Responsible NLP Checklist

Paper title: *The Relevance of Value Systems for Offensive Language Detection*
Authors: *Michael Wiegand, Elisabeth Eder, Josef Ruppenhofer*

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?
t is difficult to anticipate risks that our work may yield.

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

☑ B1. Did you cite the creators of artifacts you used?
We use many existing resources for our work (i.e. datasets and tools). We always cite the creators of a specific resource (either by the related research publication or URL) when we first mention it in our paper.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We only release data that were produced from scratch, i.e. we do not release data that are part of previous datasets. The inference of protected categories is not possible with the information that we provide. An extensive discussion can be found in the data sheet we provide as part of the supplementary material.

☑ 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)?
Among the cross-dataset classifiers we use in our experiments , we include a state-of-the-art classifier for general offensive language detection, i.e. PerspectiveAPI (https://perspectiveapi.com/), and the most recent transformer specially fine-tuned for implicitly abusive language detection, i.e. HateBERT fine-tuned on ToxiGen (Hartvigsen et al., 2022). These plug-and-play classifiers (i.e. they do not require any training) were specifically designed for the above tasks. We, therefore think that these classifiers are also meant to be applicable for our specific classification scenario, i.e. a specific subset of implicitly abusive language. Other existing datasets we used in our experiments were utilized in the way in which they were intended. We specified the use of the data we created as part of our research in the data sheet we provided as part of the supplementary material.

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.

☑ 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 discuss this in Section 7 (Ethical Considerations). Our dataset does not include the names of individuals other than public figures.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 2 (Data) we describe how our datasets have been built in detail. The same section and also Appendix D include demographic information on the crowdworkers who contributed to our datasets.

☑ 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 some descriptive statistics on our new datasets in Tables 2-4 and Table 8. Information on the set-up of supervised classifiers (i.e. training and test splits) is provided in Section 4 and Appendix A.

☑ **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 provide some of this information in Appendix A.1 (Computing Infrastructure and Running Time). Our work does not introduce new deep-learning models or deep-learning architectures. Existing classifiers/architectures were used in their standard configuration, which is also documented in Appendix A (Hyperparameters of Statistical Models). The number of parameters is specified in Appendix A.4 (DeBERTa).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Essential information on the set-up of supervised classifiers is provided in Section 3. Information on hyperparameters is stated in Appendix A (Hyperparameters of Statistical Models).

☑ 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?
Most of the classifiers we used are based on deep learning. Whenever their output is non-deterministic, we always report the average over 5 training runs. For these classifiers, we also report standard deviation. We also clarify this in Section 3.1, first paragraph (The Different Sentence Classifiers).

☑ 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?
For the classifiers based on deep learning, we use FLAIR (https://github.com/flairNLP/flair). For logistic regression, we used LIBLINEAR (Fan et al., 2008). We always use the default configuration of the tools. This information is also provided in Section 3.1, first paragraph (The Different Sentence Classifiers) and Appendix A (Hyperparameters of Statistical Models).

☑ **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 all instructions to all taks (i.e. annotation guidelines) as part of the supplementary material.

☑ 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)?
In Section 2 (Data), we mention the crowdsourcing platform for our tasks, i.e. Prolific (www.prolific.com). In Section 7 (Ethical Considerations), we also state that we compensated the crowdworkers following the wage recommended by Prolific (i.e. $12 per hour).

☑ 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 made it clear in our instructions that the tasks the crowdworkers were to perform were part of some linguistic research. We refrained from detailing the actual classification task and classification approach that we were about to pursue with the data to be created since this might have biased the crowdworkers in their annotation.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The legal department of the institution of the authors of this submission has been informed about the type of research that is conducted (i.e. data collection of abusive content).

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
In Section 2 (Data), we specify that all crowdworkers engaged in this study were required to be native speakers of English. Additionally, these individuals were expected to self-identify with one of the following groups: practicing Christians, conservatives, environmentally-conscious individuals, or liberals. To ensure a uniform cultural context, which is crucial given the varying interpretations of concepts such as liberalism or conservatism in different countries, we exclusively recruited crowdworkers residing in the United States of America. Detailed information regarding our recruitment process is thoroughly documented in Appendix D (How Crowdworkers were Recruited).

☑ **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 only used ChatGPT for rephrasing individual sentences.