Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Oreen Yousuf

Natural Language Processing
Autumn 2020

1 Introduction

An ever growing body of research analyzing many facets of "bias" in natural language processing (NLP) systems have been born in recent years, as well as work produced to analyze similar biases in NLP systems created with different intents such as sentiment analysis, dialogue generation, and embedding spaces. Contributions to the understanding and further research of bias stemming from NLP systems from these works have been indispensable, revealing oversights during the development process/cycle of NLP systems, whether intended or not by the researchers. Even so, the majority of these papers go only as far as providing inadequate, unclear, or inconsistent motivations or proposed quantitative techniques for measuring or mitigating "bias." Moreover, they do not take part in, or engage with, the relevant literature corresponding to the thrust of their analyses of "bias" in NLP systems. There is a noticeable lack of collective agreement on the concept and definition of biases as well, leading to differing conclusions for almost indistinguishable abstracts between independent works. Recommendations are expounded upon within this critical survey of "bias"; the brunt of their inclusion rests on the more imperative acknowledgment of the relationship between language and social hierarchies, and urging practitioners and researchers to elucidate on their perception of "bias(es)."

2 Process of Critique

The survey under review compiled an extensive list of "all papers known to [them] analyzing "bias" in NLP systems - 146 papers in total."[1] Sorting through and compiling relevant academic and industry literature was done with the stipulation of solely analyzing works conducted on written text, excluding research about speech. Works published before May of 2020 containing keywords of

[1] Language (Technology) is Power: A Critical Survey of "Bias" in NLP - https://arxiv.org/pdf/2005.14050.pdf?

"bias" and/or "fairness" were taken from the ACL Anthology[2] and discarded works not focused on social "bias" and works discussing topics with other forms bias such as inductive bias or hypothesis bias. To guarantee there was not oversight in the initial search of relevant papers, the researchers traversed references in citation graphs of each paper and included all cited papers analyzing "bias." All papers analyzing "bias" in NLP systems from the biggest conferences and workshops, i.e., NeurUPS, AIES, ICML, etc., were also investigated, but were already found included in early steps of the researchers' compilation procedure. Already existing taxonomy of "harms" were used in this study to categorize the 146 papers under study. These descriptors distinguish between what are called *allocational* harms and *representational* harms, the former essentially being when automated systems reserve, or allocate, resources to one group over others, while the latter essentially being when those same automated systems unfairly generate, present, or represent a social group in a demeaning or less favorable light than others, and/or unbalanced in numbers. Further definitions of categories in which motivation and proposed techniques were to be distributed across were presented as the following:

(1) *Representational harms*:

(1a) *Stereotyping* that propagates negative generalizations about particular social groups.

(1b) Differences in *system performance* for different social groups, language that *misrepresents* the distribution of different social groups in the population, or language that is *denigrating* to particular social groups.

(2) *Questionable correlations* between system behavior and features of language that are typically associated with particular social groups.

(3) *Vague descriptions* of "bias" (or "gender bias" or "racial bias") or *no description* at all.

(4) *Surveys, frameworks, and meta-analyses.*

Table 1, below, shows where motivations and techniques fall within the above definitions of categorical harms:

[2]https://www.aclweb.org/anthology/

	Papers	
Category	Motivation	Technique
Allocational harms	30	4
Stereotyping	50	58
Other representational harms	52	43
Questionable correlations	47	42
Vague/unstated	23	0
Surveys, frameworks, and meta-analyses	20	20

Table 1: Breakdown of where the 146 papers under review fall into categorically.

The sums for motivation and technique do not total to 146 due to papers overlapping in their proposed motivational harms they wish to discuss. The same applies to techniques, with the addition of instances of papers also failing to provide a quantitative technique.

3 Discoveries

Unsurprisingly, something of note is that works structured as surveys, frameworks, and meta-analyses of "bias" in NLP systems more often than not provide motivations in their papers. They often leave very little unstated in the matter of who is harmed and which/how different social groups may go through dissimilar experiences with NLP systems than those kept in mind ("core demographic") in the development of them. The spectrum of concise and clear motivational works for other papers, however, is drastic. Works range from no motivations or vague motivations, to multiple motivations, of which none are without a majority of incompleteness.

"Other biases can be inappropriate and result in negative experiences for some groups of people. Examples include, loan eligibility and crime recidivism prediction systems...and resume sorting systems that believe that men are more qualified to be programmers than women (Bolukbasi et al., 2016). Similarly, sentiment and emotion analysis systems can also perpetuate and accentuate inappropriate human biases, e.g., systems that consider utterances from one race or gender to be less positive simply because of their race or gender, or customer support systems that prioritize a call from an angry male over a call from the equally angry female."[3]

- Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings

[3]Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems - https://arxiv.org/pdf/1805.04508.pdf

of the Joint Conference on Lexical and Computational Semantics, pages 43–53, New Orleans, LA.

The motivation behind this falls in allocational harms, and within ”other representational harms” - system performance differences in regards to text written by different social groups. However, the provided techniques are questionable considering they focus on the scoring of sentiment analysis *about* different social groups, instead of text written *by* them.

”However, embeddings trained on human-generated corpora have been demonstrated to inherit strong gender stereotypes that reflect social constructs...Such a bias substantially affects downstream applications...This concerns the practitioners who use the embedding model to build gender-sensitive applications such as a resume filtering system or a job recommendation system as the automated system may discriminate candidates based on their gender, as reflected by their name. Besides, biased embeddings may implicitly affect downstream applications used in our daily lives. For example, when searching for 'computer scientist' using a search engine...a search algorithm using an embedding model in the backbone tends to rank male scientists higher than females' [sic], hindering women from being recognized and further exacerbating the gender inequality in the community.”[4]

- Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and KaiWei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pages 4847–4853, Brussels, Belgium.

Although the motivation behind this work falls within three separate categories of harm - allocational, stereotyping, and ”other” - it only provides quantitative techniques for addressing stereotyping. Moreover, this paper falls prey to a very common attribute of works analyzing stereotyping - stating that there are ”downstream effects” caused by NLP systems. It is not to say that there are not, as there definitely are. It is that this is used as a way to give grounds for the analysis of system behaviors even when there is no measurement of these ”downstream effects.” This leaves incomplete work in the eyes of this survey, which urges thoroughness through standardized procedures of ”bias” analysis. Normative questions such as ”what kinds of system behaviors are harmful, in what ways, to whom, why” are presented as what should be the backbone for such analytical works.

The researchers consistently and confidently state that one of the more pervasive issues hindering the creation of a cohesive procedure of ”bias” analysis is when works about NLP systems developed for the same task often conceptualize ”bias” in different ways, leading to little continuity and greater inconsistency

[4]Learning Gender-Neutral Word Embeddings - https://arxiv.org/pdf/1809.01496.pdf

in how to approach the subject. Works on machine translation[5] reviewed by the survey have different definitions of "bias" in the same task and come to drastically differing techniques, while papers on type-level embeddings come to much more similar conclusions on proposed plans of rectification.[6]

4 Proposal

The researchers outlined three recommendations that would assist researchers and practitioners of "bias" analysis in NLP going forward, and would aid lowering the chances of hitting the same stumbling blocks as described previously in this paper. They are as follows:

- Recommendation 1 (R1): Ground work analyzing "bias" in NLP systems in the relevant literature outside of NLP that explores the relationships between language and social hierarchies. Treat representational harms as harmful in their own right.

- Recommendation 2 (R2): Provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom. Be forthright about the normative reasoning (Green, 2019) underlying these statements.

- Recommendation 3 (R3): Examine language use in practice by engaging with the lived experiences of members of communities affected by NLP systems. Interrogate and reimagine the power relations between technologists and such communities.

According to the researchers, 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. As language is the means by which all forms of communication occur, the spread of these misrepresentations allow for the continued oppression of these same groups. However, although social change can come from a change in language, obstacles stemming from a couple directions arise. The offense taken by the dominant social group to seemingly accommodate language change is high, as they wish to be the ones in control of what words mean. As is evident by how deep retaliation can be when attempting to alter language use in favor of a minority social group.[7] The second being what is being as for from practitioners and researchers - reorienting how you think about the analytical process you are accustomed to when it comes to "bias" in NLP - is substantial. 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

[5]see Table 3, row 3 and row 4 of the main survey)
[6]see Table 3, row 5 and row 6 of the main survey
[7]The Everyday Language of White Racism by Jane H. Hill

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. 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. A popular choice for case studies analyzing "bias" in NLP (including this survey by Blodgett et al. (2020) centers around African-American English (AAE). Blodgett et al. (2020) found that their recommendations (R1-R3) are helpful when pinpointing where and why the analysis procedure of several of these case studies fail to meet "proper" or "complete" thoroughness.

5 Conclusion

In this paper, several instances of subpar or incomplete analysis of "bias" in NLP were presented and elucidated upon. Motivations of works were often vague, insufficient, or incorrectly conflated, while proposed quantitative techniques were inadequately paired with their motivations. Blodgett et al. (2020) aimed to make practitioners and researchers of the field aware of suggestions birthed by consistent findings of well-meaning but comprehensively poor analysis. It would behove those authoring future works to take these recommendations seriously, not only for the insight gained into the technological aspects of their research, but also for the betterment they may aid in bringing due to their technology's impact on the world.

6 References

Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In Proceedings of the Joint Conference on Lexical and Computational Semantics, pages 43–53, New Orleans, LA.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and KaiWei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In Proceedings of Empirical Methods in Natural Language Processing (EMNLP), pages 4847–4853, Brussels, Belgium.

Jane H. Hill. 2008. The Everyday Language of White Racism. Wiley-Blackwell.