Responsible NLP Checklist

Paper title: *VN-MTEB: Vietnamese Massive Text Embedding Benchmark*
Authors: *Loc Pham, Tung Luu, Thu Vo, Minh Nguyen, Viet Hoang*

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?
There is no risk

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

☑ B1. Did you cite the creators of artifacts you used?
I already cited the artifacts used in this research in the appendix

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
I already cited the artifacts used in this research, also their licenses in the appendix

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)?
(left blank)

☒ 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?
There is no PII or Offensive Content

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
I already cited the artifacts used in this research, their licenses and documentation in the appendix

☑ 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?
I already reported relevant statistics of all 41 datasets in the appendix

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?
I already reported the number of parameters in the models used, the total computational budget in the appendix

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
I already discussed the experimental setup in the appendix

☑ 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?
I already reported descriptive statistics about your results in the appendix

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?
(left blank)

☑ **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.?
(left blank)

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)?
(left blank)

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)?
(left blank)

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

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

☒ **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?
I didn't use AI Assistant in progress of write the paper