Responsible NLP Checklist

Paper title: *K-LegalDeID: A Benchmark Dataset and KLUEBERT-CRF for De-identification in Korean Court Judgments*
Authors: *Wooseok Choi, Hyungbin Kim, Yon Dohn Chung*

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?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4.1 Datasets, and section 4.1.1 Data Collection for Thunder-DeID Dataset \Section 4.1.1 Data Collection for raw SNS multi-turn conversation dataset \Section 4.2 KLUEBERT-CRF for KLUEBERT model \Section 4.2 KLUEBERT-CRF for Conditional Random Field \Section 5.1 Experiment Setting for Thunder-DeID Model

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 4.1.1 Data Collection for Thunder-DeID Dataset License \Section 4.1.1 Data Collection for raw SNS multi-turn conversation dataset License \Section 4.2 KLUEBERT-CRF for KLUEBERT model License

☑ 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)?
Section 3.2 Conditional Random Field and section 4.2 KLUEBERT-CRF for Conditional Random Field \Section 4.1.1 Data Collection for Thunder-DeID Dataset License \Section 4.1.1 Data Collection for raw SNS multi-turn conversation dataset License \Section 4.2 KLUEBERT-CRF for KLUEBERT model \Section 4.2 KLUEBERT-CRF for our KLUEBERT-CRF model

☑ 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?
In section 4.1.2 and section 4.1.3, we discussed the methodology for anonymizing personal identifiable information (PII) in raw data and convert PII to annotation symbol.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.1 KLUEBERT-CRF for KLUEBERT model \Section 3.2 KLUEBERT-CRF for Conditional Random Field \Section 4.1.1 Data Collection for Thunder-DeID Dataset \Section 4.1.1 Data Collection for raw SNS multi-turn conversation dataset \Section 5.1 Experiment Setting for Thunder-DeID Model

☑ 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?
Section 4.1.1 Data Collection for raw SNS multi-turn conversation dataset statistics \Section 4.1.5 Court Judgement PII Dataset for court judgement dataset statistics \Section 4.1 Datasets, and section 4.1.1 Data Collection for Thunder-DeID dataset statistics \Section 5.1 Experiment Setting for details of train/validation/test splits.

☑ **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?
In Appendix I, table 11 shows details.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Appendix I, table 11 shows details.

☑ 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?
Section 5.2 Main result \Section 5.3 Robustness to Unseen Data \Section 5.4 Comparative Analysis

☑ 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 F. we reported package that we used (KSS).

☑ **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 mainly considered Supreme Court Trial Regulation No. 1778 as instruction, which is publicly available. So, we just mentioned document name in section 4.1.2. So, we provide only the regulation reference in Section 4.1.2 without reporting the complete regulatory text.

☑ 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 Appendix G, we reported it.

☑ 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)?
In Appendix G, we reported it.

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?
We used it to analyze debugging logs when errors occurred in the training code.