Fairness, Bias, and Stereotypes: Metrics

9/12/2023

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Grading of Responses

- Full credit: response meets all expectations
- Half credit: response is relevant but does not fully meet expectations (doesn't reference all the readings reading, has little content, strictly summarizes without raising ideas or questions)
- No credit: response is not submitted, not relevant, or excessively short

Takeaways from readings

- Understanding some of the common fairness metrics and when they are used (equalized odds/opportunity, predictive parity, statistical parity, disparate impact)
- Difference between group and individual fairness
- Lack of compatibility between fairness metrics
  - Depending on data properties, it is impossible to satisfy every fairness criteria. A model that is ”fair” under one criteria may be unfair under another
  - Defining criteria depends on context and impact
- Some practice with readings and notation
- Some background on types of problems examined in fairness literature (recidivism prediction, targeted advertising, loan approval, admissions)

Discussion Time

Next readings

- Thursday: Classification (Prediction)
  1. Zhao et al. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP. 2017.
  2. Sap et al. "The Risk of Racial Bias in Hate Speech Detection", ACL. 2019.
  3. (optional) Field et al. "Examining risks of racial biases in NLP tools for child protective services" FAccT. 2023.
- Tuesday: Generation
  1. Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. ACL 2023.
  2. Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." FAccT 2023.

More examples of bias in classification we don’t have time for

- Co-reference resolution: NLP models tend to assume, for example, that “she” refers to “nurse” while “he” refers to “doctor”
  - Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL
  - Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. NAACL
- Machine translation: when translating into languages with grammatical gender, models assume doctors are male
  - Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. ACL
- Political orientation (and how biases in pre-training data perpetuate to downstream classification tasks)
  - Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. ACL

Intrinsic vs. Extrinsic Metrics of bias

Intrinsic: bias in internal model representations

Extrinsic: bias in downstream applications

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. [Man is to computer programmer as woman is to homemaker? debiasing word embeddings](). NeuRIPS

When is “de-biasing” intrinsic bias useful?

- “Debiasing” hides bias without actually removing it (it can be recovered) [Gonen and Goldberg. 2019. ACL]
- “We find that intrinsic and extrinsic metrics do not necessarily correlate in their original setting” [Cao et al. 2022 ACL]
- Even “extrinsic” bias metrics may not be measuring impact [Chouldechova, 2017]

Dimensions of Bias

- Large focus on gender and race
  - Reflects general overrepresentation of U.S. in research [Sambasivan et al. 2021. FAccT]
- Religion [Abid et al. 2021. Nature Machine Intelligence]
- Disability [Hutchinson et al. 2020. ACL]
- Political orientation
- Intersectionality
  - Race and gender [Jiang and Fellbaum. 2020. Workshop in Gender Bias in NLP]
  - Mental health and gender [Lin et al. 2022. EMNLP]

Higher-level: What is “Bias”

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. ACL
    - Work on “bias” often has poor engagement with external literature and metrics for measuring “bias” are often not aligned with claimed motivations
    - Conceptualization of “bias” needs to incorporate sociotechnical context
- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. ACL
    - Pitfalls in building benchmark data sets for “bias”

Next readings

- Thursday: Classification (Prediction)
  1. Zhao et al. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP. 2017.
  2. Sap et al. "The Risk of Racial Bias in Hate Speech Detection", ACL. 2019.
  3. (optional) Field et al. "Examining risks of racial biases in NLP tools for child protective services" FAccT. 2023.