Does Double-Hard Debias Keep its Promises? - A Replication Study

Kristina Kobrock[1], Meike Korsten[1], & Sonja Börgerding[1]

[1] University of Osnabrück

Author Note

The authors made the following contributions. Kristina Kobrock: müssen wir natürlich nicht ausfüllen; Meike Korsten: ...; Sonja Börgerding: ....

Does Double-Hard Debias Keep its Promises? - A Replication Study

Just edit the text! There are several websites that help with (r)markdown, e.g. https://rmarkdown.rstudio.com/. If you don't want to edit directly, but rather comment use: Cite using: Wang et al. (2020) said that… or simply blablabla (Wang et al., 2020). References must be added to r-references.bib file.

Introduction

- shall we include an abstract?

Bias in Embeddings.

- some text on bias in embeddings and why it's useful to do research in that field
- summary of related approaches (or put it in the discussion ?)

Motivation.

- rough description of the specific study (Wang et al.) & motivation of replication attempt (maybe some general remarks on the need for replication studies)

Implementation

- task description
- explanation of model and training choices

Datasets and Preliminaries.

Hard Debias.

Double-Hard Debias.

Evaluation

- stick to the paper

Baselines.

Evaluation of Debiasing Performance.

Debiasing in Downstream Applications.

Coreference Resolution.

- no replication possible, no code provided

Debiasing at Embedding Level.

The Word Embeddings Association Test (WEAT).

Neighborhood Metric.

- discuss that this is shady in the discussion part

Analysis of Retaining Word Semantics

Word Analogy.

Concept Categorization.

Discussion

- analysis of results and evaluation of performance evaluation
- ablation studies (not applicable)
- discuss the results and what could be (partly) replicated and what not

Conclusion

References

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V., & Xiong, C. (2020). Double-hard debias: Tailoring word embeddings for gender bias mitigation. *Association for computational linguistics (acl).*