Responsible NLP Checklist

Paper title: *TR-MTEB: A Comprehensive Benchmark and Embedding Model Suite for Turkish Sentence Representations*
Authors: *Mehmet Selman Baysan, Tunga Gungor*

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?
We discuss potential risks and limitations in Section 6. In particular, we highlight risks related to the weak supervision of the training corpus, which may contain noisy or biased pairs, and note that the monolingual nature of the models may limit performance in cross-lingual tasks such as bitext mining. Additionally, we note that no exhaustive filtering was applied for personally identifiable information (PII) or offensive content due to the large scale of the dataset.

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

☑ B1. Did you cite the creators of artifacts you used?
We used a large number of publicly available datasets hosted on Hugging Face. All datasets used in the construction of the sentence pair corpus and TR-MTEB benchmark are cited in Appendix A and D, with detailed references provided in the bibliography.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Appendix D provides Hugging Face links to each dataset, most of which are publicly released under open-source or research-friendly licenses. We ensured compliance with terms of use, and no proprietary or restricted datasets were included in our training corpus. Licenses for artifacts are discussed in Appendix A and D

☑ 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)?
As discussed in Section A and D in the appendix, all datasets were used for research purposes in accordance with their original license conditions and access terms. Our use is consistent with intended purposes, and the derived corpus and models are released for research use 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

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.

taken to protect/anonymize it?
We discuss data filtering and quality control in Section 3.2. While the datasets were sourced from public repositories, we performed basic filtering to remove malformed or low-quality instances. However, no exhaustive PII or toxicity screening was performed due to the scale of the corpus. This is mentioned as a limitation in Section 6.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Documentation of the datasets, including schema, domain types, and data statistics, is provided in Appendix D and throughout Section 3.2. TR-MTEB benchmark datasets are documented in Section 3.1 and Appendix A.

☑ 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?
We report dataset statistics in tables inside Appendix D and other benchmark-specific tables in Appendix A, as well as in the corpus description in Section 3.2

☑ **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?
Section 4 discusses model initialization, architecture (BERTurk), and training budget. GPU usage (Google Colab A100, 80 hours for pretraining, 2 hours for fine-tuning) is also reported there.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.3 and Appendix E describes the training procedure, including loss functions, batch size (32,768), optimizer details, and learning setup (contrastive pretraining and supervised fine-tuning).

☑ 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?
Results are presented in Table 2, reporting mean scores across task types and task categories. As this is a benchmarking paper, we report performance from single runs per model following MTEB protocols.

☑ 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 used the Hugging Face Datasets and Transformers library and sentence-transformers for training and Turkish adapted version of MTEB for evaluation mentioned in Section 4, with hyperparameters and loss functions explained in Section 3.

☑ **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.?
Instructions for manual annotation during translation evaluation are given in Appendix B, including the PASS/FAIL schema and prompt calibration process.

☒ 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)?
Manual annotations were conducted by the authors without external crowdworkers or payments. This is not applicable to external annotator recruitment.

☒ 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)?
All datasets were sourced from public Hugging Face repositories under open licenses. No direct interaction or consent from data subjects was required.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new user data was collected. All datasets were sourced from previously published and publicly released resources. D5 Characteristics Of Annotators: No

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Manual annotations were performed by the authors. No demographic information was collected or relevant.

☑ **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?
AI assistants were used only for minor text editing and language refinement purposes. No factual content, analysis, or model-related information was generated or sourced from AI tools. All technical contributions, experimental results, and claims in the paper are entirely original and authored by the researchers.