NLP Reproducibility Checklist

Version 1.3, 2021-10-13

For all reported experimental results:

- ☐ A clear description of the mathematical setting, algorithm, and/or model
- ☐ A description of computing infrastructure used
- ☐ The total computational budget used (e.g. GPU hours), average runtime for each model or algorithm, or estimated energy cost
- ☐ The number of parameters in each model
- ☐ Corresponding validation performance for each reported test result
- ☐ A clear definition of the specific evaluation measure or statistics used to report results

For all results involving multiple experiments, such as hyperparameter search:

- ☐ The exact number of training and evaluation runs
- ☐ The bounds for each hyperparameter
- ☐ The hyperparameter configurations for best-performing models
- ☐ The method of choosing hyperparameter values (e.g. manual tuning, uniform sampling, etc.) and the criterion used to select among them (e.g. accuracy)
- ☐ Summary statistics of the results (e.g. expected validation performance, mean, variance, error bars, etc.)

For all datasets used:

- ☐ Relevant statistics such as number of examples and label distributions
- ☐ Details of train/validation/test splits
- ☐ An explanation of any data that were excluded, and all pre-processing steps
- ☐ For natural language data, the name of the language(s)
- ☐ A link to a downloadable version of the dataset or simulation environment
- ☐ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control

Curated by Jesse Dodge, direct questions and feedback to jessed@allenai.org
http://www.cs.cmu.edu/~jessed/NLP_Reproducibility_Checklist_V1.2.pdf