

BIASPAD: A BIAS-PROGRESSIVE AUTO-DEBIASING FRAMEWORK

Anonymous authors

Paper under double-blind review

A APPENDIX

A.1 HYPERPARAMETERS FOR THE REPRODUCED METHODS IN ANALYSIS

For the work by Sanh et al. (2020), we fine-tune a BERT-tiny model with the following hyperparameters: 3 epochs of training with a learning rate of $3e-5$, and a batch size of 32. The learning rate is linearly increased for 2000 warming steps and linearly decreased to 0 afterward. We use an Adam optimizer with default hyperparameters.

For the work by Utama et al. (2020), we fine-tune a BERT-base model with the following hyperparameters: 2000 examples and 3 epochs of training with a learning rate of $2e-5$, and a batch size of 32. The learning rate is linearly increased for 2000 warming steps and linearly decreased to 0 afterward. We use an Adam optimizer with default hyperparameters.

A.2 LIST OF NEGATION WORDS

We use the following list to filter the negation words bias,

[no, not, none, nothing, never, aren't, isn't, weren't, neither, don't, didn't, doesn't, cannot, hasn't, won't.]

A.3 CALCULATION OF THE WORD OVERLAP RATE

We use the following formula to calculate the overlap rate,

$$\text{overlap rate} = \frac{\#words(sentence1 \cap sentence2)}{\min(\#words sentence1, \#words sentence2)}. \quad (1)$$

REFERENCES

- Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*, 2020.
- Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Towards debiasing NLU models from unknown biases. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7597–7610. Association for Computational Linguistics, 2020. URL <https://doi.org/10.18653/v1/2020.emnlp-main.613>.