Datasets for Measuring Social Phenomena

Indira Sen

University of Konstanz

Research Projects in Computational Studies of Social Phenomena

https://github.com/Indiiigo/css_research_projects/tree/main

Agenda

- ❖ Ethical Aspects of Working with Digital Traces
- ❖ Datasets for Projects
- ❖ Exploring the datasets

recap

Understanding Social Phenomena

Prototypical CSS Studies Leveraging Digital Trace Data

‘Readymade’ and ‘Custommade’ data

Examples of Research with Readymade data: repurposes existing data, like web or social media data (but could also be other types of content — books, newspaper articles…)

Examples of Research with Custommade data: creates surveys or survey experiments to test perceptions of politicians based on their gender

[As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians]()

WebSci ’19, June 30–July 3, 2019, Boston, MA, USA

As the Tweet, so the Reply? Gender Bias in Digital Communication with Politicians

Armin Mertens
Cologne Center for Comparative Politics
Cologne, Germany
mertens@wiso.uni-koeln.de

Franziska Pradel
Cologne Center for Comparative Politics
Cologne, Germany
pradel@wiso.uni-koeln.de

Ayjeren Rozyjumayeva
Faculty of Management, Economics and Social Sciences
Cologne, Germany
rozyju

Jens Wäckerle
Cologne Center for Comparative Politics
Cologne, Germany

ABSTRACT
This study investigates
ital platforms by consi
on Twitter and how the
social identity theory,
in individual tweets c
in 2017. Besides sentim
of personal- vs. job-rel
with structural topic m
communication on Twi
gender. However, we fi
directed at politicians:
likely to be reduced to
compared to male polit

Article

The Price of Power: Power Seeking and Backlash Against Female Politicians

Personality and Social Psychology Bulletin
36(7) 923–936

Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146167210371949
http://pspb.sagepub.com
SAGE

Tyler G. Okimoto[1] and Victoria L. Brescoll[1]

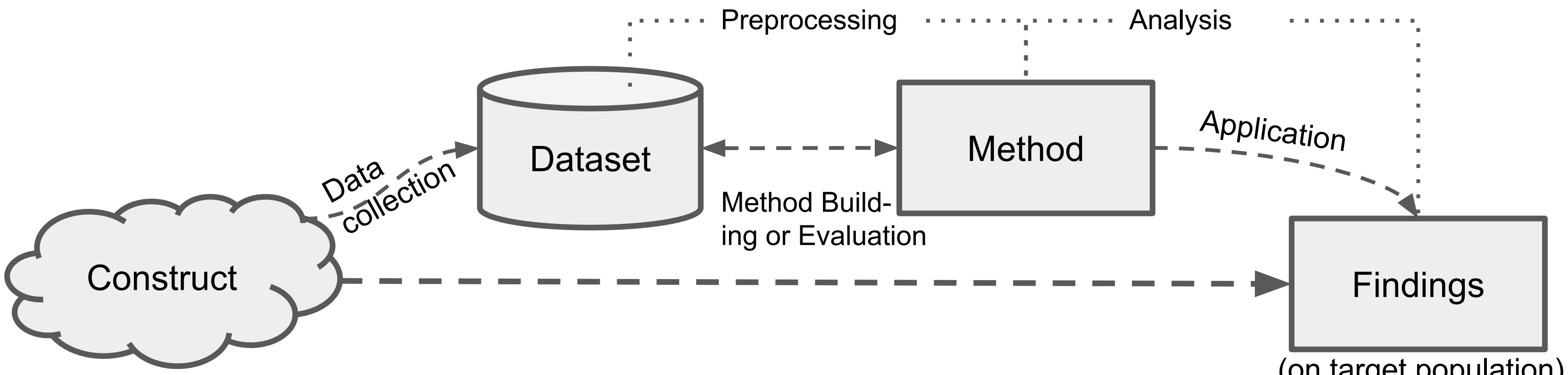
Abstract
Two experimental studies examined the effect of power-seeking intentions on backlash toward women in political office. It was hypothesized that a female politician’s career progress may be hindered by the belief that she seeks power, as this desire may violate prescribed communal expectations for women and thereby elicit interpersonal penalties. Results suggested that voting preferences for female candidates were negatively influenced by her power-seeking intentions (actual or perceived) but that preferences for male candidates were unaffected by power-seeking intentions. These differential reactions were partly explained by the perceived lack of communality implied by women’s power-seeking intentions, resulting in lower perceived competence and feelings of moral outrage. The presence of moral-emotional reactions suggests that backlash arises from the violation of communal prescriptions rather than normative deviations more generally. These findings illuminate one potential source of gender bias in politics.

Keywords
gender stereotypes, backlash, power, politics, intention, moral outrage

Received June 5, 2009; revision ac

[The Price of Power: Power Seeking and Backlash Against Female Politicians]()

Prototypical Pipeline - Artifacts and Steps



Prototypical Pipeline: From Construct to Measurement

From Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, January). The meaning and measurement of bias: lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 706-706).

Prototypical Pipeline: Data Collection

collect data potentially containing tangible signals regarding our construct

Content

videos

text

images

A rough list of (inter-connected) methods

- Network analysis
- Simulations
- NLP / Text-as-data
- Content analysis
- Human Computer Interaction (‘social computing’)
- Causal inference
- Surveys + digital traces
- …?

Let’s do a another quick activity

- Go back to the paper that you had picked (or pick a new one).

- What type of method did the authors use? Is it appropriate for what is being studied?

- Discuss

Ethics

New data, new methods, new challenges…

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

Experimental evidence of massive-scale emotional contagion through social networks

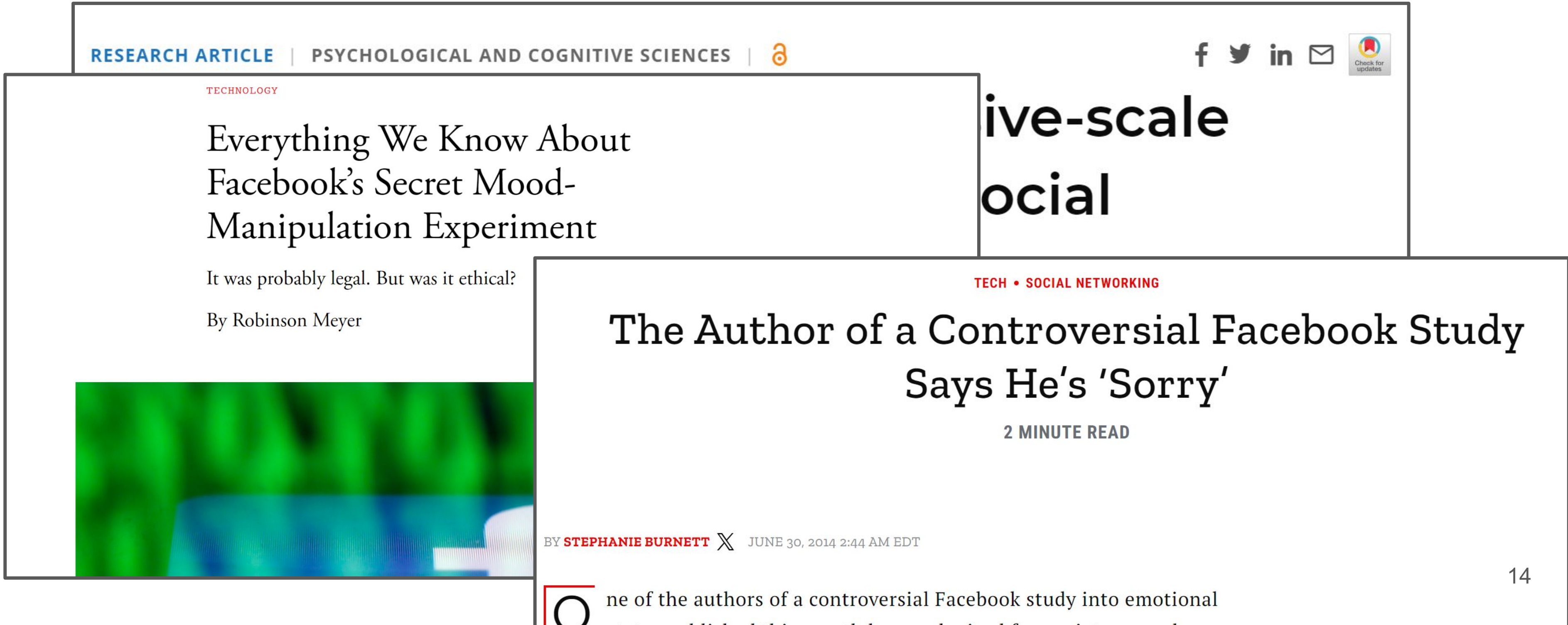
Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock Authors Info & Affiliations

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

June 2, 2014 | 111 (24) 8788-8790 | https://doi.org/10.1073/pnas.1320040111

[**Experimental** evidence of massive-scale emotional **contagion** through social networks]

New data, new methods, new challenges…


RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

ive-scale

ocial

TECHNOLOGY

Everything We Know About Facebook's Secret Mood-Manipulation Experiment

It was probably legal. But was it ethical?

By Robinson Meyer

TECH • SOCIAL NETWORKING

The Author of a Controversial Facebook Study Says He's 'Sorry'

2 MINUTE READ

BY STEPHANIE BURNETT JUNE 30, 2014 2:44 AM EDT

One of the authors of a controversial Facebook study into emotional

Potential stakeholders, harms

- Analyses
  - Experimental approaches can have deeper repercussions
  - Can set narratives in research community and beyond
  - Findings used for policy-making
- Datasets
  - Can be used unreflectly, propagating biases, or simply wrong conclusions
- Methods
  - Biased ML methods can encode substantial biased, but even simpler methods are likely to do so (cf. sentiment, hate speech lexicons)

Reproducibility: Ephemerality

- Evolution of
    - Platform access, affordances and recorded data
    - User behaviour (bc. of rules, affordances..) and composition
    - General social context / topics / trends
    - The construct and its measurement
    - Methods (especially ML/NLP)
- Data loss: deleted accounts / posts
- …

Reproducibility continued

- Accessibility/Transparency - not least for interdisciplinary audiences
- Reusability: Code available and runnable? Data available for all steps or only some / aggregated. Correctly described?
- Maintenance for datasets: Foreign keys, dependence on redownloading data (e.g., Tweet “rehydration”)
- Maintenance for methods: dependencies, specific environments
- …

Datasets

Suggested Ideas

1. **Reddit Posts and Comments about Politicians:** How do people discuss female politicians?
2. **Stance Detection Benchmark:** How well do current computational models perform at detecting stance towards different entities and topics?
3. **Annotator (Dis)agreement:** Characterizing differences in annotator perspective for subjective constructs
4. **Lost in Simplification? English vs. Simple Wikipedia:** How does content and framing diverge in Simplified Wikipedia?
5. **X (Twitter) and Reddit discussion about Football:** How do people talk about non-white players?
6. **One Day on Twitter**: What goes on in one day on Twitter?

Suggested Ideas

1. **Reddit Posts and Comments about Politicians:** How do people discuss female politicians?
2. **Stance Detection Benchmark:** How well do current computational models perform at detecting stance towards different entities and topics?
3. **Annotator (Dis)agreement:** Characterizing differences in annotator perspective for subjective constructs
4. **Lost in Simplification? English vs. Simple Wikipedia:** How does content and framing diverge in Simplified Wikipedia?
5. **X (Twitter) and Reddit discussion about Football:** How do people talk about non-white players?
6. **One Day on Twitter**: What goes on in one day on Twitter?

1. Reddit Posts and Comments about Politicians

Full data:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YWRXEP&version=1.0

Quantifying gender biases towards politicians on Reddit

Sara Marjanovic, Karolina Stańczak, Isabelle Augenstein

Published: October 26, 2022 • https://doi.org/10.1371/journal.pone.0274317

Article | Authors | Metrics | Comments | Media Coverage | Peer Review

Abstract

Despite attempts to increase gender parity in politics, global efforts have struggled to ensure equal female representation. This is likely tied to implicit gender biases against women in authority. In this work, we present a comprehensive study of gender biases that appear in online political discussion. To this end, we collect 10 million comments on Reddit in conversations *about* male and female politicians, which enables an exhaustive study of automatic gender bias detection. We address not only misogynistic language, but also other manifestations of bias, like benevolent sexism in the form of seemingly positive sentiment and dominance attributed to female politicians, or differences in descriptor attribution. Finally, we conduct a multi-faceted study of gender bias towards politicians investigating both linguistic and extra-linguistic cues. We assess 5 different types of gender bias, evaluating coverage, combinatorial, nominal, sentimental and lexical biases extant in social media language and discourse. Overall, we find that, contrary to previous research, coverage and sentiment biases suggest equal public interest in female politicians. Rather than overt hostile or benevolent sexism, the results of the nominal and lexical analyses suggest this interest is not as professional or respectful as that expressed about male politicians. Female politicians are often named by their first names and are described in relation to their body, clothing, or family; this is a treatment that is not similarly extended to men. On the now banned far-right subreddits, this disparity is greatest, though differences in gender biases still appear in the right and left-leaning subreddits. We release the curated dataset to the public for future studies.

[Quantifying gender biases towards politicians on Reddit]

Data Source

Reddit:

Between 2018-2020

Collected using PushShift

Cross-referenced with **Wikidata**: metadata about the politicians themselves (gender, etc)

S1 Table. Subreddits Included.

Subreddit	Number of comments	Partisan-affiliation
politics	9744853	—
The_Donald	1664335	alt-right
news	556783	—
neoliberal	340533	left
canada	285667	—
Libertarian	207109	right
Conservative	200772	right
unitedkingdom	197881	—
europe	158342	—
australia	107966	—
india	87367	—
democrats	53381	left
ireland	40964	—
teenagers	33311	—
newzealand	32847	—
socialism	18241	left
TwoXChromosomes	15734	—
MensRights	13664	—
Republican	13014	right
Liberal	10503	left
uspolitics	8873	—
SocialDemocracy	1977	left
alltheleft	837	left
feminisms	108	—

Augmenting the data

- Example of a comment: ‘*Clinton*’s emails were a red herring, it was a ploy to distract us from *her* scandals in Benghazi’

- “To identify the politician discussed in each post, a state-of-the-art lightweight entity linker is used to mark each comment with the associated wikidata ID.”

- “However, it should be noted that the correct female entity is only caught in 50% of the labelled cases.”

Can you think of potential issues in the dataset and how that might bias the analyses?

Methods in the paper

- Coverage of politicians of different genders (overall statistics, network analysis)
    - Mentions
    - Centrality

- Linguistic analysis
    - Naming and reference conventions
    - Sentiment (using the NRC **V**alence **A**rousal **D**ominance Lexicon)
    - Framing and topics

Things you could study with this data

- More comprehensive techniques for measuring positive stereotypes (benevolent sexism)
- Backlash effects
- Other types of sexism: stereotypes and frames
- Intersection sexism
- Sexism across political lines
- Affective polarization?

2. Stance Detection Benchmark

One popular stance dataset:

https://drive.google.com/drive/u/0/folders/1so8lY1XKpnhUtTvb15edEz6aeHt7CSuh

I am a CSS researcher who has tweets about **Donald Trump**. I want to measure the opinions towards him from these tweets. What do I do?

Defining the Measurement: Approval in Political Science

“Do you **approve** or **disapprove** of how X has done their job as the president?”

UPDATED OCT. 19, 2020 AT 1:03 PM

How popular is Donald Trump?

An updating calculation of the president's approval rating, accounting for each poll's quality, recency, sample size and partisan lean. How this works »

I am a CSS researcher who has tweets about **Donald Trump**. I want to *measure approval* towards him from these tweets. What do I do?

I am a CSS researcher who has tweets about **Donald Trump**. I want to *measure approval* towards him from these tweets. What do I do?

Use **off-the-shelf** (OTS) NLP methods to measure **positive, negative, or neutral** attitudes

RQ: Do current NLP methods <u>accurately</u> capture approval?

How do we measure **approval**?

O'Connor, B., Balasubramanyan, R., Routledge, B.R. and Smith, N.A., 2010, May. **From tweets to polls: Linking text sentiment to public opinion time series.** In *Fourth international AAAI conference on weblogs and social media.*

Conrad, F.G., Gagnon-Bartsch, J.A., Ferg, R.A., Schober, M.F., Pasek, J. and Hou, E., 2019. **Social media as an alternative to surveys of opinions about the economy.** Social Science Computer Review, p.0894439319875692.

hate	negative
honest	positive
inefficient	negative
love	positive
destroy	negative
encourage	positive
...	...

Are we Measuring approval towards Trump?

- Usually approval is defined as ‘sentiment’

Tweet	Untargeted Sentiment	Approval
Trump is the only candidate I fully support	positive	approval

Are we Measuring approval towards Trump?

- This sentiment is untargeted, therefore, if multiple entities are mentioned, untargeted sentiment =/= approval

Tweet	Untargeted Sentiment	Targeted Sentiment	Approval
What makes me angry is the media unnecessarily attacking Trump	negative	positive	approval

Are we Measuring approval towards Trump?

- Finally, targeted sentiment doesn’t work when entities are ***not directly*** mentioned

Tweet	Untargeted Sentiment	Targeted Sentiment	Stance	Approval
Jeb Bush is the best choice in the republican lineup	positive	none	against	disapproval

What is ‘Stance’ Detection?

- “the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target.” [Mohammad et al 2017]

- **Indirect stance:** “the target is referred to in indirect ways such as through pronouns, epithets, honorifics, and relationships.”

Stance as a Theoretical Proxy for Approval

Tweet	Untargeted Sentiment	Targeted Sentiment	Stance	Approval
Trump is the only candidate I fully support	positive	positive	favor	approval
What makes me angry is the media unnecessarily attacking Trump	negative	positive	favor	approval
Jeb Bush is the best choice in the republican lineup	positive	none	against	disapproval

RQ: Do current NLP methods <u>accurately</u> capture approval?

How? Measure the performance of **12 different methods** across a benchmark (**7 targets, 4 datasets**) of stance annotated by humans

"On the reliability and validity of detecting approval of political actors in tweets"

Methods

8 Off-the-shelf methods including:

Gilbert, C.H.E. and Hutto, E., 2014, **Vader: A parsimonious rule-based model for sentiment analysis of social media text.** ICWSM

VADER: Untargeted Sentiment

Augenstein, I., Rocktäschel, T., Vlachos, A. and Bontcheva, K., 2016. **Stance detection with bidirectional conditional encoding.** EMNLP

DSSD: Stance

Tang, D., Qin, B. and Liu, T., 2016. **Aspect level sentiment classification with deep memory network.** EMNLP

TD-LSTM: Targeted Sentiment

Methods

8 Off-the-shelf methods including:

Gilbert, C.H.E. and Hutto, E., 2014, **Vader: A parsimonious rule-based model for sentiment analysis of social media text.** ICWSM

VADER: Untargeted Sentiment

Augenstein, I., Rocktäschel, T., Vlachos, A. and Bontcheva, K., 2016. **Stance detection with bidirectional conditional encoding.** EMNLP

DSSD: Stance

Tang, D., Qin, B. and Liu, T., 2016. **Aspect level sentiment classification with deep memory network.** EMNLP

TD-LSTM: Targeted Sentiment

Trained on Trump data

Coverage of Test Sets

Unfamiliar Target	Familiar Target, Familiar Dataset	Familiar Target, Unfamiliar Dataset
Other Politicians	**Trump (SemEval)**	**Trump (Other)**

Results

Metric: Macro F1

Generalizability to Unseen Targets

Method	Other Politicians
VADER	43.9
TD-LSTM	36.3
DSSD	30.3

Generalizability to Seen Targets

Method	Other Politicians
VADER	43.9
TD-LSTM	36.3
DSSD	30.3

Generalizability to Seen Targets

Method	Other Politicians	Trump (SemEval)
VADER	43.9	38.0
TD-LSTM	36.3	38.8
DSSD	30.3	**60.6**↑↑

Generalizability to Seen Targets

Method	Other Politicians	Trump (SemEval)	Trump (Other)
VADER	43.9	38.0	35.1
TD-LSTM	36.3	38.8	34.4
DSSD	30.3	**60.6**↑↑	31.9 ↓↓

Takeaways

- **Stance** is a good theoretical proxy for approval, compared to targeted and untargeted sentiment

- But, current targeted methods like stance detection have room for improvement, even for **familiar targets**

How well do OTS models, especially LLM-based techniques identify stance towards targets?

Stance Detection Datasets

P-Stance: A Large Dataset for Stance Detection in Political Domain

Authors	Target(s)	Source	Type	Size
Mohammad et al. (2016a)	Atheism, Climate change is a real concern, Feminist movement, Hillary Clinton, Legalization of abortion, Donald Trump	Twitter	Target-specific	4,870
Ferreira and Vlachos (2016)	Various claims	News articles	Claim-based	2,595
Sobhani et al. (2017)	Trump-Clinton, Trump-Cruz, Clinton-Sanders	Twitter	Multi-target	4,455
Derczynski et al. (2017)	Various claims	Twitter	Claim-based	5,568
Swami et al. (2018)	Demonetisation in India in 2016	Twitter	Target-specific	3,545
Gorrell et al. (2019)	Various claims	Twitter, Reddit	Claim-based	8,574
Conforti et al. (2020b)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna, Disney-Fox	Twitter	Target-specific	51,284
Conforti et al. (2020a)	Merger of companies: Cigna-Express Scripts, Aetna-Humana, CVS-Aetna, Anthem-Cigna	News articles	Target-specific	3,291
P-STANCE	Donald Trump, Joe Biden, Bernie Sanders	Twitter	Target-specific	21,574

Table 2: Comparison of English stance detection datasets.

3. Annotator (Dis)agreement

Why is it important to study annotator disagreement?

- Annotating data is *interpretive*
- People's perceptions of constructs (especially toxicity, hate speech) is affected by their backgrounds
    - Demographics
    - Lived experience (e.g.., if they have faced harassment in the past or not)
- It is important our computational measurement models are 'representative'

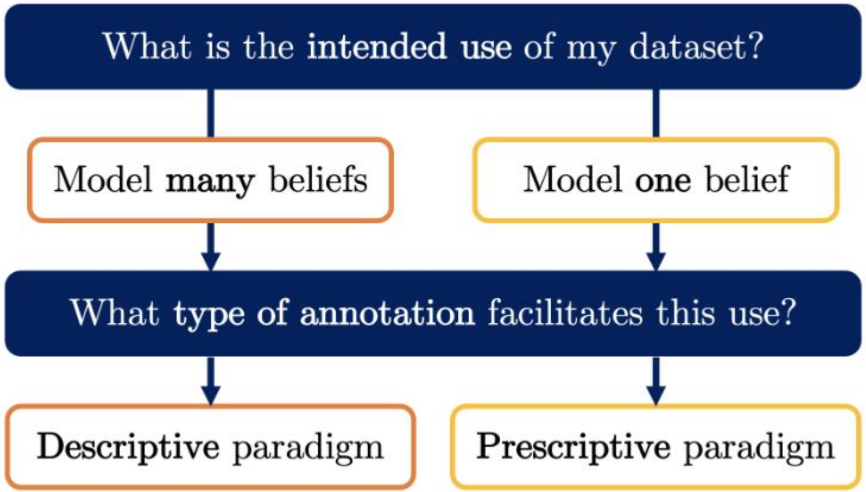


Figure 1: Two key questions for dataset creators.

[Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks]

We know that annotators perceive some constructs differently.

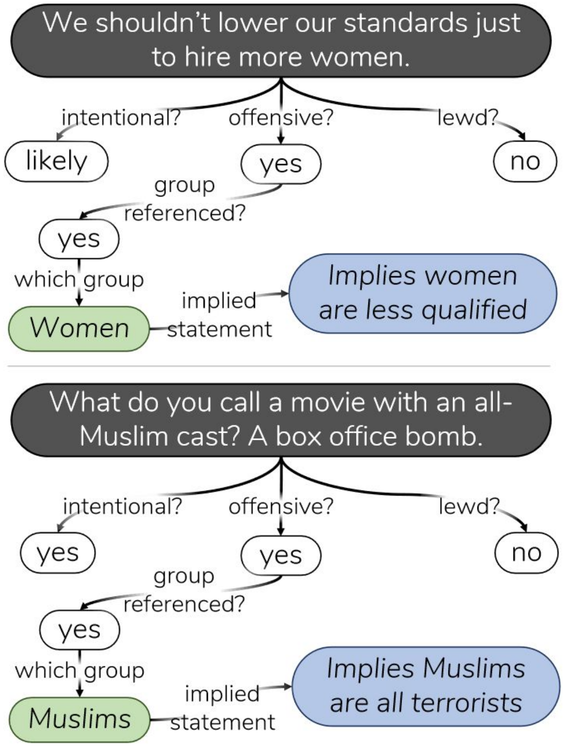
We know that annotators perceive some constructs differently.

But, what <u>characteristics</u> of the subjective instance drive this disagreement?

Datasets

Paper	Dataset	NLP Task	Which demographics?
Social Bias Frames	https://huggingface.co/datasets/social_bias_frames	Offensiveness, lewdness, sexual content	Gender, minority, political leaning,
Annotators with Attitudes	Contact authors	Toxicity (Anti-black toxicity)	Race, gender, political leaning, other beliefs (empathy, altruism, attitudes towards free speech, traditionalism)
NLPPositionality	http://nlpositionality.cs.washington.edu/	Social acceptability, hate speech	Gender, age, religion, country (residence, longest), education, ethnicity, native language
Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application	https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech	Hate speech	Age, disability, religion, sexuality, race, origin, gender
Designing Toxic Content Classification for a Diversity of Perspectives	https://data.esrg.stanford.edu/study/toxicity-perspectives (encrypted, need to contact authors)	Toxicity	Gender, age, race/ethnicity, LGBTQ+ status, Religion importance, political attitude, parental status
POPQUORN	https://github.com/Jiaxin-Pei/Potato-Prolific-Dataset/tree/main/dataset	Offensiveness, politeness, *email writing, question answering*	Gender, age, race, education
DICES Dataset: Diversity in Conversational AI Evaluation for Safety	https://github.com/google-research-datasets/dices-dataset/	Safety risk	Race, gender
Don't Take It Personally: Analyzing Gender and Age Differences in Ratings of Online Humor	No link to data	Humor and offense	Age, gender

Social Bias Frames



category_type	category	count	percentage
gender	woman	74337	51.98%
gender	man	68661	48.02%
race	white	115506	83.43%
race	hisp	8905	6.43%
race	asian	8597	6.21%
race	black	5444	3.93%
mixed	white man	57272	39.59%
mixed	white woman	58227	40.25%
mixed	black man	6	0.00%
mixed	black woman	5435	3.76%
mixed	asian man	5049	3.49%
mixed	asian woman	3548	2.45%
mixed	hisp man	3667	2.54%
mixed	hisp woman	5234	3.62%

Now, let's explore some of these datasets

- Download and open the notebook:
- Create one exploratory data visualization based on one or more datasets
- Examples
    - Who are the most frequently mentioned politicians in 1?
    - How many instances of indirect stance are there in 2?
    - Is there higher disagreement for the toxic/hateful class in 3?