**Language (Technology) is Power: A Critical Survey of “Bias” in NLP:**

Hej hej! My name’s Oreen Yousuf and I’m gonna talk to yall about a critical survey of analyses of “bias” in NLP systems. So you can think of it as an analysis on analyses.

**Background:**

Although papers analyzing “bias” in NLP systems have laid really vital groundwork by bringing some of the ways NLP systems can be harmful to the foreground, the majority of them fall short of engaging critically with what constitutes “bias” in the first place. Despite the fact that analyzing “bias” is an inherently normative process—in which some system behaviors are deemed good and others harmful—papers on “bias” in NLP systems are rife with unstated assumptions about what kinds of system behaviors are harmful, in what ways, to whom, and why.

Last note: And in this survey, researchers attempt to provide recommendations for practitioners and researchers moving forward after a critical analysis of analytical work done on “bias” in NLP systems.

**Dataset Compilation:**

* The survey under review compiled an extensive list of ”all papers known to [them] analyzing ”bias” in NLP systems – which totaled to 146 papers

\* One of my favorite parts was seeing how the researchers compiled a “complete” dataset that they intend to base their study on. I liked seeing their method to obtain what they deemed to be relevant papers.

* Works published before May of 2020 containing keywords of ”bias” and/or ”fairness” were taken from the ACL Anthology. And the ACL Anthology is this enormous digital archive of research papers that hosts something like a little more than 62,000 (62,486) papers on the study of computational linguistics and NLP.)
* The researchers discarded works not focused on social ”bias” and works discussing topics with other forms of bias (i.e., inductive bias, hypothesis bias)
  + Inductive bias (a.k.a. learning bias) is mainly for the assumption set for the user to further use to predict outputs of given inputs that is hasn’t encountered
  + Traversed citation graphs of each initially compiled paper to capture all relevantly cited papers within said papers to incorporate into the working list
* Included papers analyzing ”bias” in NLP systems from the biggest conferences and workshops, i.e., NeurIPS, AIES, ICML, etc., were also investigated, but were already found to be included
* And also it’s really important to note that sorting through and compiling relevant academic and industry literature was done with the stipulation of solely analyzing works conducted on written text, so they excluded research about speech.

**Taxonomy of Categorization:**

To categorize motivations and proposed quantitative techniques for measuring or mitigating “bias”, the authors of the critical survey made use of pre-existing taxonomy or nomenclature.

These are also known as “harms” – they isolate specific aspects of the greater “bias” label.

**Initial Findings:**

**Quantitative Findings:**

Works structured as surveys, frameworks, and meta-analyses of ”bias” in NLP systems more often than not provide motivations in their papers. Understandably state multiple motivations.

However, 33% of reviewed papers not structured this way also state multiple motivations

**Further Findings:**

1: These examples leave unstated what it might mean for an NLP system to “discriminate,” what constitutes “systematic biases,” or how NLP systems contribute to “social injustice” which itself is undefined

2: It was found that some papers (32%) are not motivated by any apparent normative concerns, often focusing instead on concerns about system performance. For example, the first quote below includes normative reasoning—namely that models should not use demographic information to make predictions—while the other focuses on learned correlations impairing system performance.

3: Found that even papers with clear motivations often fail to explain what kinds of system behaviors are harmful, in what ways, to whom, and why. These examples leave unstated what “problematic biases” or non-ideal user experiences might look like, how the system behaviors might result in these things, and who the relevant stakeholders or users might be.

**Techniques:**

* This is where the statistic of only 4 papers of the 21% of papers citing allocational harms actually proposing techniques for measuring or mitigating allocational harms
* Most papers focus on system predictions as the potential source of “bias”.
* Most papers don’t question the normative implications of other decisions made during the development and deployment lifecycle— perhaps unsurprising given that their motivations sometimes include no normative reasoning.
* Sap et al. (2019) focus on the effect of priming annotators with information about possible dialectal differences when asking them to apply toxicity labels to sample tweets, finding that annotators who are primed are significantly less likely to label tweets containing features associated with African-American English as offensive.

**Proposal:**

1 - R1 aids in creating a much more complete understanding of the unintended and consequential misrepresentations of some social groups by NLP systems are in and of themselves a dangerous harm to make

2 - I believe R2 is included as the most needed suggestion from an academic standpoint - the need to essentially state everything is required if a consistent and collective agreement on analyzing ”bias” in NLP is to come to fruition. It reduces the chance of papers with the same task from having conclusions at odds with each other and can assist in moving towards the goal of collective analytical agreement mentioned previously.

3 - R3’s desire of engagement with the social groups affected by NLP systems seeks to place those same groups at the core of this dreamt of collective agreement when analyzing ”bias” in NLP. To make this the center of your work may help propel advancements in how researchers understand the full effect of these systems.

**Conclusion:**

It would only benefit those authoring future works to take these recommendations seriously, not only for the insight gained into the technological aspects of their research, like the every focused on system performance, but also for the betterment their technology may bring from its impact on the world.

**References:**

**Question:**

How do social hierarchies and language ideologies infuence the decisions made during the development and deployment lifecycle? What kinds of NLP systems do these decisions result in, and what kinds do they foreclose?

General assumptions: To which linguistic norms do NLP systems adhere (Bender, 2019; Ruane et al., 2019)? Which language practices are implicitly assumed to be standard, ordinary, correct, or appropriate?

Task defnition: For which speakers are NLP systems (and NLP resources) developed? (See Joshi et al. (2020) for a discussion.) How do task defnitions discretize the world? For example, how are social groups delineated when defning demographic attribute prediction tasks (e.g., Koppel et al., 2002; Rosenthal and McKeown, 2011; Nguyen et al., 2013)? What about languages in native language prediction tasks (Tetreault et al., 2013)?

Data: How are datasets collected, preprocessed, and labeled or annotated? What are the impacts of annotation guidelines, annotator assumptions and perceptions (Olteanu et al., 2019; Sap et al., 2019; Geiger et al., 2020), and annotation aggregation processes (Pavlick and Kwiatkowski, 2019)?

Evaluation: How are NLP systems evaluated? What are the impacts of evaluation metrics (Olteanu et al., 2017)? Are any non-quantitative evaluations performed?