“Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science”

Emily M. Bender, Batya Friedman

Student: Bonomi Silvia

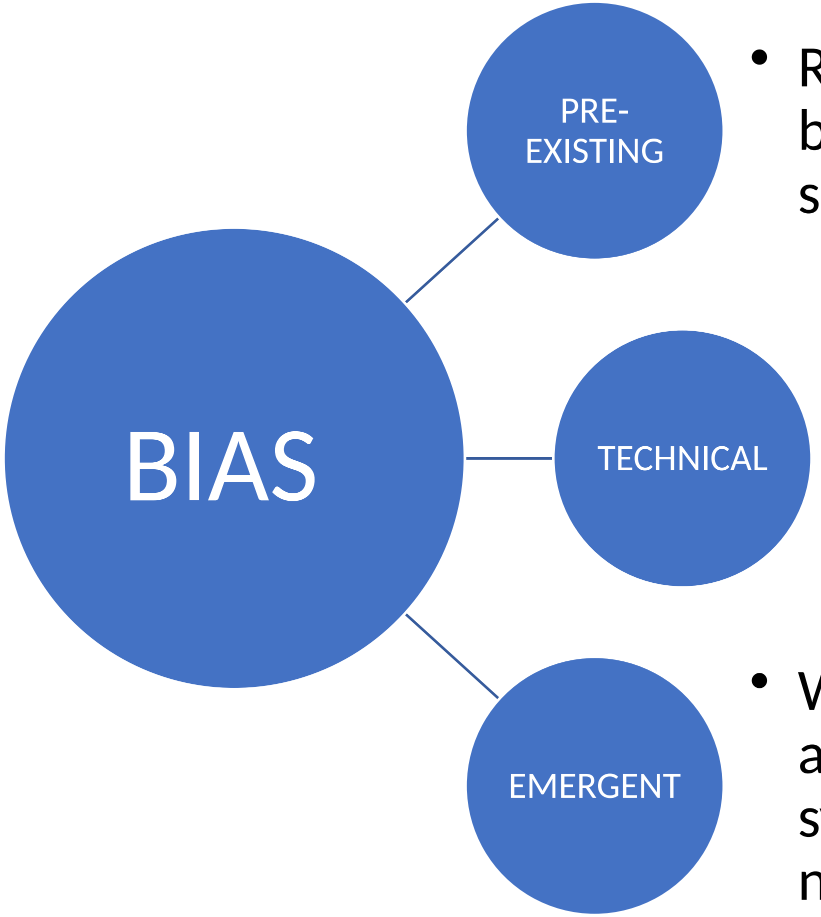
Outlines

1. What is a Bias?
2. What is a Data Statement?
3. Why does NLP need Data Statement?
4. Challenge
5. Data Statement: Long Form
6. Data Statement: Short Form
7. Availability of Information
8. Possible Costs
9. Application
10. Conclusion

What is a **Bias**?

Cases where computer systems «*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others»

(Friedman and Nissenbaum, 1996)



- Reflects biases in society

- Stem from technical constraints and decisions

- When we apply a system to a new context

What is a **Data Statement**?

- Characterization of a dataset that provides context to allow developers and users to:
  - better understand how experimental results might generalize
  - what biases might be reflected in systems

Why does NLP need Data Statements?

1. The linguistic data we use will always include pre-existing biases
2. It’s not possible to build an NLP system that is immune to emergent bias

- We must seek additional strategies for **mitigating** the scientific and ethical shortcomings that follows from imperfect dataset
- Highlighting the features of datasets it’s possible to reason about what the likely effects may be

Challenge

- Information Gap: filled by including data statements in every NLP publication
- **Detailed** and **Concise**:
    - **Long** Form
    - **Short** Form
- How to use them?
    - Long Form for academic papers presenting the datasets or in systems documentation
    - Short Form should be included in research papers using datasets with existing long form data statement

Data Statement: Long Form

1. Curation Rationale
2. Language Variety
3. Speaker Demographic
4. Annotator Demographic
5. Speech Situation
6. Text Characteristics
7. Recording Quality
8. Provenance Appendix
9. Other

Speaker: the individual who produced some segment of linguistic behaviour included in the dataset

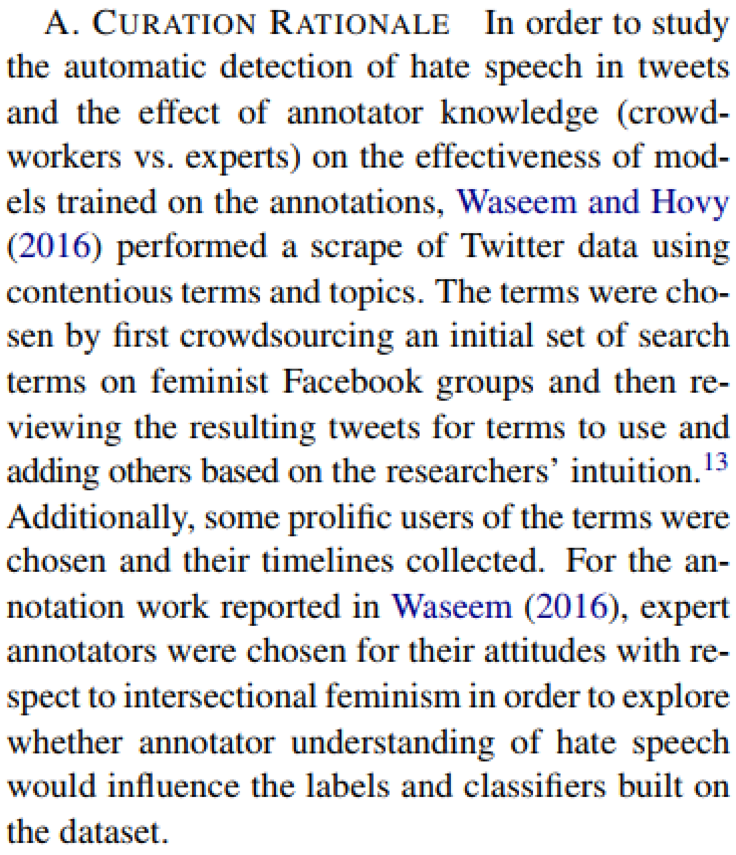
Annotator: who assign annotations to the raw data, including transcribers of spoken data (crowdworkers or highly trained researchers)

- The Hate Speech Twitter Annotations collection is a set of labels for ∼19,000 tweets collected by Waseem and Hovy (2016) and Waseem (2016)

Data Statement: Long Form

1. Curation Rationale

- Explanation of texts included and goals in selection
- Provides transparency for users to infer potential generalizations of systems trained with the dataset

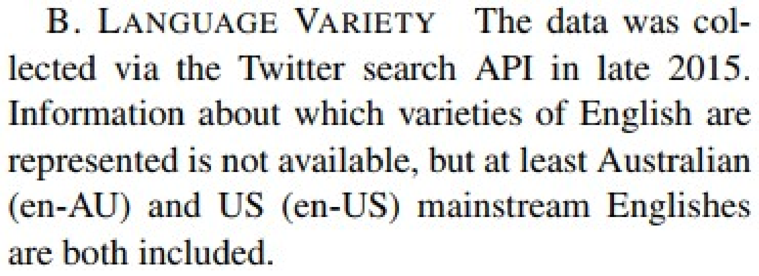


A. CURATION RATIONALE In order to study the automatic detection of hate speech in tweets and the effect of annotator knowledge (crowd-workers vs. experts) on the effectiveness of models trained on the annotations, Waseem and Hovy (2016) performed a scrape of Twitter data using contentious terms and topics. The terms were chosen by first crowdsourcing an initial set of search terms on feminist Facebook groups and then reviewing the resulting tweets for terms to use and adding others based on the researchers' intuition.[13] Additionally, some prolific users of the terms were chosen and their timelines collected. For the annotation work reported in Waseem (2016), expert annotators were chosen for their attitudes with respect to intersectional feminism in order to explore whether annotator understanding of hate speech would influence the labels and classifiers built on the dataset.

Data Statement: Long Form

2. Language Variety

- Since languages differ from each other in structural ways that can interact with the NLP system
- Requires:
  - Tag for language identification
  - Description of language variety

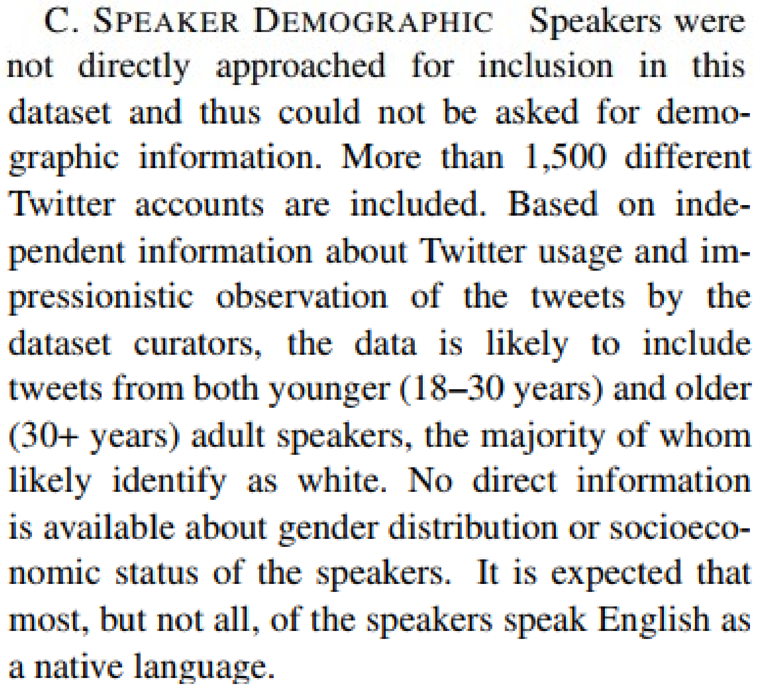


B. LANGUAGE VARIETY The data was collected via the Twitter search API in late 2015. Information about which varieties of English are represented is not available, but at least Australian (en-AU) and US (en-US) mainstream Englishes are both included.

Data Statement: Long Form

3. Speaker Demographic

- Variation in pronunciation, prosody, word choice, and grammar correlates with speaker demographic characteristics, reflecting identity construction
- Specifications include:
  - Age
  - Gender
  - Race/ethnicity
  - Native language
  - Socioeconomic status
  - Number of different speakers represented

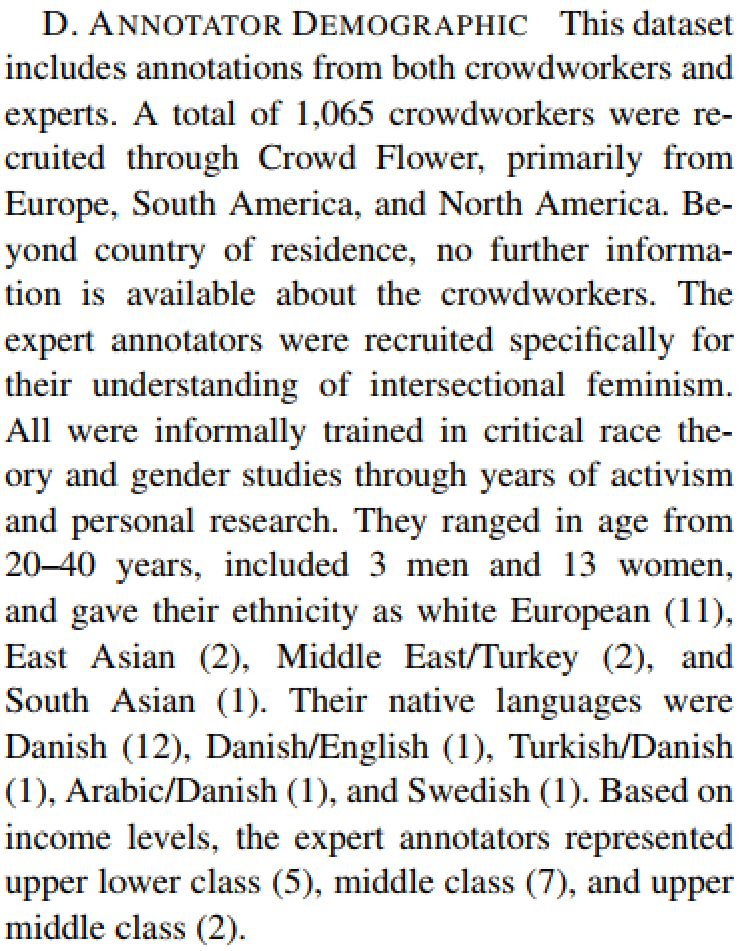


C. SPEAKER DEMOGRAPHIC Speakers were not directly approached for inclusion in this dataset and thus could not be asked for demographic information. More than 1,500 different Twitter accounts are included. Based on independent information about Twitter usage and impressionistic observation of the tweets by the dataset curators, the data is likely to include tweets from both younger (18–30 years) and older (30+ years) adult speakers, the majority of whom likely identify as white. No direct information is available about gender distribution or socioeconomic status of the speakers. It is expected that most, but not all, of the speakers speak English as a native language.

Data Statement: Long Form

4. Annotator Demographic

- Demographic characteristics of annotators and guideline developers, including their "social address," influence their language experience and perception during the annotation process
- Specifications include:
  • Age
  • Gender
  • Race/ethnicity
  • Native language
  • Socioeconomic status
  • Training in linguistics/other relevant discipline

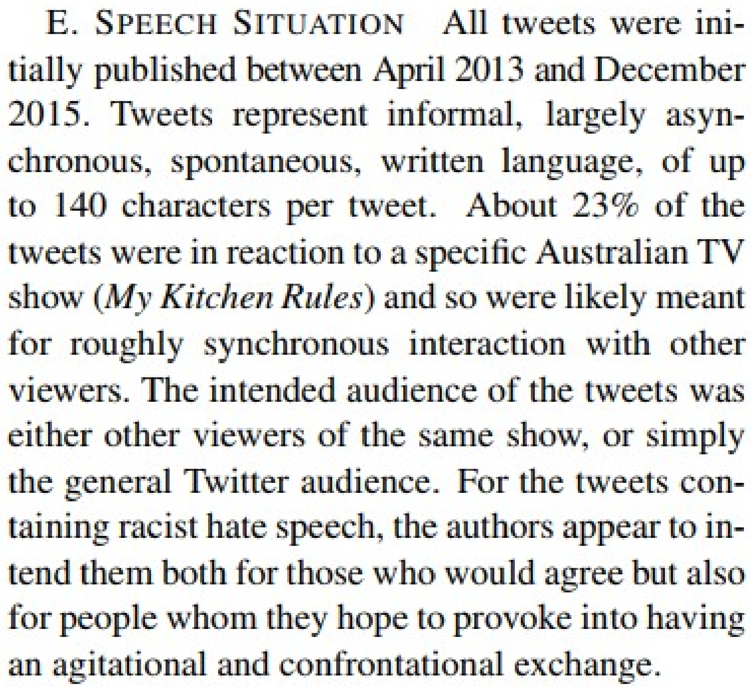


D. ANNOTATOR DEMOGRAPHIC This dataset includes annotations from both crowdworkers and experts. A total of 1,065 crowdworkers were recruited through Crowd Flower, primarily from Europe, South America, and North America. Beyond country of residence, no further information is available about the crowdworkers. The expert annotators were recruited specifically for their understanding of intersectional feminism. All were informally trained in critical race theory and gender studies through years of activism and personal research. They ranged in age from 20–40 years, included 3 men and 13 women, and gave their ethnicity as white European (11), East Asian (2), Middle East/Turkey (2), and South Asian (1). Their native languages were Danish (12), Danish/English (1), Turkish/Danish (1), Arabic/Danish (1), and Swedish (1). Based on income levels, the expert annotators represented upper lower class (5), middle class (7), and upper middle class (2).

Data Statement: Long Form

5. Speech Situation

- Elements that influence linguistic structure and patterns in collected texts:
  - Time
  - Place
  - Modality: scripted/spontaneous
  - Synchronous/Asynchronous interaction
  - Intended audience

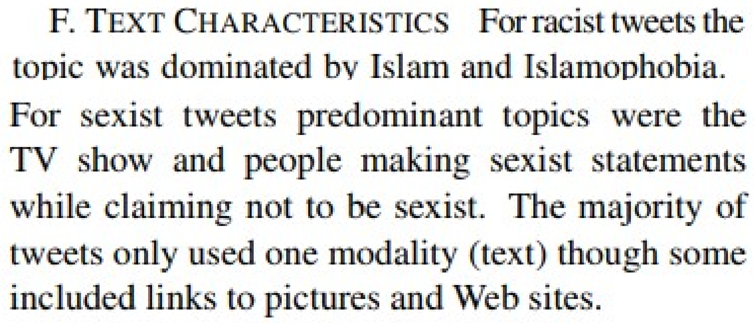


E. SPEECH SITUATION All tweets were initially published between April 2013 and December 2015. Tweets represent informal, largely asynchronous, spontaneous, written language, of up to 140 characters per tweet. About 23% of the tweets were in reaction to a specific Australian TV show (*My Kitchen Rules*) and so were likely meant for roughly synchronous interaction with other viewers. The intended audience of the tweets was either other viewers of the same show, or simply the general Twitter audience. For the tweets containing racist hate speech, the authors appear to intend them both for those who would agree but also for people whom they hope to provoke into having an agitational and confrontational exchange.

Data Statement: Long Form

6. Text Characteristics

- Genre and topic, as key influencers because they impact vocabulary and structural features of texts

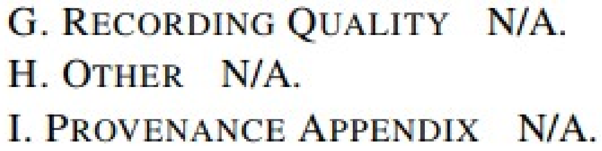


F. TEXT CHARACTERISTICS For racist tweets the topic was dominated by Islam and Islamophobia. For sexist tweets predominant topics were the TV show and people making sexist statements while claiming not to be sexist. The majority of tweets only used one modality (text) though some included links to pictures and Web sites.

Data Statement: Long Form

7. Recording Quality

- For datasets involving audiovisual recordings, the recording equipment's quality and relevant aspects of the recording situation are indicated



G. Recording Quality N/A.

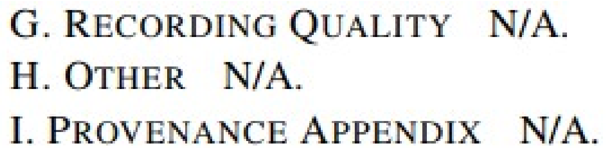
H. Other N/A.

I. Provenance Appendix N/A.

Data Statement: Long Form

8. Provenance Appendix

- For datasets derived from existing ones, the data statements of the source datasets should be included as an appendix to establish provenance



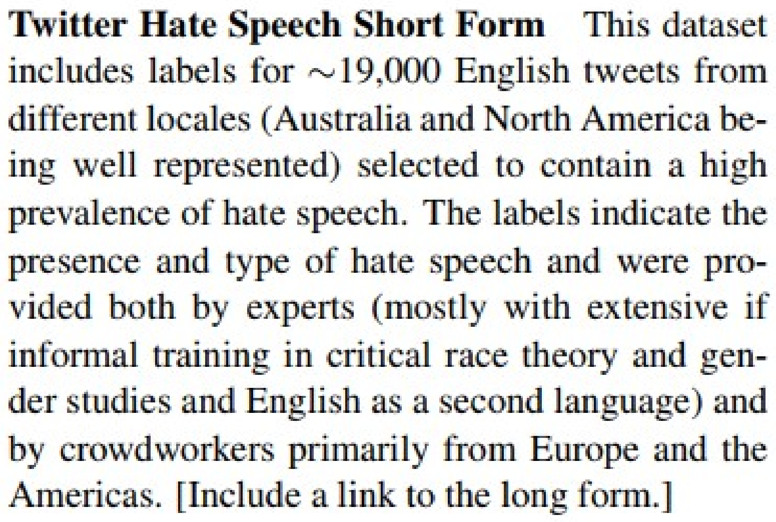
G. RECORDING QUALITY N/A.

H. OTHER N/A.

I. PROVENANCE APPENDIX N/A.

Data Statement: Short Form

- Facilitates accessibility to **essential dataset information** without overwhelming readers
- Do not replace long form ones: they serve as concise supplements
- Include a direct reference or pointer t the comprehensive long form version
- Envisioned as 60–100 word summaries, capturing key details from the long form description



Twitter Hate Speech Short Form This dataset includes labels for ~19,000 English tweets from different locales (Australia and North America being well represented) selected to contain a high prevalence of hate speech. The labels indicate the presence and type of hate speech and were provided both by experts (mostly with extensive if informal training in critical race theory and gender studies and English as a second language) and by crowdworkers primarily from Europe and the Americas. [Include a link to the long form.]

Availability of information

- Full specification of all these information <u>may not be feasible </u>in all cases
- Precise demographic information may be unavailable
- Sometimes it may be preferable to provide ranges rather than precise values (privacy of annotators)
- Nonetheless the explicit statement of lack of availability provides a more informative picture of the dataset

Possible Costs

FOR WRITERS

- Time:
  - 2–3 hours to write it
  - time for collecting information
  - institutional review board approval for exempt status, adding procedural time
- Space: propose exemption of data statements from page limits, addressing potential space constraints

FOR READERS

- Readers (reviewers and users) need to scrutinize data statements for appropriate dataset selection

Application of Data Statement

- Public Health and NLP for Social Media: how data statement could help
- Anticipating and Mitigating Barriers, potential negative outcome of data statements:
  - Desk rejection
  - Stop new dataset developments
- Possible solution: mentoring

Conclusion

- Starting point
- The impact of data statements cannot be fully demonstrated a priori
- Propose the formation of a working group that:
  - Develop best practices for data statement production, including steps before dataset collection and addressing privacy concerns
  - Create training materials: digital templates, conference tutorials, mentoring network, and an online guide
  - Establish standardized form after a few years of usage
- Potential Benefits:
  - Scientific community
  - Industry
  - Public at large

Thank You!