Addressing Bias: Algorithmic Fairness & Better Data Practices

Ethics in NLP: Including Society in Discourse and Design

February 28th, 2019

Ryan Georgi

UNIVERSITY OF WASHINGTON

PROFESSIONAL MASTER'S IN COMPUTATIONAL LINGUISTICS

Quick Thoughts on Fixing our Algorithms

- "Pipeline Problems"
  - Yao & Huang (2017) discuss fairness in recommendation systems
  - Sampling from existing models of gender representation in tech
  - Systems would be likely to under-recommend tech as a field to women, based on existing sampling

Addressing Bias in Data

- **Being aware of possible biases in data:**
  - Data Statements & Model Statements (Bender & Friedman, 2018; Mitchell et. al 2019)
- **Programmatically dealing with bias in the data**
  - Identifying bias
  - "de-biasing"
    - Alterations to the data set
    - Alterations to the training algorithms
    - Alterations to the algorithmic predictions

Some Metrics for Fairness

- Disparate Impact and the "Eighty Percent Rule" (via Feldman et al, 2014)
  - Recommended by the US EEOC
  - For data set $D = (X, Y, C)$
    - X = protected attribute
    - Y = remaining attributes
    - C = treatment (e.g. "Will be hired")

$$\frac{Predicted(C = YES \mid X = 0)}{Predicted(C = YES \mid X = 1)} \leq \tau = 0.8$$

Balanced Error Rate (BER)

- In addition to (overall) accuracy
- Average of the proportion of wrong classifications in each class
  - Smaller classes weight for more in this metric than overall accuracy

Balanced Error Rate (BER)

- Unfortunately, as Chouldechova (2016) points out:
  - *"…when the [target variable] prevalence differs between two groups, a test-fair score S_c cannot have equal false positive and negative rates across those groups."*

Algorithmic Fairness Methods

Algorithmic Fairness:
Preprocessing

- Disparate impact removal:
  - Take the same data set D
  - Modify to set $\bar{D} = (X, \bar{Y}, C)$
  - s.t. that any classification algorithm, predicted C' will not show disparate impact
- Note that this approach treats Y as a binary variable.

Algorithmic Fairness:
Algorithmic Modification

- Building a fairness measure into the cost function (regularization)
  - For Logistic Regression (Maximum Entropy) — (Kamishima et. al, 2012)
  - Enforce demographic parity

Better Data Practices

Data Statements (Bender & Friedman, 2018)

A. Curation Rationale
   - Why was the data gathered? (Relevant to HS protections)

B. Language Variety
   - What language, dialect, other focus?

C. Speaker Demographic

D. Annotator Demographic

E. Speech Situation
   - Was this a public speech? On the phone?

Data Statements (Bender & Friedman, 2018)

F. Text Characteristics
   - Genre, topic(s), vocabulary

G. Recording Quality

H. Other

I. Provenance Appendix

Example

- Hate Speech Twitter Annotation (Waseem and Hovy, 2016)
  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 researcher's intuition. 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.*

(Bender & Friedman, 2018, p. 6)

Example

- Hate Speech Twitter Annotation (Waseem and Hovy, 2016)

  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.*

(Bender & Friedman, 2018, p. 6)

Example

- Hate Speech Twitter Annotation (Waseem and Hovy, 2016)

  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.*

(Bender & Friedman, 2018, p. 6)

Key Thoughts

- You don't need to know everything about the speakers/language variety, etc
  - *But make an effort to state what you do and don't know!*

Example of Poor Practices

- Larson (2017) points out a number of cases related to gender:
  - Labeling BNC by gender, without discussing how labeling was done (Koppel et al 2002)
  - Labeling tweets (Rao et al 2010)
    - "*For gender, the seed set for the crawl came from initial sources including sororities, fraternities, and male and female hygiene products. This produced around 500 users in each class*"

Example of Poor Practices