Qualitative Insights from Social Media Data: Seeing the Forest and the Trees

AAPOR, 2022

Johann Gagnon-Bartsch, University of Michigan
Frederick Conrad, University of Michigan
Michael Schober, The New School
Robyn Ferg, Westat
Juejue Wang, University of Michigan
Rebecca Dolgin, The New School
Mao Li, University of Michigan
Paul Beatty, US Census Bureau

May 11, 2022

Our Team

Johann Gagnon-Bartsch
(U. Michigan)

Fred Conrad
(U. Michigan)

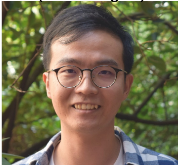
Michael Schober
(New School)

Robyn Ferg
(U. Michigan, Westat)

Juejue Wang
(U. Michigan)

Rebecca Dolgin
(New School)

Mao Li
(U. Michigan)



Paul Beatty
(US Census Bureau)

Friday, 1:15pm

Session 7: Using Google Searches and Tweets for Social Research

New Thinking About When Social Media and Survey Responses May Align

Frederick G. Conrad[1]
Michael F. Schober[2]
Johann A. Gagnon-Bartsch[1]
Robyn Ferg[3]
Mao Li[1]
Paul C. Beatty[4]

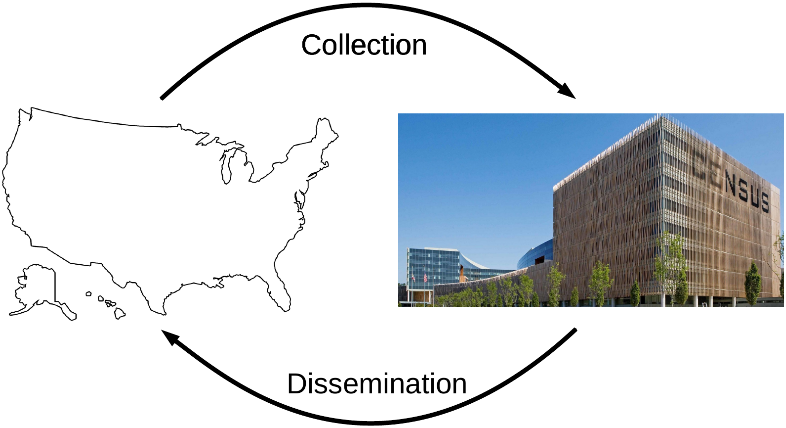
[1]University of Michigan
[2]New School
[3]Westat
[4]US Census Bureau

Note: Views and ideas expressed here are those of the authors and not of the US Census Bureau

Our Goal

Use social media data to improve Census Bureau activities, broadly considered

Our Goal



Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.
- Improve questionnaire design

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.
- Improve questionnaire design
  - Question phrasing, word choice, etc.

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.
- Improve questionnaire design
  - Question phrasing, word choice, etc.
  - Questions not asked

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.
- Improve questionnaire design
  - Question phrasing, word choice, etc.
  - Questions not asked
- Sampling strategies

Collection

Examples of how insights from social media may improve data collection:

- Improve outreach, reduce non-response, e.g., by addressing trust concerns
  - Data won't be shared with other agencies
  - Data won't be hacked
  - Data will be used for good, etc.
- Improve questionnaire design
  - Question phrasing, word choice, etc.
  - Questions not asked
- Sampling strategies
  - Oversample demographic groups or geographical regions where responses may be rapidly changing

Dissemination

Examples of how insights from social media may improve data analysis and dissemination:

- Understand how data is used, and by whom

Dissemination

Examples of how insights from social media may improve data analysis and dissemination:

- Understand how data is used, and by whom
- Discover unmet needs

Dissemination

Examples of how insights from social media may improve data analysis and dissemination:

- Understand how data is used, and by whom
- Discover unmet needs
- Improve data quality, e.g., survey precision

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)
  - Cody, Reagan, Dodds, and Danforth (2016)

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)
  - Cody, Reagan, Dodds, and Danforth (2016)
  - Daas and Puts (2014)

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)
  - Cody, Reagan, Dodds, and Danforth (2016)
  - Daas and Puts (2014)
- Further investigation revealed that some of the reported associations may have been spurious

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)
  - Cody, Reagan, Dodds, and Danforth (2016)
  - Daas and Puts (2014)
- Further investigation revealed that some of the reported associations may have been spurious
  - Conrad, et al. (2019)

Background (from our point of view)

- Considerable enthusiasm for using social media data to augment, or even replace, traditional surveys
  - AAPOR task force on big data (Murphy, et al., 2014)
  - O'Connor, Balasubramanyan, Routledge, and Smith (2010)
  - Cody, Reagan, Dodds, and Danforth (2016)
  - Daas and Puts (2014)
- Further investigation revealed that some of the reported associations may have been spurious
  - Conrad, et al. (2019)
- Open question: When and how can social media data be used for public opinion research?

Paradigms

- Social media as analogous to a survey

Paradigms

- Social media as analogous to a survey
- Social media as analogous to a focus group

This talk: Qualitative insights

- What concerns are expressed?

This talk: Qualitative insights

- What concerns are expressed?
- What misinformation is present?

This talk: Qualitative insights

- What concerns are expressed?
- What misinformation is present?
- Who is citing Census data, and how?

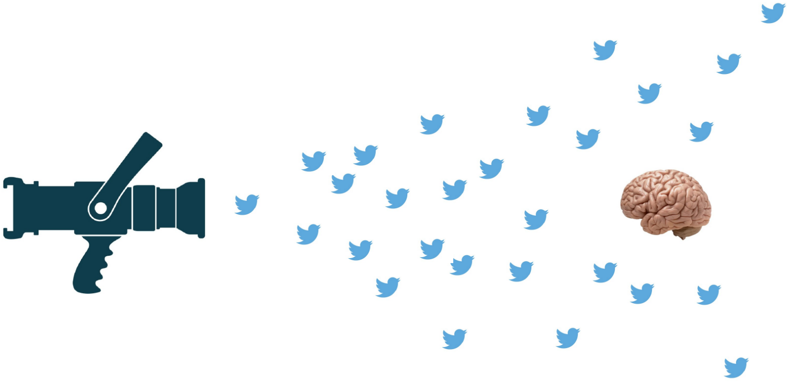
This talk: Qualitative insights

- What concerns are expressed?
- What misinformation is present?
- Who is citing Census data, and how?
- More generally: **What is being said?**

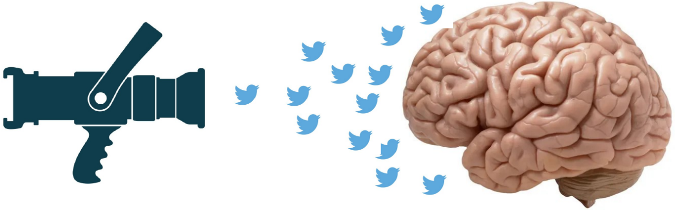
This talk: Qualitative insights

- What concerns are expressed?
- What misinformation is present?
- Who is citing Census data, and how?
- More generally: **What is being said?**
- Just read the tweets!

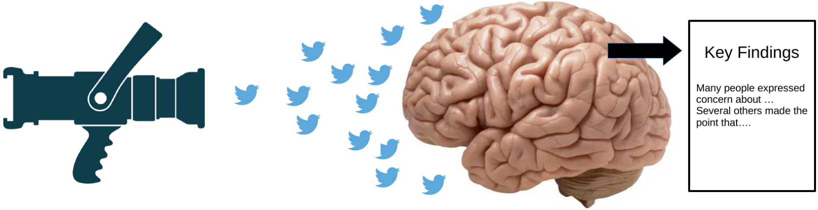
Key Challenge



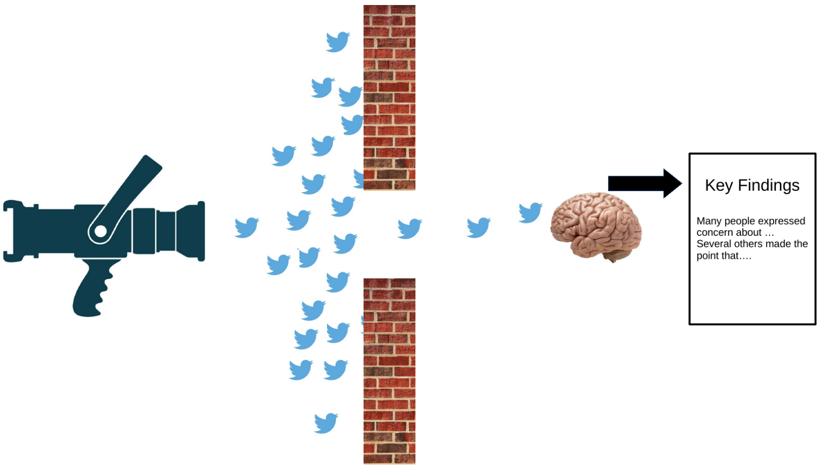
Ideal Solution



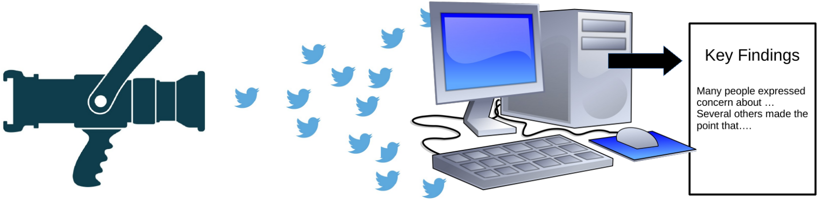
Ideal Solution



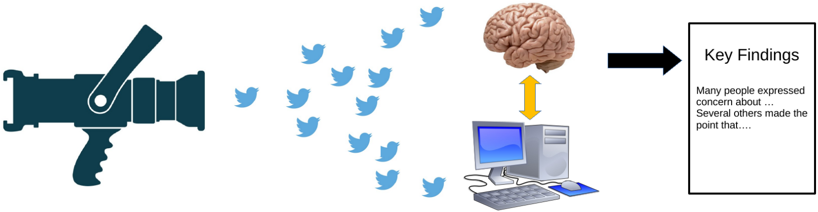
Random Sampling



Computer Analysis



Interactive Analysis



Questions

1. What is the optimal level and type of automation to allow a human analyst to obtain novel, timely, and credible insights?

Questions

1. What is the optimal level and type of automation to allow a human analyst to obtain novel, timely, and credible insights?
2. To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Questions

1. What is the optimal level and type of automation to allow a human analyst to obtain novel, timely, and credible insights?
2. To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?
3. To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Question 1

What is the optimal level and type of automation to allow a human analyst to obtain novel, timely, and credible insights?

Tweet Browser:

- Interactive tool to allow a human to explore and digest a large social media corpus.
- Goal is to see *"both the forest and the trees."*

Tweet Browser

- Top down:

Tweet Browser

- Top down:
    - Specify keywords, hashtags, etc.

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated "topics"
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling
  - Either way, the process may be recursive, so users can “drill down” into a topic

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated "topics"
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling
  - Either way, the process may be recursive, so users can "drill down" into a topic
- Browse

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling
  - Either way, the process may be recursive, so users can “drill down” into a topic
- Browse
  - Random subsets of tweets

Tweet Browser

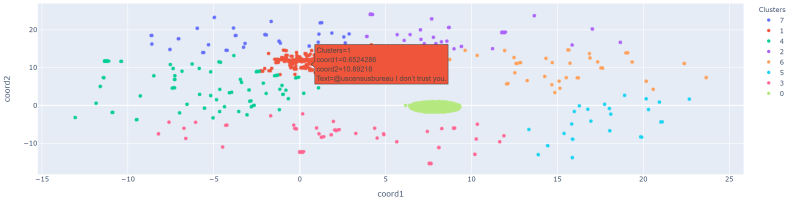
- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated “topics”
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling
  - Either way, the process may be recursive, so users can “drill down” into a topic
- Browse
  - Random subsets of tweets
  - Graphical depictions, e.g., of important words

Tweet Browser

- Top down:
  - Specify keywords, hashtags, etc.
  - Dates, whether retweet, etc.
  - Sentiment, etc.
- Bottom up:
  - Computer-generated "topics"
  - User specified parameters (algorithm, number of topics, etc.)
- Exploration can be **iterative** and **interactive**
  - May start with computer-generated topics, and refine manually (keywords, dates, etc.)
  - May start with manually specified subsets, and then apply automated topic modeling
  - Either way, the process may be recursive, so users can "drill down" into a topic
- Browse
  - Random subsets of tweets
  - Graphical depictions, e.g., of important words
  - Distribution of tweets over time, geography

Tweet Browser

The plot below contains 26311 total tweets.



Most common words in each cluster:

Cluster	Proportion of Tweets	Number of Tweets	Top 5 Stemmed Words
0	0.557	14644	but californian questionnaires repkatieport struggl
4	0.053	1397	bug debt jimmyrayn trust unionist
6	0.023	617	com news theguardian trust us
3	0.043	1124	case judg mail probe trust
1	0.241	6346	2020censu count mail trust uscensusbureau
5	0.019	511	crime given minggao26 propens trust
2	0.037	983	cel good mail sun trust
7	0.026	689	baddcompani hansilowang right st trust

Longer-term goal

- Systematically explore different workflows

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.
  - Hypothesis driven?

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.
  - Hypothesis driven?
  - Spend time reading many tweets, or glance at summaries?

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.
  - Hypothesis driven?
  - Spend time reading many tweets, or glance at summaries?
- Requires systematic evaluation of insights

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.
  - Hypothesis driven?
  - Spend time reading many tweets, or glance at summaries?
- Requires systematic evaluation of insights
  - Consistent format for reporting insights

Longer-term goal

- Systematically explore different workflows
  - Purely top-down vs. purely-bottom up vs. begin top-down, etc.
  - Hypothesis driven?
  - Spend time reading many tweets, or glance at summaries?
- Requires systematic evaluation of insights
  - Consistent format for reporting insights
  - Scoring for quality, novelty, credibility

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked
  - Whether data is accurate

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked
  - Whether data is accurate
  - Whether data will be used for good

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked
  - Whether data is accurate
  - Whether data will be used for good
- Two approaches:

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked
  - Whether data is accurate
  - Whether data will be used for good
- Two approaches:
  - Read a random sample of "census" tweets from 2020

Question 2

To what extent do the insights obtained from social media posts reproduce those obtained from traditional focus groups?

Case study:

- 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS)
- Can attitudes and opinions revealed in the focus group also be discovered on Twitter?
- Focus specifically on trust concerns:
  - Whether data might be shared with other agencies
  - Whether data might be hacked
  - Whether data is accurate
  - Whether data will be used for good
- Two approaches:
  - Read a random sample of “census” tweets from 2020
  - “Search” for tweets using BERT-derived distance metrics

Similar content

Quote from CBAMS:

"The government has always been intrusive as it is, and it's probably a level of intrusion. That's why people are like, 'Hold on, what you want to know what's in my bed, at my house, and who's using my toilet? You should go mind your business."

Tweets:

- *"@Mededitor A surprisingly lean census. I recall in other years being asked to report the number of toilets in my home!"*
- *"The census is full of questions that I'm not sure I'd want a government to know about me."*

Similar content

- We explored several other quotes from CBAMS
- In all cases, we were able to find tweets expressing similar concerns / opinions
- We conclude that, *at least on this topic*, attitudes and opinions revealed in the focus group can also be discovered on Twitter
- Still an open question:
  - Would the *insights* derived from an analysis of social media replicate those from a focus group?
  - For example, would a "blind" analysis have also highlighted the same trust concerns?

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of “census” tweets from 2020

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of "census" tweets from 2020
  - Read a random sample of "census + citizenship" tweets from 2020

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of “census” tweets from 2020
  - Read a random sample of “census + citizenship” tweets from 2020
  - Sampled in various ways; several hundred tweets

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of "census" tweets from 2020
  - Read a random sample of "census + citizenship" tweets from 2020
  - Sampled in various ways; several hundred tweets
- Findings:

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of “census” tweets from 2020
  - Read a random sample of “census + citizenship” tweets from 2020
  - Sampled in various ways; several hundred tweets
- Findings:
  - Several opinions were prominent on Twitter that did not appear in CBAMS

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of "census" tweets from 2020
  - Read a random sample of "census + citizenship" tweets from 2020
  - Sampled in various ways; several hundred tweets
- Findings:
  - Several opinions were prominent on Twitter that did not appear in CBAMS
  - Example: Refusal to participate in the Census if it did *not* contain a citizenship question

Question 3

To what extent do social media posts provide insights that complement or are entirely distinct from those obtained from traditional focus groups?

Case study:

- CBAMS
- Approach:
  - Read a random sample of "census" tweets from 2020
  - Read a random sample of "census + citizenship" tweets from 2020
  - Sampled in various ways; several hundred tweets
- Findings:
  - Several opinions were prominent on Twitter that did not appear in CBAMS
  - Example: Refusal to participate in the Census if it did *not* contain a citizenship question
  - Highlights ability of social media to provide insights on opinions of otherwise hard-to-reach populations

Thank You

We gratefully acknowledge support from the US Census Bureau through a cooperative agreement with the University of Michigan and the New School, "New Approaches to Analyzing Social Media Content for Enhancing Census Bureau Data" (award #CB20ADR0160002)