D. Reproducibility

Our code is publicly available at `https://github.com/sfschouten/court-of-xai`. We conducted our experiments on Amazon Web Services `g4dn.xlarge` EC2 instances using an NVIDIA T4 GPU with 16GB of RAM. The version of PyTorch was `1.6.0+cu101`. We refer to Table 7 for the average time to train each model on each dataset.

The DistilBERT model contained 66955779 trainable parameters and the BiLSTM model contained 12553519 trainable parameters, as reported by the AllenNLP library [51]. Table 8 lists the number of instances in each split of each dataset and Table 9 details the accuracy of our models on the validation sets during training.

Links to download versions of all datasets are included in our code repository. For posterity, links to all datasets are listed here:

- **SST-2**: `https://github.com/successar/AttentionExplanation/tree/master/preprocess/SST`
- **IMDb**: `https://github.com/successar/AttentionExplanation/tree/master/preprocess/IMDB`
- **SNLI**: `https://nlp.stanford.edu/projects/snli/`
- **MNLI**: `https://cims.nyu.edu/~sbowman/multinli/`
- **XNLI**: `https://cims.nyu.edu/~sbowman/xnli/`
- **Quora Question Pair**: `https://drive.google.com/file/d/12b-cq6D45U5c-McPoq2wsFjzs6QduY_y/view?usp=sharing`

Table 7. Number of minutes (average ± standard deviation) required to train each model on each dataset reported across three seeds.

	BiLSTM	DistilBERT
MNLI	8.65 ± 0.635	296.228 ± 48.859
Quora	7.567 ± 1.404	380.056 ± 124.911
SNLI	31.495 ± 5.618	126.395 ± 22.909
IMDb	1.122 ± 0.107	24.2 ± 1.212
SST-2	0.216 ± 0.029	2.833 ± 0.65

Table 8. Number of instances in each split of each dataset before any exclusions based on length (see Section 4.1). Since MultiNLI has no publicly available test set, we use the English subset of the XNLI dataset.

	Training	Validation	Test
MNLI	392702	10000	5000
Quora	323426	40429	40431
SNLI	550152	10000	10000
IMDb	17212	4304	4363
SST-2	8544	1101	2210

Table 9. Validation accuracy (average ± standard deviation) of the selected model epoch reported across three seeds.

	BiLSTM	DistilBERT
MNLI	67.088 ± 0.190	77.338 ± 0.251
Quora	83.232 ± 0.139	88.801 ± 0.055
SNLI	81.535 ± 0.041	87.679 ± 0.075
IMDb	87.975 ± 1.375	88.587 ± 0.489
SST-2	80.696 ± 0.403	83.066 ± 0.692