12 Reproducibility Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1 and Section 5

   (b) Did you describe the limitations of your work? [Yes] See Section 5

   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The study does not have any direct potential negative effect.

   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? https://automl.cc/ethics-accessibility/ [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All assumptions are described in the paper as well as the detail in Appendix section.

   (b) Did you include complete proofs of all theoretical results? [Yes] Explained in Section 3 and 4

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results, including all requirements (e.g., `requirements.txt` with explicit version), an instructive `README` with installation, and execution commands (either in the supplemental material or as a URL)? [Yes] Added all required hyper-parameters, seeds, link to data sets, link to publicly available MLMs; see Section 4, Appendix A to C, and GitHub repository files.

   (b) Did you include the raw results of running the given instructions on the given code and data? [Yes] Each experiment configurations saved as a JSON file that includes all parameters, raw data sets or encoded dataset and it is available in GitHub Repository.

   (c) Did you include scripts and commands that can be used to generate the figures and tables in your paper based on the raw results of the code, data, and instructions given? [Yes] We added all detailed as JSON and CSV file in GitHub repository.

   (d) Did you ensure sufficient code quality such that your code can be safely executed and the code is properly documented? [Yes] We explained the detail in Algorithm 1, Appendix A-C; we believe that based on all publicly available resources and provided instructions and explained algorithm, readers can reproduce our results. We also added raw and BERT-Sort encoded data sets that submitted to AutoML tools among seeds and the evaluation results (e.g., see *outputsout_roberta*) in GitHub Repository.

   (e) Did you specify all the training details (e.g., data splits, pre-processing, search spaces, fixed hyperparameter settings, and how they were chosen)? [Yes] All details are explained in Section 4 and Appendix section and it is linked to GitHub Repository.

   (f) Did you ensure that you compared different methods (including your own) exactly on the same benchmarks, including the same datasets, search space, code for training and hyperparameters for that code? [Yes] We strictly follow a fair comparison to compare our approach against others; see Section 4.

(g) Did you run ablation studies to assess the impact of different components of your approach? [Yes] See Appendix D as an example where we consider different types of MLMs and different input formats.

(h) Did you use the same evaluation protocol for the methods being compared? [Yes] We ensured to report a fair comparison as explained in Section 3 and 4 (i.e., pre-process step is conducted on both OrdinalEncoder and BERT-Sort encoder; however, we expect that those pre-process step is not apply to OrdinalEncoder in a real-world scenario that increases the performance of BERT-Sort.

(i) Did you compare performance over time? [Yes] See Section 4.3 where we limit only an hour for each AutoML tool with one seed, and a 30-minutes, 15-minutes for 4 seeds and 5 seeds, respectively.

(j) Did you perform multiple runs of your experiments and report random seeds? [Yes] The random seeds for Figure 8 is reported in GitHub repository.

(k) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The detail of all results are reported in GitHub repository that includes the ranges of outputs.

(l) Did you use tabular or surrogate benchmarks for in-depth evaluations? [Yes] All detailed results and summary reports are available in GitHub Repository.

(m) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.3

(n) Did you report how you tuned hyperparameters, and what time and resources this required (if they were not automatically tuned by your AutoML method, e.g. in a NAS approach; and also hyperparameters of your own method)? [Yes] All hyper-parameters are reported in Section 4 and Appendix section as well as listed in each experiment folder(*outputs*) in GitHub Repository.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] Added reference and link to the resources; See Appendix Section

(b) Did you mention the license of the assets? [No] the licence of all data sets and pre-trained MLMs are available in provided link as explained in Appendix section.

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We did not develop a model or a data set, we use only publicly available resources and provide the link to those resources.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We used only publicly available resources and we expect that the consent has been acquired by the provider, if it is applied.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Not applicable, we do not believe that the external resources contains personally identifiable information or offensive content since has been widely evaluated/used by ML community.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use any crowdsourcing or conduct any research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use any crowdsourcing or conduct any research with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use any crowdsourcing or conduct any research with human subjects.