuring Public Opinion with Twitter Data

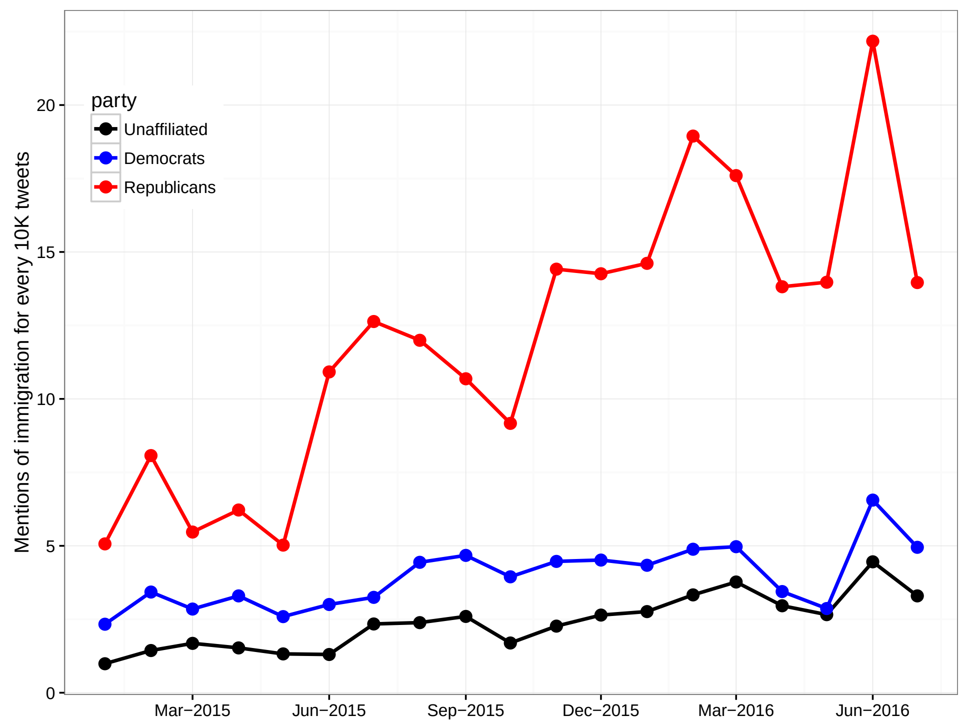
Building a panel of U.S. Twitter users:

- Random sample of N=500,000 in the U.S.
- Collect all tweets and friends from API
  - $\sim$ 400 million tweets since 01/01/2015
  - $\sim$ 90 million friends
  - ... and counting
- Predict sociodemographic traits (network features only)

Application 1: Measuring issue salience across demographic groups



Health Care
Immigration
Gun Control
Terrorism
Same−Sex Marriage
Any topic
All 2.3 3.2 1.2 8.1 3.6 18.4
Men 2.3 3.5 1.4 9.4 2.7 19.3
Women 2.2 2.9 1.0 6.1 4.9 17.2
Unaffiliated 1.6 2.4 0.9 6.9 3.0 14.7
Democrats 3.2 4.3 1.4 7.9 5.1 22.0
Republicans 4.6 6.7 3.6 19.0 2.7 36
Age 18−25 1.0 1.4 0.7 4.9 4.1 12.0
Age 26−40 2.2 3.1 1.3 7.7 5.1 19.5
Age >40 2.8 4.0 1.4 9.4 2.9 20.4
Low income 1.2 2.9 0.8 6.4 2.8 14.1
Middle income 2.1 2.9 1.2 7.7 3.5 17.4
High income 5.0 6.3 1.8 12.5 5.3 30.8
African−Am. 1.3 2.2 0.8 4.7 2.4 11.4
Hispanic 0.4 2.2 0.2 5.9 2.2 10.8
White 2.8 3.6 1.5 8.9 4.0 20.8
Non−voters 1.1 1.8 0.7 5.8 3.5 12.9
Voters 2.6 3.6 1.4 8.7 3.6 19.9
Mentions of political topic for every 10,000 tweets



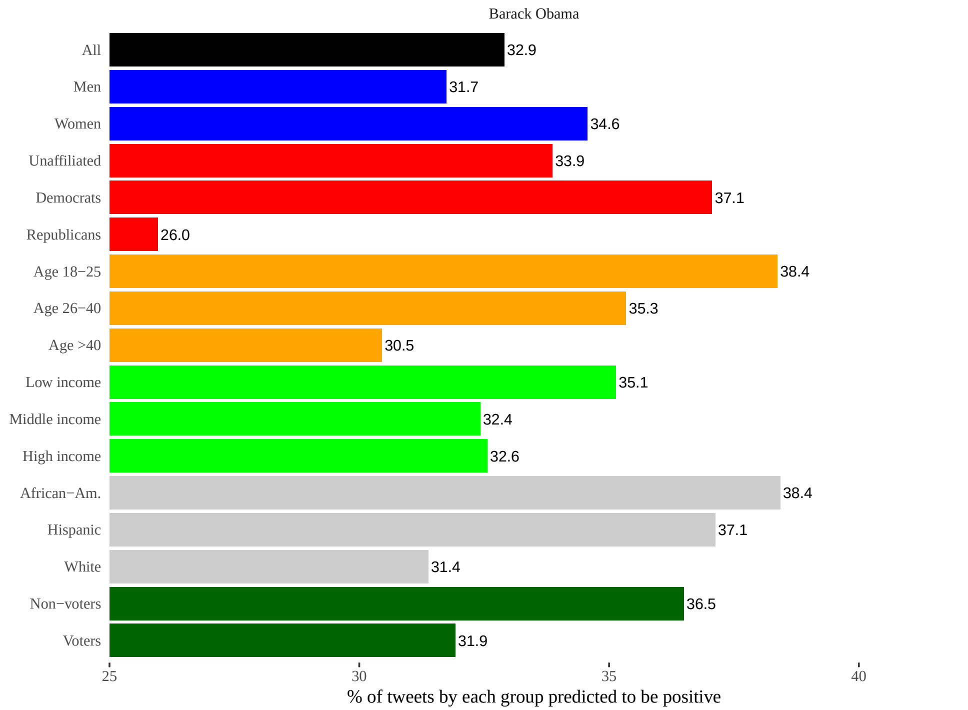
party
Unaffiliated
Democrats
Republicans
Mentions of immigration for every 10K tweets
0
5
10
15
20
Mar−2015
Jun−2015
Sep−2015
Dec−2015
Mar−2016
Jun−2016

Application 2: Measuring presidential job approval

Application 2: Measuring presidential job approval

Steps similar to previous studies:

- Collect tweets mentioning “obama”
- Take random sample and code their sentiment (“does this tweet express support for the president?”)
- Use supervised learning to estimate sentiment of rest of tweets



Barack Obama
All
32.9
Men
31.7
Women
34.6
Unaffiliated
33.9
Democrats
37.1
Republicans
26.0
Age 18−25
38.4
Age 26−40
35.3
Age >40
30.5
Low income
35.1
Middle income
32.4
High income
32.6
African−Am.
38.4
Hispanic
37.1
White
31.4
Non−voters
36.5
Voters
31.9
25
30
35
40
% of tweets by each group predicted to be positive

Multilevel regression with post-stratification

MRP (Lax and Phillips 2009; Park et al 2004; Wang et al 2014)

1. Partition the population into J cells of size N_j based on sociodemographic characteristics

Multilevel regression with post-stratification

MRP (Lax and Phillips 2009; Park et al 2004; Wang et al 2014)

1. Partition the population into J cells of size N_j based on sociodemographic characteristics
2. Regression models to estimate sentiment for each tweet i within each cell:

$$y_{ij} = \alpha + \beta_1 \text{gender}_j + \beta_2 \text{race}_j + \beta_3 \text{party}_j + \beta_4 \text{age}_j + \beta_5 \text{inc.}_j$$

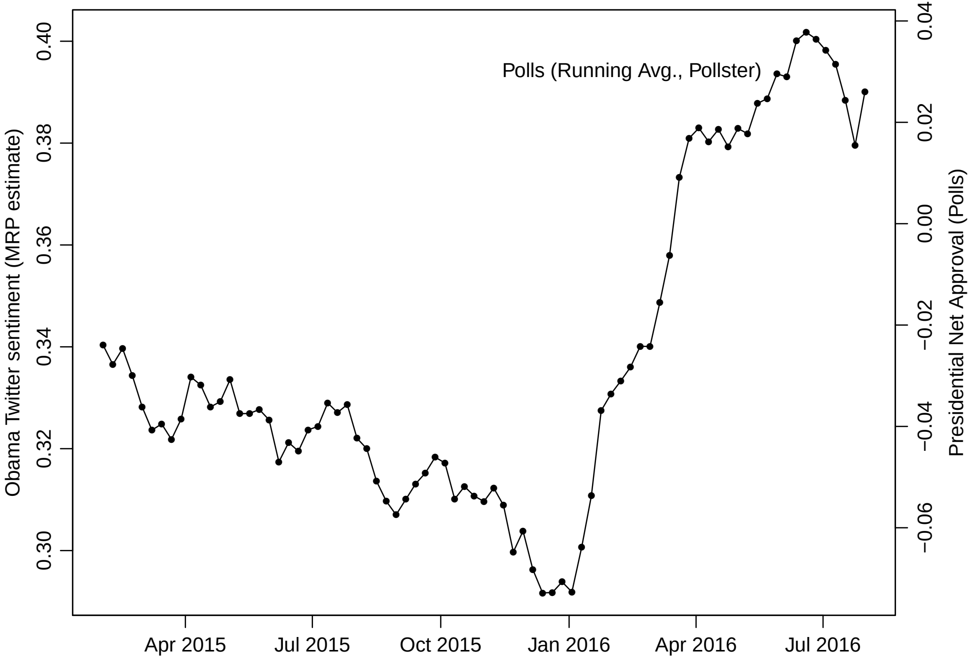
Multilevel regression with post-stratification

MRP (Lax and Phillips 2009; Park et al 2004; Wang et al 2014)

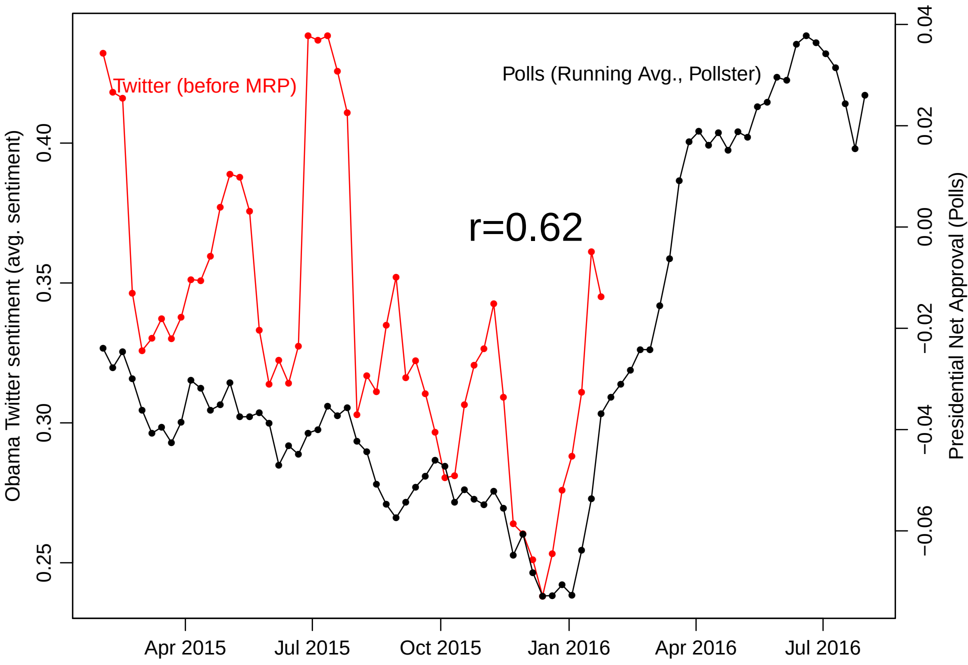
1. Partition the population into J cells of size N_j based on sociodemographic characteristics
2. Regression models to estimate sentiment for each tweet i within each cell:

$$y_{ij} = \alpha + \beta_1 \text{gender}_j + \beta_2 \text{race}_j + \beta_3 \text{party}_j + \beta_4 \text{age}_j + \beta_5 \text{inc.}_j$$

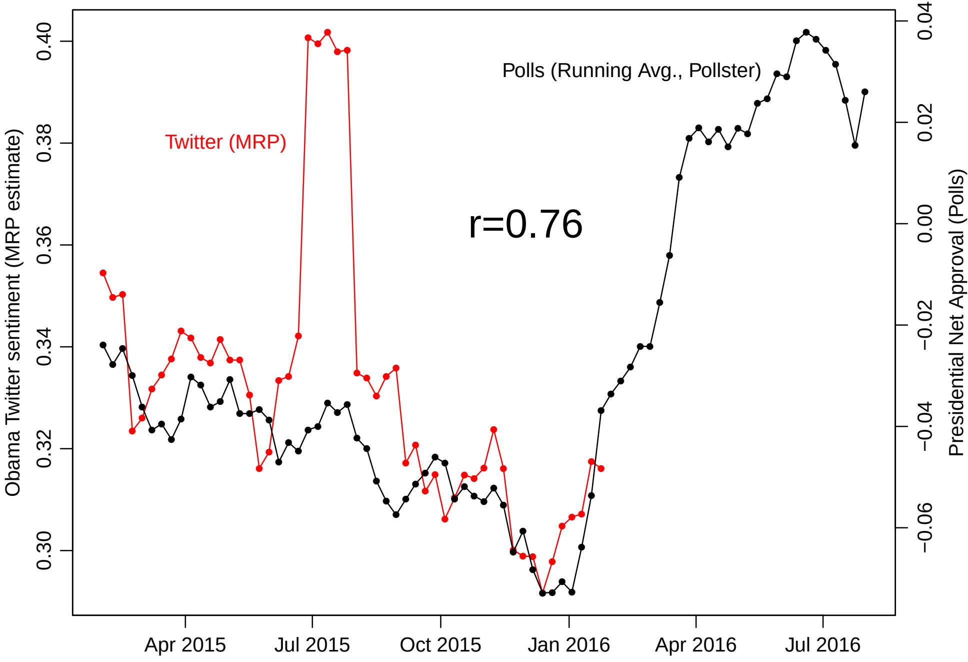
3. Aggregate to cell level and weight by proportion of electorate in each cell (using CCES 2012 data)



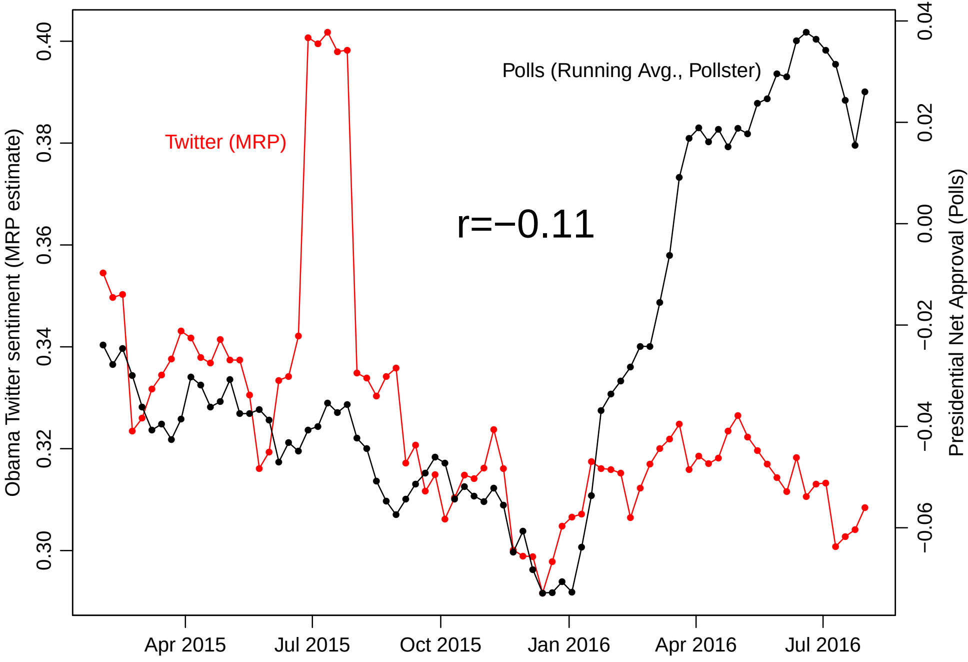
Polls (Running Avg., Pollster)
Obama Twitter sentiment (MRP estimate)
Presidential Net Approval (Polls)
0.30
0.32
0.34
0.36
0.38
0.40
−0.06
−0.04
−0.02
0.00
0.02
0.04
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016

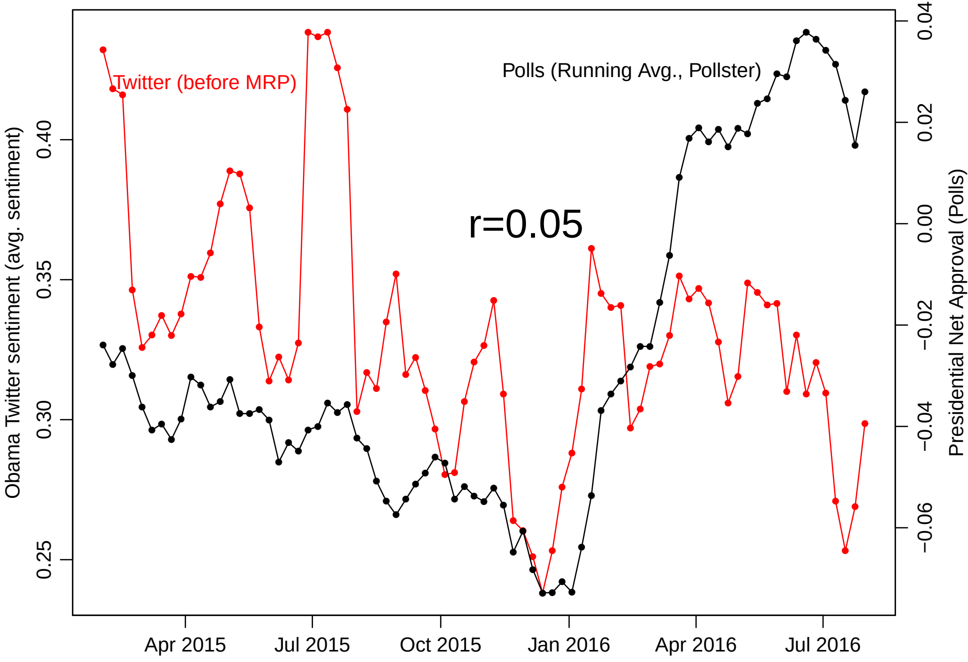


Twitter (before MRP)
Polls (Running Avg., Pollster)
r=0.62
Obama Twitter sentiment (avg. sentiment)
Presidential Net Approval (Polls)
0.25
0.30
0.35
0.40
0.04
0.02
0.00
−0.02
−0.04
−0.06
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016



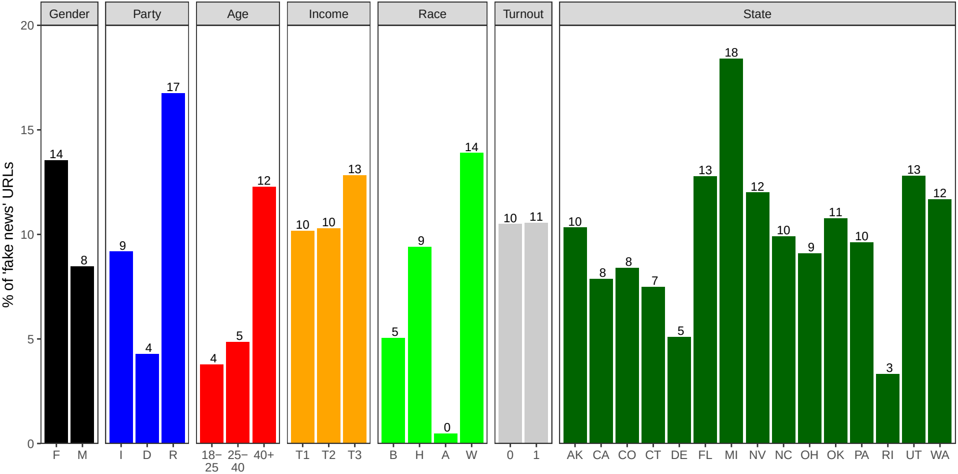
Polls (Running Avg., Pollster)
Twitter (MRP)
r=0.76
Obama Twitter sentiment (MRP estimate)
Presidential Net Approval (Polls)
0.30
0.32
0.34
0.36
0.38
0.40
−0.06
−0.04
−0.02
0.00
0.02
0.04
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016


Polls (Running Avg., Pollster)
Twitter (MRP)
r=−0.11
Obama Twitter sentiment (MRP estimate)
Presidential Net Approval (Polls)
0.30
0.32
0.34
0.36
0.38
0.40
−0.06
−0.04
−0.02
0.00
0.02
0.04
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016


Twitter (before MRP)
Polls (Running Avg., Pollster)
r=0.05
Obama Twitter sentiment (avg. sentiment)
Presidential Net Approval (Polls)
0.25
0.30
0.35
0.40
0.04
0.02
0.00
−0.02
−0.04
−0.06
Apr 2015
Jul 2015
Oct 2015
Jan 2016
Apr 2016
Jul 2016

Application 3: Spread of misinformation during 2016 presidential election campaign

Spread of misinformation during 2016 election



Data: URLs shared by panel that correspond to 145 domains manually annotated as spreading mostly misinformation, 10/02 to 11/09/2016.

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample
  - Ethical implications

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample
  - Ethical implications
- Other applications:

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample
  - Ethical implications
- Other applications:
  - Public opinion on “unpolled” topics

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample
  - Ethical implications
- Other applications:
  - Public opinion on “unpolled” topics
  - Inequality in exposure to information

Discussion

Contributions:

- Method to predict sociodemographic traits of Twitter users and track behavior over time
- Useful to measure public opinion and make meaningful comparisons about attention to issues (panel design)
- Limitations:
  - Can we treat tweets as survey responses?
  - Training dataset not a random sample
  - Ethical implications
- Other applications:
  - Public opinion on “unpolled” topics
  - Inequality in exposure to information
  - Offline vs online geographical segregation

Thanks!

website: pablobarbera.com

github: pablobarbera

twitter: @p_barbera