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

Mohammad Arvan

October 19th, 2020

Department of Computer Science
University of Illinois at Chicago Illinois

Introduction

Introduction

"bias" in

- Embedding spaces
- Language modeling
- Coreference resolution
- Machine translation
- Sentiment analysis
- Hate speech/toxicity detection

Pitfalls

- Fail to engage critically with what constitutes “bias”
- Unstated assumptions about what kinds of system behaviors are harmful in what ways, to whom, and why

An Example for the Imprecise Terminology

“racial bias”:

- embedding spaces in which embeddings for names associated with African Americans are closer to unpleasant words than pleasant words
- sentiment analysis systems yielding different intensity scores for sentences containing names associated with African Americans and sentences containing names associated with European Americans
- toxicity detection systems scoring tweets containing features associated with African-American English as more offensive than tweets without these features

Result of their Analysis

- motivations are often vague and inconsistent
- lack any normative reasoning for why the system behaviors that are described as “bias” are harmful, in what ways, and to whom
- do not engage with the relevant literature outside of NLP to ground normative concerns when proposing quantitative techniques for measuring or mitigating “bias”
- techniques are poorly matched to their motivations, and are not comparable to one another

Recommendations

- examine the relationships between language and social hierarchies
- articulate their conceptualizations of “bias” in order to enable conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why
- deeper engagements between technologists and communities affected by NLP systems

Method

Which Papers?

- All papers with the keywords “bias” or “fairness” in ACL Anthology that were made available prior to May 2020
- Traversed the citation graph of our initial set of papers, retaining any papers analyzing “bias” in NLP systems that are cited by or cite the papers in our initial set
- Papers analyzing “bias” in NLP systems from leading conferences and workshops

Papers Group by Tasks

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Taxonomy

Built upon the work of (Barocas et al., 2017; Crawford, 2017).

Allocational harms
Arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups

Representational harms
Arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether

Taxonomy (cont.)

- Allocational harms
- Representational harms
- Questionable correlations
- Vague descriptions
- Surveys, frameworks, and meta-analyses

Papers Categorized by Taxonomy

	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

Findings

Identified Commonalities

- Motivation
- Technique

Multiple, Vague, or No Motivation at All

- 33%: multiple
- 16%: vague, or none

Examples of Vague Motivations

Kaneko and Bollegala (2019)
No human should be discriminated on the basis of demographic attributes by an NLP system.

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

No Normative Reasoning

Zhang et al. (2020a)
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.

Saunders and Byrne (2020)
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.

Lack of Details

Brunet et al. (2019)
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.

Liu et al. (2019)
If the systems show discriminatory behaviors in the interactions, the user experience will be adversely affected.

Inconsistent Bias Conceptualization

- Same task, different conceptualizations of “bias,” leading to different proposed techniques
- Same task, different motivation, but same proposed techniques

Conflate Allocational and Representational Harms

Name immediate representational harms, alongside more distant allocational harms in imagined as downstream effects of stereotypes

Techniques

Papers' techniques are not well grounded in the relevant literature outside of NLP exception the papers on stereotyping.

- Word Embedding Association Test from Implicit Association Test from the social psychology literature
- "Angry Black Woman" stereotype "double bind" faced by women from Black feminist scholarship on intersectionality

Poorly Matched to their Motivation

21% of the papers include allocational harms but only four papers actually propose techniques for measuring or mitigating it

Narrowed Focus

Papers focus on a narrow range of potential sources of “bias.”

Mostly limited to system predictions and “bias” in datasets

A Path Forward

Three Recommendations

- Ground work in the relevant literature outside of NLP
- Provide explicit statements for their arguments
- Examine language use in practice by engaging with affected communities

Language and social hierarchies

- sociolinguistics, linguistic anthropology, sociology, and social psychology
- Group labels can serve as the basis of stereotypes and thus reinforce social inequalities
- many have sought to bring about social changes through changes in language, disrupting patterns of oppression and marginalization via so-called “gender-fair” language

Language Attitudes and Ideologies

- Which language varieties or practices are taken as standard, ordinary, or unmarked?
- Which are rendered invisible?

The Role of Language

in maintaining social hierarchies

Important Question

How are social hierarchies, language ideologies, and NLP systems coproduced?

Future Work

Concrete Research Questions

Future Work

Concrete Research Questions

Future Work

Concrete Research Questions

Future Work

Concrete Research Questions

Conceptualizations of “bias”

Provide explicit statements of why the system behaviors that are described as “bias” are harmful, in what ways, and to whom, as well as the normative reasoning underlying these statements.

should take into account the relationships between language and social hierarchies

Future Work

Concrete Research Questions

Language use in practice

- center work analyzing “bias” in NLP systems around the lived experiences of members of communities affected by these systems
- power relations between technologists and such communities be interrogated and reimagined.

Examples

- work on language reclamation to support decolonization and tribal sovereignty
- work in sociolinguistics focus- ing on developing co-equal research relationships with community members and supporting linguistic justice efforts

Future Work

Concrete Research Questions

Case Study

A Case Study

Covering work on African-American English (AAE) in part-of-speech taggers and toxicity detection

Main Focus

Not only on system performance

Issues

None of these papers engage with the literature on AAE, racial hierarchies in the U.S., and raciolinguistic ideologies

Viewed as “bad”

Not considered consumers who matter

Thank you for your attention!

Questions?