Responsible NLP Checklist

Paper title: *GASE: Generatively Augmented Sentence Encoding*
Authors: *Manuel Frank, Haithem Afli*

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.

N/A A2. Did you discuss any potential risks of your work?
We do not see any risk related to this work.

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

☑ B1. Did you cite the creators of artifacts you used?
2, Appendix D, Appendix E, Appendix F

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Data-wise we use a standard benchmarking dataset (MTEB). Model-wise we only run inference using standard ("off-the-shelf") models. No models have been trained or shared and no datasets have been generated that are being shared.

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)?
Data-wise we use a standard benchmarking dataset (MTEB). Model-wise we only run inference using standard ("off-the-shelf") models. No models have been trained or shared and no datasets have been generated that are being shared.

☒ 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?
Our work is based on a standard benchmarking dataset (MTEB).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2, For languages: Table 17 (part 1) and Table 17 (part 2)

☒ 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?

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.

We used a commonly-used benchmark dataset (MTEB) for which only the test splits are made fully available and provided references for all included datasets which provide the necessary details.

☑ **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?
The hardware used is listed in Appendix F, the model-provider APIs are listed in Appendix F too, and a runtime analysis provided in Appendix H. Beyond that, this was not in scope of our work and we used a standard ("off-the-shelf") computational setup which is not computationally intense.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
2, Appendix B, Appendix C, Appendix D, Appendix F

☑ 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?
3, Appendix G

☑ 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 report the exact model names in Table 1 (embedding models) and Appendix F (generative models). The Sentence Transformers package is referenced in Appendix D. The plattform for the non-API generative models (Ollama) is provided in Appendix F. Where not explicitly mentioned, we used default (hyper)parameter settings. The authors are happy to provide any additional information on request via e-mail to any interested reader..

☒ **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?
Acknowledgments