Responsible NLP Checklist

Paper title: *Social Bias in Multilingual Language Models: A Survey*
Authors: *Lance Calvin Lim Gamboa, Yue Feng, Mark G. Lee*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

□ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
The survey paper has minimal risks, which we discuss in our Limitations section as well.

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B1. Did you cite the creators of artifacts you used?
We examine several multilingual benchmarks across Sections 3 to 6, where we also make appropriate citations to these benchmarks' developers.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We analyze open-source benchmarks that have been released by the NLP research community in open platforms (e.g., ACL Anthology, GitHub). These resources are academic in nature and are available for use in research contexts. Although we do not explicitly discuss the specific license details of every artifact in the paper due to space considerations, we include links to the datasets and their respective license details in the annotation files we will release with the paper.

N/A B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We do not use the multilingual bias benchmarks in any computational analysis. Rather, our survey takes a step back and examines the benchmarks for linguistic diversity, cultural awareness, and bias conceptualization and operationalization.

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Sections 4 and 5 discuss methods multilingual bias benchmarks developers take to create bias benchmarks and make sure their contents, which are somtimes potentially offensive, are culturally appropriate.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sections 3 to 6 and Appendix C provides a comprehensive documentation of the artifacts we examined.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Sections 3 to 6 and Appendix C provides a comprehensive documentation of the artifacts we examined.

☒ **C. Did you run computational experiments?**

N/A C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We ran no computational experiments.

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We ran no computational experiments.

N/A C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We ran no computational experiments.

N/A C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
We ran no computational experiments.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Only the authors worked on examining the articles and benchmarks the survey analyzes.

N/A D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Only the authors worked on examining the articles and benchmarks the survey analyzes.

N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Only the authors worked on examining the articles and benchmarks the survey analyzes.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Only the authors worked on examining the articles and benchmarks the survey analyzes.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Only the authors worked on examining the articles and benchmarks the survey analyzes.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
We did not use AI assistants.