Bias at a Second Glance: A Deep Dive into Bias for German Educational Peer-Review Data Modeling

THE 29TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS

Thiemo Wambsganss[1] ,Vinitra Swamy[1], Roman Rietsche[2] and Tanja Käser[1]

[1]EPFL, Lausanne, CH
[2]University of St.Gallen, Sankt Gallen, CH

Abstract

Natural Language Processing (NLP) has become increasingly utilized to provide adaptivity in educational applications. However, recent research has highlighted a variety of biases in pre-trained language models. While existing studies investigate bias in different domains, they are limited in addressing fine-grained analysis on educational and multilingual corpora.

In this work, we analyze bias across text and through multiple architectures on a corpus of 9,165 German peer-reviews collected from university students over five years. Notably, our corpus includes labels such as helpfulness, quality, and critical aspect ratings from the peer-review recipient as well as demographic attributes. We conduct a Word Embedding Association Test (WEAT) analysis on (1) our collected corpus in connection with the clustered labels, (2) the most common pre-trained German language models (T5, BERT, and GPT-2) and GloVe embeddings, and (3) the language models after fine-tuning on our collected data-set. In contrast to our initial expectations, we found that our collected corpus does not reveal many biases in the co-occurrence analysis or in the GloVe embeddings. However, the pre-trained German language models find substantial conceptual, racial, and gender bias and have significant changes in bias across conceptual and racial axes during fine-tuning on the peer-review data. With our research, we aim to contribute to the fourth UN sustainability goal (quality education) with a novel dataset, an understanding of biases in natural language education data, and the potential harms of not counteracting biases in language models for educational tasks.

Problem and Solution Suggestion

Problem:

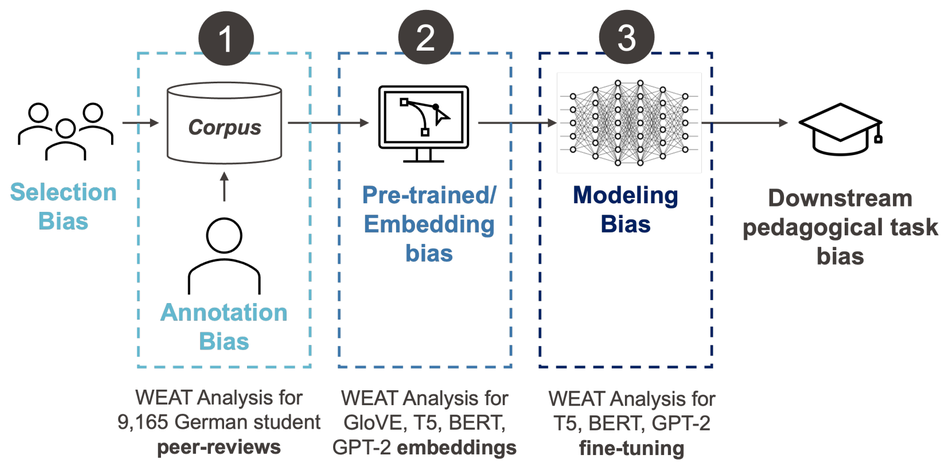
- Natural Language Processing (NLP) has become increasingly utilized to provide adaptivity in educational applications.
- However, recent research has highlighted a variety of biases in pre-trained language models.

Research Objective :

- Investigating bias along the entire NLP pipeline of an exemplary to see how bias evolves in student-written text
- Bias analysis through multiple architectures on a corpus of 9,165 German peer-reviews collected from our university students over five years.

Subsets	High (>= 6)	Low (< 6)
Total	9165	
Constructive Suggestions	5656	3509
Critical Aspects	5514	3651
Helpfulness	5886	3279
Quality	5391	3774

Table 2: Distribution of student peer-reviews in each reviewer rating subset. The scores have been rated by the receiver of the review. High and Low ratings correspond to scores on the Likert 7-point scale.

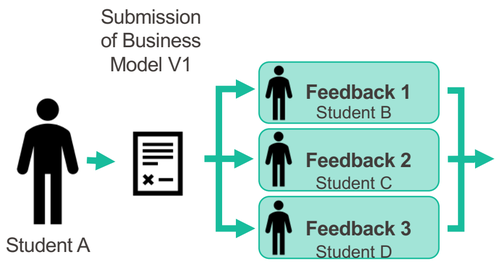


Methodology

Data

Overview of data collection from 2015 to 2019

Data collection of peer-review corpus



Subsets	High (>= 6)	Low (< 6)
Total	9165	
Constructive Suggestions	5656	3509
Critical Aspects	5514	3651
Helpfulness	5886	3279
Quality	5391	3774

Table 2: Distribution of student peer-reviews in each reviewer rating subset. The scores have been rated by the receiver of the review. High and Low ratings correspond to scores on the Likert 7-point scale.

Students elaborate on the strengths and weaknesses of a business model and give recommendations

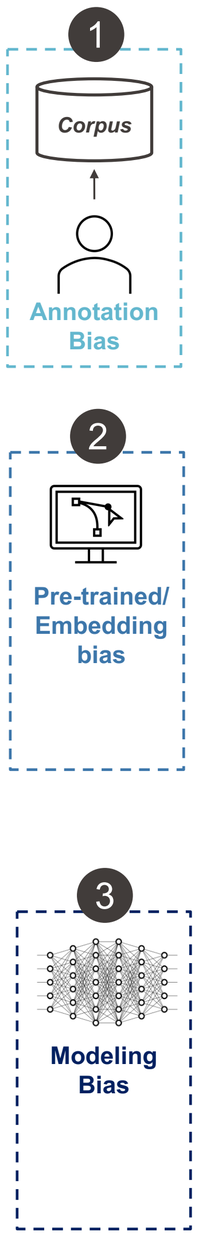
Feedback receiver rates the helpfulness of the received reviews (1-7 Likert scale).

Word Embedding Association Test (WEAT)

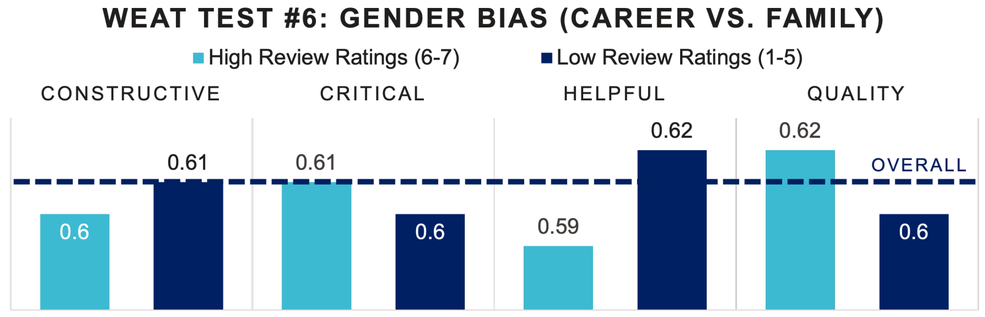
Bias	#	Targets	Attributes
Conceptual	1	Flowers vs. Insects	Pleasant vs. Unpleasant
	2	Instruments vs. Weapons	Pleasant vs. Unpleasant
	9	Mental vs. Physical Disease	Temporary vs. Permanent
Racial	3	Native vs. Foreign Names	Pleasant vs. Unpleasant
	4	Native vs. Foreign Names (v2)	Pleasant vs. Unpleasant
	5	Native vs. Foreign Names (v2)	Pleasant vs. Unpleasant (v2)
Gender	6	Male vs. Female Names	Career vs. Family
	7	Math vs. Arts	Male vs. Female Terms
	8	Science vs. Arts	Male vs. Female Terms

Table 1: Overview of our proposed measured bias categories (conceptual, race, and gender) for the WEAT analysis. WEAT compares the association between two different target word lists (i.e. Math vs. Arts) to attribute word lists (i.e. Male vs. Female terms). # indicates the original WEAT test number (Caliskan et al., 2017).

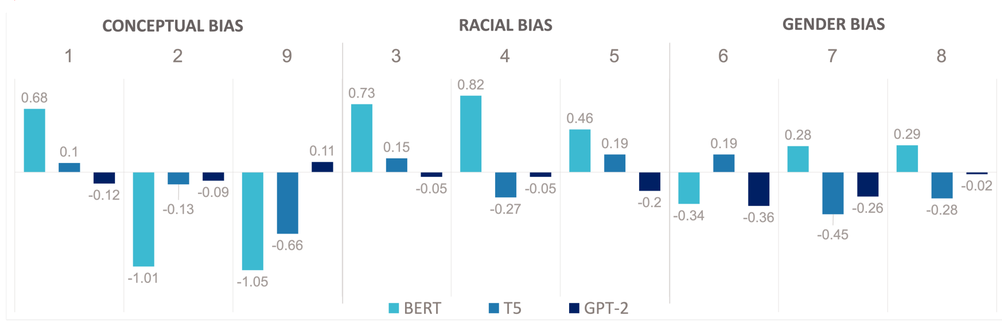
Results



We do not find significant results across any of the nine WEAT tests, with only six co-occurrences identified in total across 9,165 peer-reviews.



Overall, only one test on the gender axis is able to uncover bias using traditional word embeddings (GloVe).



Pre-trained German language models are inherently significantly biased, and fine-tuning using language models uncovers different, significant bias. BERT is the most susceptible to changes in bias of the three architectures.

Machine Learning for Education Lab
https://www.epfl.ch/labs/ml4ed/




Thiemo Wambsganss, Vinitra Swamy, Roman Rietsche, Tanja Käser

More information/ contact:
thiemo.wambsganss@epfl.ch

Access our code on GitHub:
https://github.com/epfl-ml4ed/bias-at-a-second-glance