Measuring the Strength of Attitudes in Social Media Data

October 27, 2018

Ashley Amaya, Ruben Bach, Frauke Kreuter, & Florian Keusch

The Opportunity

- The amount of stored data will be ≥ 45,000 exabytes by 2020! (Quartz 2015)
  - A significant amount of these data are on social media.

- The benefits of social media data:
  - Constant / instantaneous
  - "Free"
  - No interviewer effects

- And, the challenges:
  - Coding error
  - Missingness
  - Measurement

Research Questions

- Can social media text data be used to produce similar attitude distributions as survey data?

- If not, why not?
  - Coding
  - Missingness
  - Measurement

Can social media text data be used to produce similar attitude distributions as survey data?

Data

2016 European Social Survey

- Traditional in-person survey
- 2,852 German residents, aged 15+
  - Limited to 16+ for analysis
- RR1 = 30.61%
- Includes several attitudinal questions:
  - Politics (interest & ideology)
  - Immigration
  - EU
  - Trust in individuals
  - Gay Rights
  - Climate Change

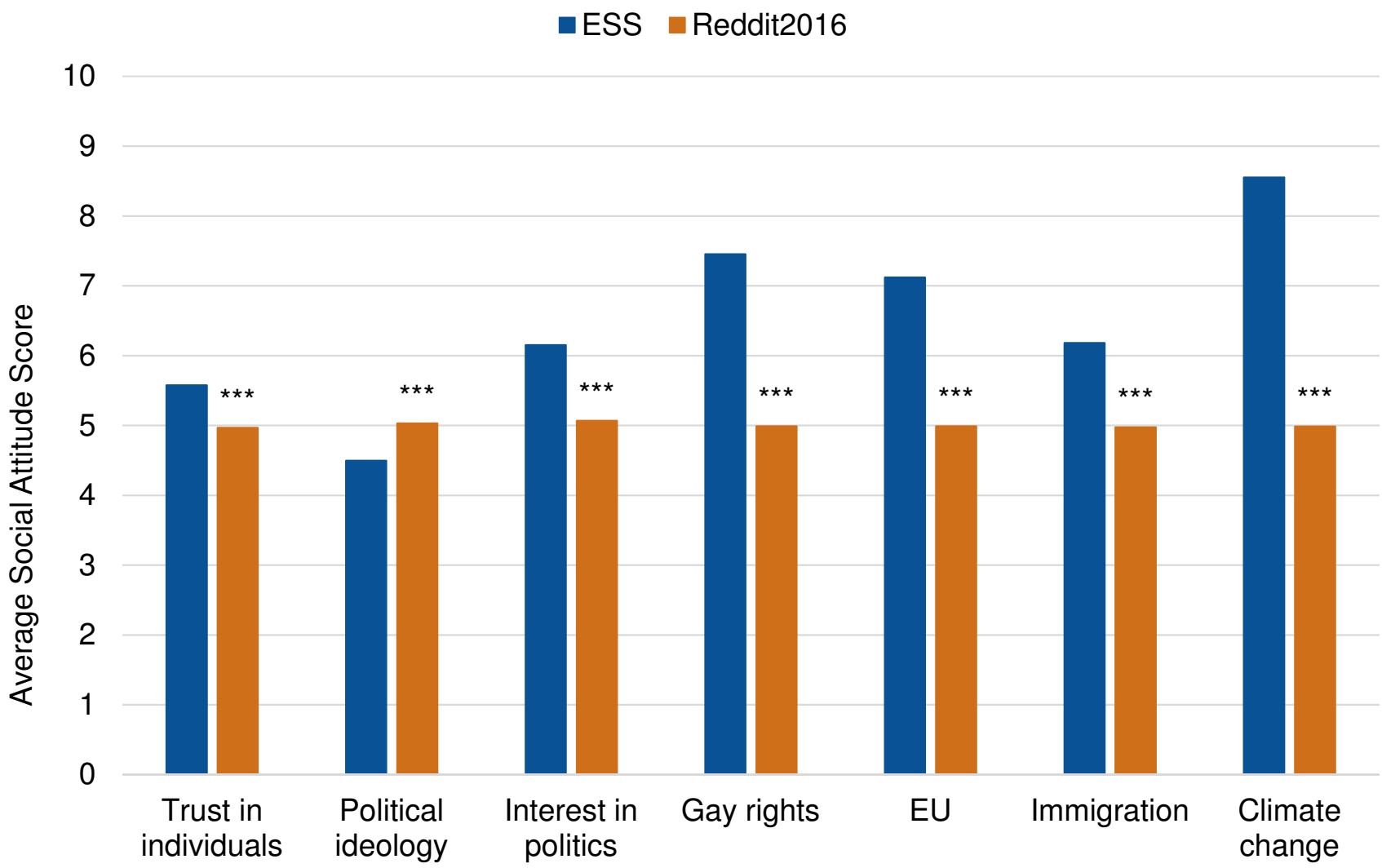
2016 Reddit

- Anonymous, public social media site
- 5th most popular site in Germany (Alexa 2018)
  - 10.7mil submissions/month (worldwide) (Reddit 2017)
- Includes all topics!
  - 367 subreddits
  - 463k total submissions
  - 16.6k topical authors
  - 4k – 12k authors per topic w/ scores

Coding Methods

- Tokenized the submissions
- Subset to English and German submissions
- Identified relevant submissions
  - Boolean search
  - Combined posts & comments for topic assignment
- Assigned sentiment score
  - Collapsed relevant submissions by author by language
  - Computed average weighted value across languages
  - Scaled scores to ESS scale (0-10)

Average Social Attitude Score by Source



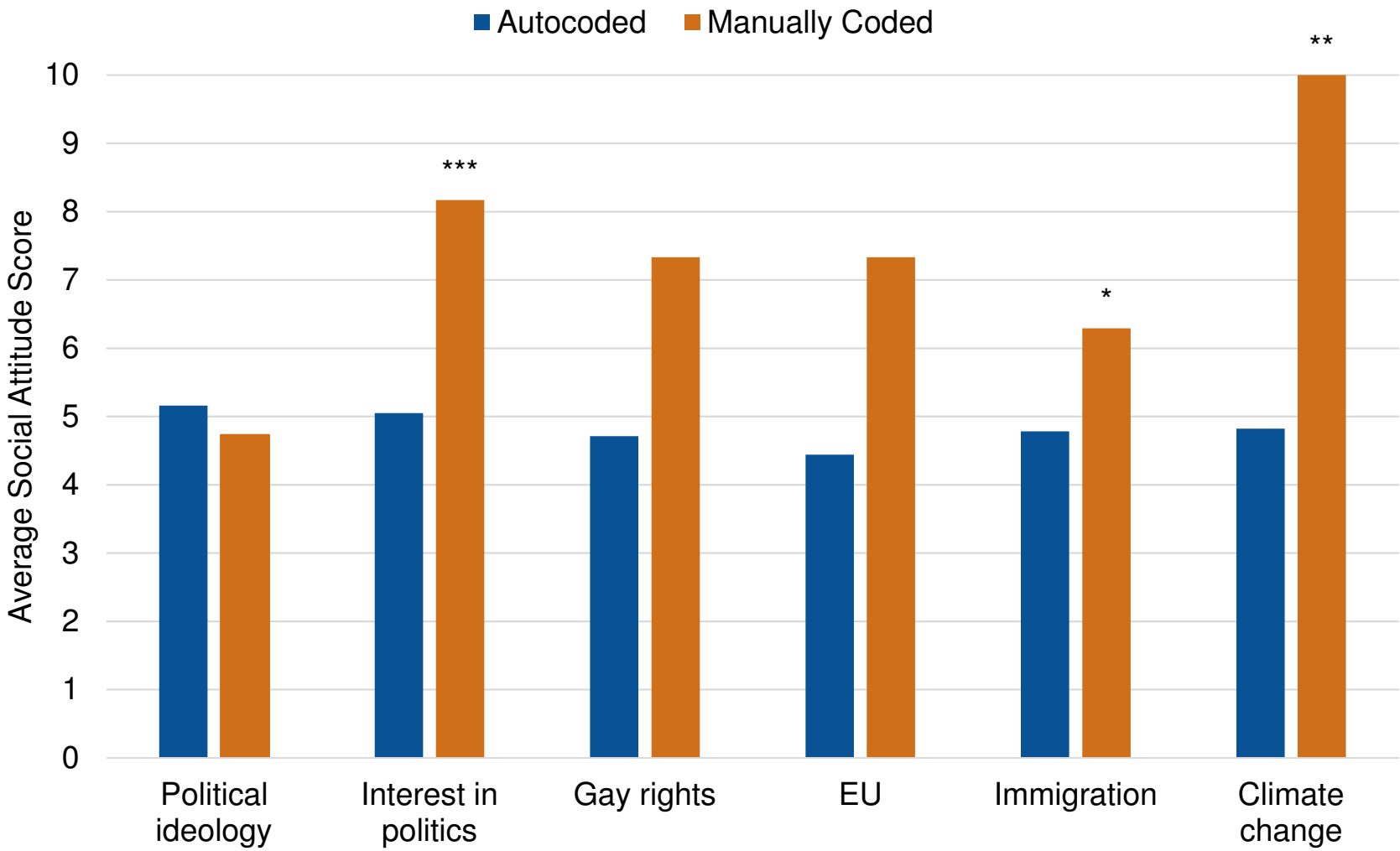
* $p<0.05$, ** $p<0.01$, *** $p<0.001$

If not, why not?

Data: Coding Error

- 2018 Reddit posts
  - ~11k submissions
  - 26 authors
- Coding
  - Autocoded using same methods as above
  - Manual coding
    - Flagged relevant submissions
    - Collapsed relevant submissions by author
    - Assigned value on ESS scale (0-10)
- Compared autocoded vs. manually coded

Average Social Attitude Score by Coding Method

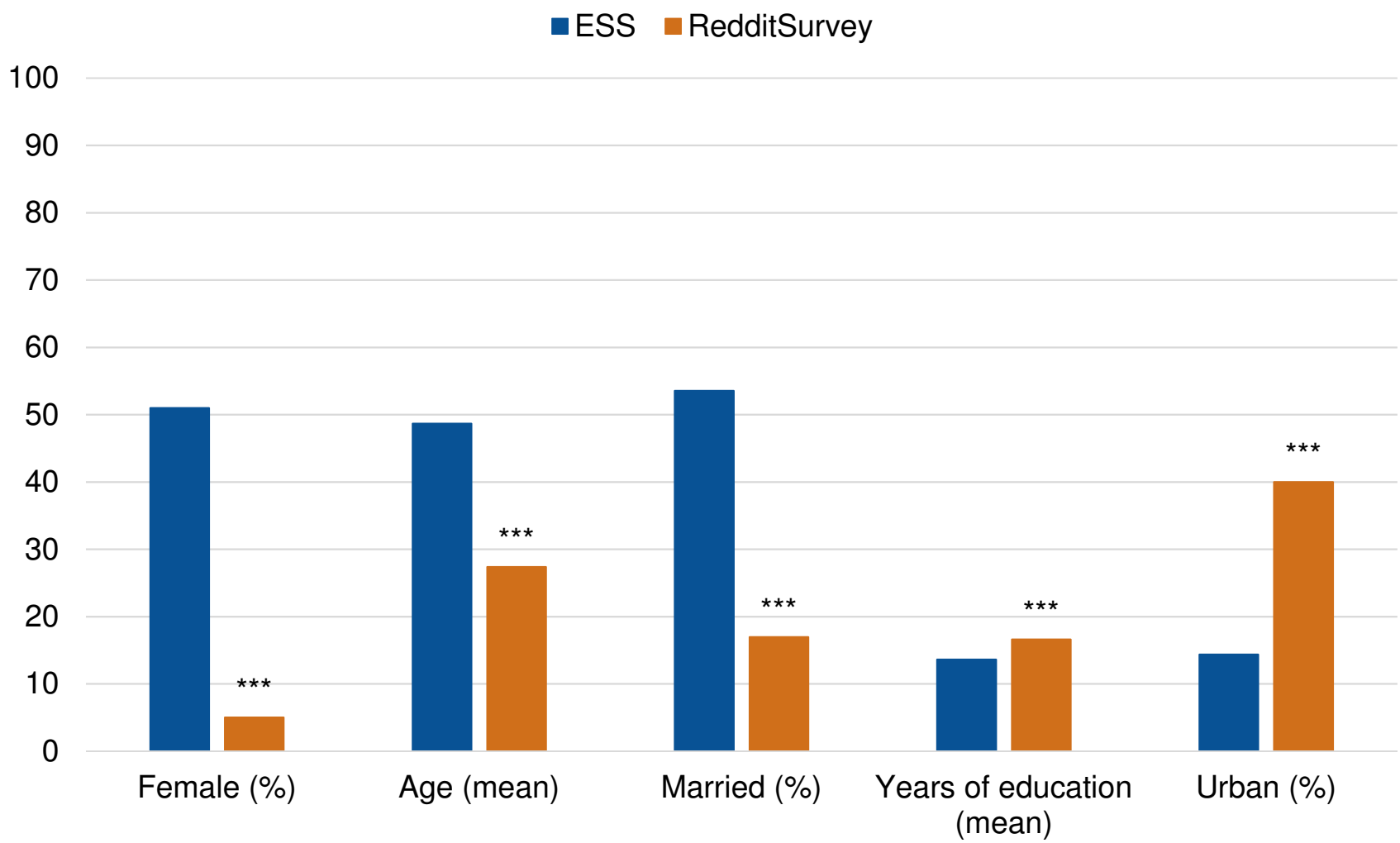


* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Data: Missingness

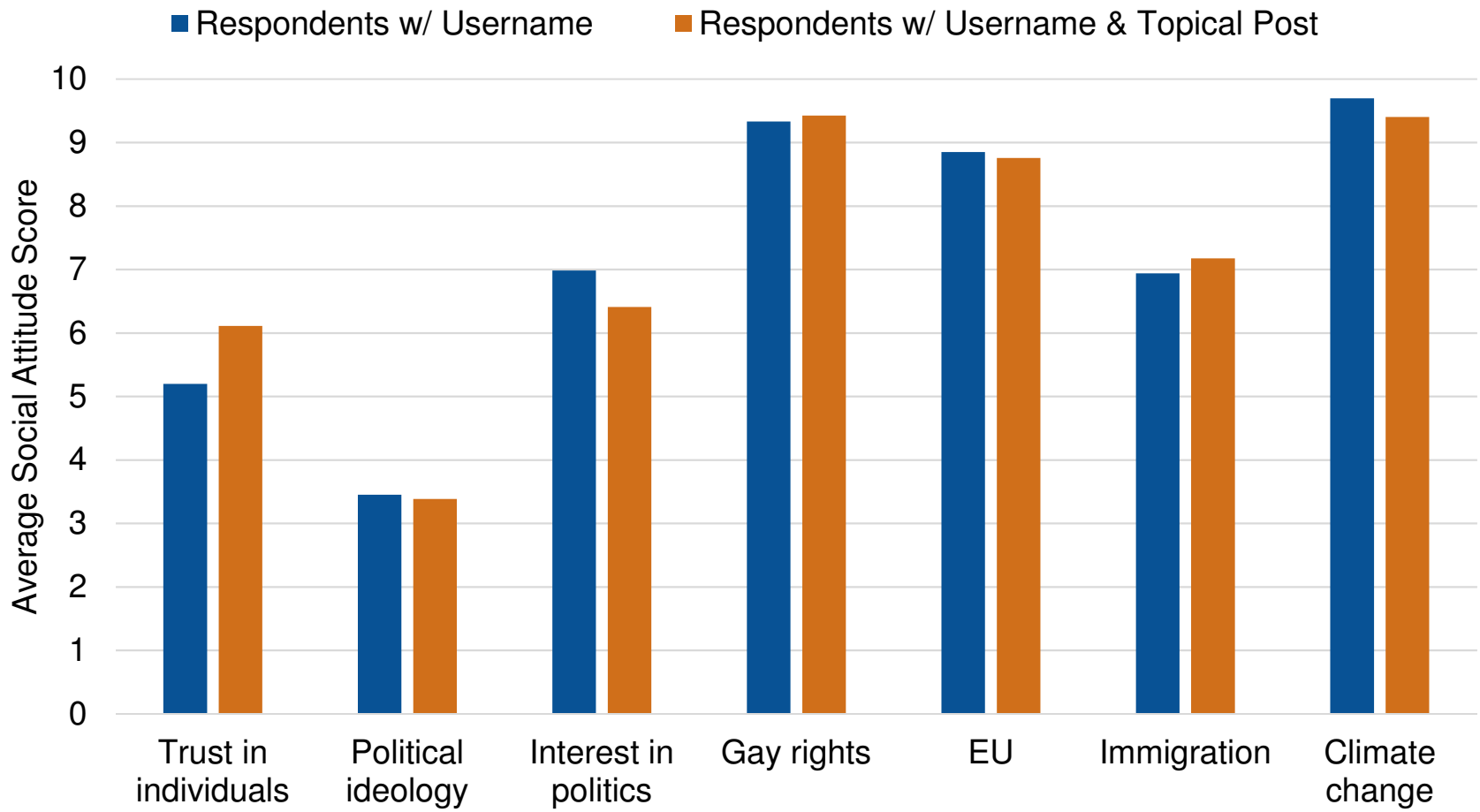
- 2018 Reddit Survey
  - 60 respondents
- ESS
- Analyses
  - Compared demographics of survey vs. ESS
  - Compared ***all*** survey respondents vs. survey respondents who **posted on a topic**

Socio-demographic Distributions by Source



* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Average Social Attitude Score by Whether they Share Attitude on Reddit



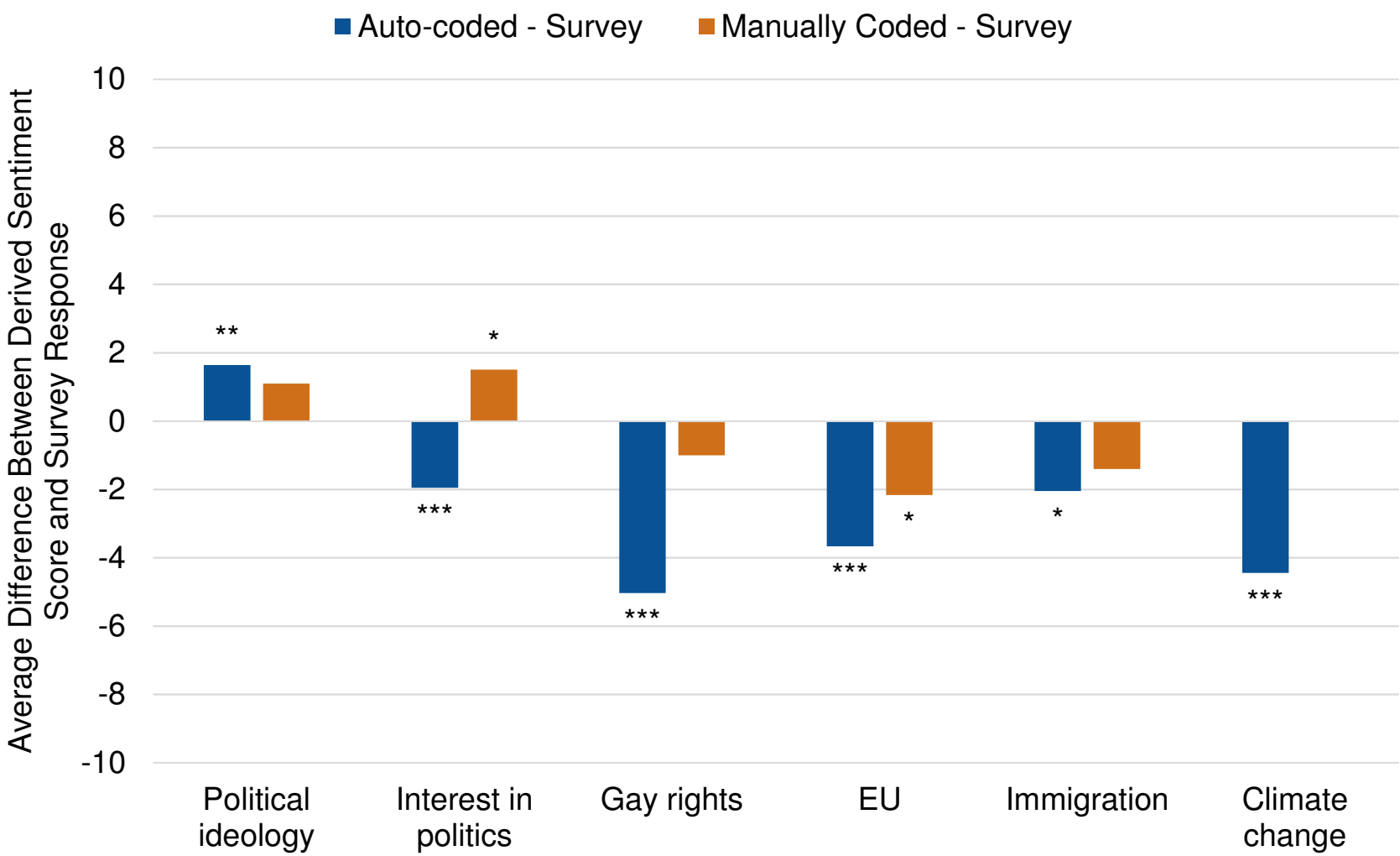
* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Data: Measurement

- 2018 Reddit Survey
  - 60 respondents

- Analyses
  - Compared survey responses vs. Reddit posts

Measurement Error Results



* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Summary

- Can social media text data be used to produce similar attitude distributions as survey data?
  - No

- If not, why not?
  - Coding error
    - Topic models did not pick up correct submissions
  - Missingness
    - Reddit is not representative of the population
  - Measurement
    - Survey responses different from both auto- and manually coded sentiment scores

Up Next

- Revise the paper
- Finish up some complementary papers
- Track over time & sources

More Information

Ashley Amaya
Research Survey Methodologist
202.728.2486
aamaya@rti.org