LANGUAGE (TECHNOLOGY) IS POWER: A CRITICAL SURVEY OF "BIAS" IN NLP

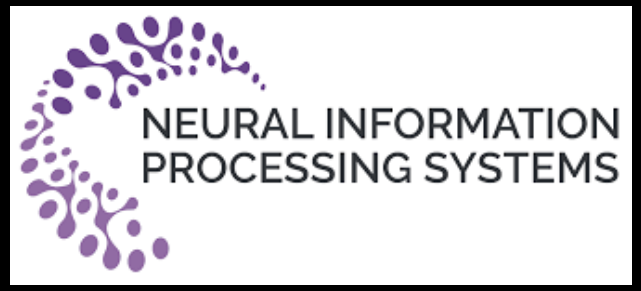
Oreen Yousuf

BACKGROUND

- Many papers in the growing body of research analyzing "bias" in natural language processing (NLP) in recent years provide inadequate, unclear or inconsistent motivations or proposed quantitative techniques for measuring or mitigating "bias."
- Techniques often don't match their motivations or engage with relevant literature.
- Lack of agreement on concept and definition of "bias(es)" leading to inconsistent conclusions for similar abstracts.
- Provide recommendations on analytical works on "bias" in NLP systems to move forward

DATASET COMPILATION

- 146 papers analyzing “bias” in NLP systems
- Papers with keywords of ”bias” and/or ”fairness” were taken from the ACL Anthology.
- Discarded works not focused on social ”bias” and works discussing topics with other forms of bias (i.e., inductive bias, hypothesis bias)
- Traversed citation graphs to fully capture relevant papers
- Papers analyzing ”bias” in NLP systems from NeurUPS, AIES, ICML, etc.
- Focused on written text; excluded research about speech.



AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE,
ETHICS, AND SOCIETY

TAXONOMY OF CATEGORIZATION

- Representational harms:
  - a) Stereotyping that propagates negative generalizations about particular social groups.
  - b) 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.
- Questionable correlations between system behavior and features of language that are typically associated with particular social groups.
- Vague descriptions of “bias” (or “gender bias” or “racial bias”) or no description at all.
- Surveys, frameworks, and meta-analyses

INITIAL FINDINGS

	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

- Tallies of either Motivation or Technique do not equate to 146 due to papers overlapping proposed motivational harms.
- Same applies to techniques, with the addition of instances of papers also failing to provide quantitative techniques.

QUANTITATIVE FINDINGS

- Papers' motivations span every harm category, with numerous papers matching every one.
- Surveys, frameworks, and meta-analyses had much higher percentage of providing motivations.
  - Common to state multiple motivations.
  - However, 33% of reviewed papers not structured this way also state multiple motivations
- 16% of papers state vague or no motivation at all.
- 32% of papers are not motivated by any apparent normative concern, opting instead to highlight concerns for system performance.
- Only 4 papers proposed techniques for allocational harms of the 21% of papers that specified said harms in their motivations.

FURTHER FINDINGS

Vague motivations/no motivations

"[N]o human should be discriminated on the basis of demographic attributes by an NLP system." - Kaneko and Bollegala (2019)

"[P]rominent word embeddings [...] encode systematic biases against women and black people [...] implicating many NLP systems in scaling up social injustice." — May et al. (2019)

No normative reasoning

"In [text classification], models are expected to make predictions with the semantic information rather than with the demographic group identity information (e.g., 'gay', 'black') contained in the sentences."
— Zhang et al. (2020a)

"An over-prevalence of some gendered forms in the training data leads to translations with identifiable errors. Translations are better for sentences involving men and for sentences containing stereotypical gender roles."
— Saunders and Byrne (2020)

Unstated information

"Deploying these word embedding algorithms in practice, for example in automated translation systems or as hiring aids, runs the serious risk of perpetuating problematic biases in important societal contexts." —Brunet et al. (2019)

"[I]f the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected." —Liu et al. (2019)

TECHNIQUES

- Don’t engage with relevant literature outside of NLP
  - Notable exception – papers on stereotyping
- Focus on narrow range of potential sources of “bias”
  - System predictions
  - “Bias” in datasets
- Small percentage questioning decisions made during development lifecycle
- Positive Example: Sap et al. (2019)
  - Analyzed effect of priming annotators on dialectal differences for toxicity labeling on African-American English (AAE)

PROPOSAL

- Recommendation 1 – 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 – 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 – 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.

CONCLUSION

- 146 papers analyzing "bias" in NLP systems:
  - Several subpar or incomplete analyses:
    - Motivations are often vague, inconsistent, lacking normative reasoning
    - Quantitative techniques poorly match their motivations
  - Recommendations made by Blodgett et al. (2020) aimed to help practitioners and researchers avoid consistent pitfalls

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.