Appendix

Appendix A: Evaluation Metrics and resources used

Mean Squared Error
https://en.wikipedia.org/wiki/Mean_squared_error
Implementation SciKit Learn:
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

Mean Absolute Error
https://en.wikipedia.org/wiki/Mean_absolute_error
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

R2 score
https://en.wikipedia.org/wiki/Coefficient_of_determination
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

MCNemar's test
https://en.wikipedia.org/wiki/McNemar%27s_test
https://www.statsmodels.org/dev/generated/statsmodels.stats.contingency_tables.mcnemar.html

Wilcoxon signed-rank test

https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html

Appendix B: Validation Scores

This appendix reports the validation scores for the metric that was used for model selection. This is accuracy in case of the first task, accept/reject prediction on PeerRead and the R^2 score in case of the second task, citation score prediction. The highest validation scores reported for these models correspond to the epoch and corresponding model checkpoint that is then selected to use as the model for generating the predictions for the test set with.

Table B1: Task 1 - PeerRead validation accuracy different models, for the model checkpoints that are also used for the test set (selected because of highest validation score).

arXiv sub-domain dataset	Average Word Embeddings	BiLSTM (re-implemented)	HAN	HAN_{ST}
artificial intelligence	68.3 ± 2.24%	90.1 ±0.28%	89.9 ± 1.71%	88.8 ± 1.29%
computation & language	70.5 ± 0.00%	79.3 ± 0.44%	77.5 ± 0.87%	77.8 ±1.16%
machine learning	69.0 ± 0.40%	77.1 ± 1.28%	75.3 ± 2.82%	76.1 ± 1.39%

Table B2: Task 2 - S2ORC dataset validation R^2 scores different models, using the model checkpoints that are also used for the test set (selected because of highest validation score).

Average Word Embeddings	BiLSTM (re-implemented)	HAN	HAN_{ST}
0.211 ± 0.0008	0.229 ± 0.0064%	0.257 ± 0.0047	0.261 ± 0.0076

Appendix C: Model Runtimes

This appendix reports both the time per example, the most normalized form of runtime measurement, as well as the total time per epoch.

For the first task, PeerRead accept/reject prediction, we used the computation and language domain as representative for time measuring.

The minority class contains 576 examples, therefore the resampled number of examples per epoch is 1152 (2 times the size of the minority class).

The average time per example is higher on the PeerRead task across systems, because the examples contain markedly more text.

However, in the second task: citation prediction, the number of examples is much higher (78894), which explains that despite the lower time per example, the total time per epochs is much higher in this task.

Table C1: Average time/example during training. Time in milliseconds.

Task	AWE	BiLSTM	HAN	HAN_{ST}
PeerRead (computation and language domain)	20.5	55.3	75.9	75.8
citation prediction	4.3	7.5	11.04	10.7

Table C2: Average time/epoch during training. Time in seconds.

Task	AWE	BiLSTM	HAN	HAN_{ST}
PeerRead (computation and language domain)	24	64	87	87
citation prediction	349	593	872	848

The main observation here is that the computational cost jumps from AWE to BiLSTM and again from BiLSTM to HAN_{ST}. However, the increase in computation cost from BiLSTM to HAN_{ST} is less than factor 1.5 for both tasks, so this is manageable.