Responsible NLP Checklist

Paper title: *IndoSafety: Culturally Grounded Safety for LLMs in Indonesian Languages*
Authors: *Muhammad Falensi Azmi, Muhammad Dehan Al Kautsar, Alfan Farizki Wicaksono, Fajri Koto*

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?
Ethics Statement

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

☑ B1. Did you cite the creators of artifacts you used?
Section 3 and 4. Our dataset extends previous works.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The artifact used is publicly available under the Apache 2.0 license. While its licensing terms were verified during our work, they were not explicitly stated in our paper.

☒ 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)?
The artifact used is publicly available under the Apache 2.0 license. While its licensing terms were verified during our work, they were not explicitly stated in our paper.

☒ 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?
We acknowledge that our data contains inappropriate content. Therefore, we put a potential risk of misuse in the section Ethical Statement.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4 Dataset Creation. Our dataset comprises several languages, such as Indonesian, colloquial Indonesian, Javanese, Sundanese, and Minangkabau.

☑ 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?
4 Dataset Creation. It explains about how many instances we have in the dataset.

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.

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

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In section 5 (Experimental Set-Up), we used several LLMs and generate outputs with those. However, we didn't document the total computation budget because we use various resources that hard to track.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In section 5 (Experimental Set-Up), we explain how we conduct the experiment. Further details are explained in the Appendix J

☒ 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 only report brief descriptive statistics in Section 6 (Results and Analysis), but we didnt explicitly state that the results are from a single run. Due to time constraints, we typically run each experiment once.

☑ 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?
In Appendix J (Training Setup), we explained about the parameter settings for LoRA. In Appendix E, we explained that most of the models we used were sourced from Hugging Face.

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

☒ 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.?
We did not report the full text of instructions. However, we report the important secti

☑ 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)?
In P

☒ 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)?
No. Formal written consent was not explained in the paper, but annotators were informed that their work would be used for research purposes. The data does not include any personal or identifiable information about the annotators.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our work involved simple annotation tasks and did not collect any personal or sensitive data from annotators. As such, the data collection protocol did not require ethics board approval.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix D.

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

☑ E1. If you used AI assistants, did you include information about their use?
In section 4 (Dataset Creation) and 5 (Experimental Set-Up). We used LLMs for dataset creation and experiments.