Responsible NLP Checklist

Paper title: *Learning Trajectories of Figurative Language for Pre-Trained Language Models*
Authors: *Nicola Arici, Luca Putelli, Ejdis Gjinika, Ivan Serina, Alfonso Gerevini*

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?
Yes, in the Ethics and Impact Statement section.

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

☑ B1. Did you cite the creators of artifacts you used?
Yes. We provided references for the datasets we employed in Section 3 and for the models we considered in Section 5.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes, in Section 3 and Appendix A.2.

☑ 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)?
Yes, in Section 3 and Appendix A.2.

☒ 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?
For the figures of speech datasets, the data considered in our study are simple sentences taken from benchmarks on figure of speech and, therefore, they dont contain names or identifiers for individuals, and neither offensive content. Moreover, we used the standard Wikipedia and Project Gutenberg corpora released by Dolma (ACL 2024), which do not contain personal information or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. Our figures of speech datasets contain examples of figurative and literal language (metaphors, hyperboles, pleonasms and oxymorons) in English. Wikipedia and Project Gutenberg are both in English. All this information is available in Sections 3 and 5.

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?
Yes, in Section 3 and 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?
Yes. The number of parameters of each model is reported in Section 5 and Appendix B. The computing infrastructure used in our experiments and the relative GPU hours are reported in the Ethics and Impact Statement section.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Section 4.1 and 5.

☑ 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?
Our results are obtained in a single run. However, the metrics we considered (the compression obtained by the MDL method) are stable to multiple seeds. This information is reported in Section 5.

☑ 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?
Yes, in Appendix B.

☑ **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.?
Our annotation process was limited only to the validation of a set of generated sentences, in order to check whether they contained an oxymoron or not. This process was done by two of the authors manually, without any formal set of instructions.

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)?
Our annotation process was done by two of the authors (a PhD student and a post-doc researcher), without any payment. This information is available in Appendix A.

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 use any data regarding individuals.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We did not use any data regarding individuals.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Yes, in Appendix 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?
Yes, we used the GPT-4 model through the official OpenAI API for generating a dataset. More details about this procedure, which has been validated by humans, is reported in Section 3 and, in a more detailed way, in Appendix A.1.