Responsible NLP Checklist

Paper title: *DisastIR: A Comprehensive Information Retrieval Benchmark for Disaster Management*
Authors: *Kai Yin, Xiangjue Dong, Chengkai Liu, Lipai Huang, Yiming Xiao, Zhewei LIU, Ali Mostafavi, James Caverlee*

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 present the potential risks in Ethics Statement part

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

☑ B1. Did you cite the creators of artifacts you used?
We give detailed license and HuggingFace link of all baseline models in Appendix H

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We mention the use of developed benchmark in Ethics Statement. And we have publicly released benchmark at https://huggingface.co/datasets/KaiYinTAMU/DisastIR

☑ 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)?
For existing artifacts (baseline models and open-source datasets), we give details of their uses in Appendix E and Appendix H. For artifact we created (DisastIR benchmark) we specified its use in Ethics Statement.

☑ 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 provide it in Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We mainly provide the benchmark in this work and we describe the analyses of the developed benchmark in Section 4

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.

☑ 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 provide detailed analyses of developed benchmark in Section 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?
We detail the selected baseline models and the model implementation in Section 5.1 and Appendix H. The cost for using GPT-4o-mini is detailed in Appendix D.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We give model implementation of all baseline models in Appendix H

☑ 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?
Evaluation results are provided in Section 6

☑ 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 A, we provide the package for web crawling and PDF extraction. In Appendix H, information of package needed for model implementation is provided.

☑ **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 provide details of instructions for human query generation in Appendix G

☑ 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)?
The information of human participant is provided in Appendix G

☑ 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 data are publicly available and how to construct the benchmark is detailed in Section 3 and Appendix A

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study uses publicly available datasets without any personally identifiable or sensitive information, so ethics review board approval was not required.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
As we are mostly focused in disaster management area. The background knowledge of disaster management is more important than the demographic and geographic characteristics of human annotators.

☑ **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 use ChatGPT API for user query generation and relevance labeling. The reliability of query generation and relevance labeling is further validated with human annotators . We have detailed their use in benchmark development in Section 3