Responsible NLP Checklist

Paper title: *NeighXLM: Enhancing Cross-Lingual Transfer in Low-Resource Languages via Neighbor-Augmented Contrastive Pretraining*
Authors: *Sicheng Wang, Wenyi Wu, Zibo Zhang*

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?
Our work does not include a dedicated section on potential risks, as it does not involve any user-generated, sensitive, or private data. Nonetheless, we acknowledge that pretrained multilingual models may still reflect underlying biases in their training corpora, especially for low-resource languages.

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

☑ B1. Did you cite the creators of artifacts you used?
1,2,3,4

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We did not include license information in the paper, as we only used publicly available datasets and pretrained models under established open-source licenses

☒ 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 did not explicitly discuss intended use in the paper, but all artifacts we used (e.g., Tatoeba, MasakhaNEWS, SQuAD, InfoXLM) are publicly available for academic research purposes, and our usage remained consistent with those purposes. We did not use any data with commercial restrictions, and all outputs were generated in an academic research context only.

☒ 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 did not explicitly discuss this aspect in the paper because all datasets used in our study (e.g., Tatoeba, MasakhaNEWS, KenSwQuAD, SQuAD) are publicly available and curated for academic use. These datasets are widely adopted in the NLP community and do not contain any known personally identifiable information or offensive content.

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.?
4

☑ 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

☑ **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?
4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☒ 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 report results from a single run for each configuration. Due to resource constraints, we did not conduct multiple runs or report variance. In future work, we plan to include more robust statistical analysis.

☒ 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 did not include a detailed breakdown of package implementations and parameter settings. For sentence similarity and neighbor mining, we used cosine similarity over mean pooled embeddings using standard PyTorch operations. Evaluation metrics such as F1 and exact match were computed using standard scripts consistent with prior work. No non-default or package-specific parameters were used.

☒ **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.?
N/A

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)?
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)?
We did not explicitly discuss consent because all datasets used in our work (e.g., Tatoeba, MasakhaNEWS, KenSwQuAD, SQuAD) are publicly available and were released for academic research purposes. We did not collect or curate any data from individuals ourselves.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
N/A

☑ **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?

We used ChatGPT to improve the fluency and clarity of the papers English phrasing, such as making the writing more natural and accurate. All technical content, including methods, experiments, and analysis, was written and verified by the human authors. The use of AI tools was limited to surface-level language revision and did not contribute to the scientific content.