Political Discussion at Scale: Argument Mining with Deep Learning

Supplementary Information

Isaac D. Mehlhaff*

August 28, 2022

Contents

*The University of North Carolina at Chapel Hill; mehlhaff@live.unc.edu.

S1 Hyperparameter Tuning

Table S1 presents the results of a grid search over three hyperparameters in the random forests with extreme gradient boosting: the learning rate, the proportion of data sampled in each tree, and the proportion of data sampled at each node. All three hyperparameters were allowed to take values in {0.2, 0.4, 0.6, 0.8}. Table S2 presents the results of a grid search over five hyperparameters in the fine-tuned BERT neural networks: the initial learning rate in {0.00005, 0.00001, 0.00015}, the weight decay rate in {0.01, 0.05, 0.1}, the proportion of the training data used for warm-up in {0.05, 0.1, 0.2}, the proportion of nodes dropped by the dropout layer in {0.2, 0.3, 0.4}, and the batch size in {32, 64, 128}.

Table S1: Results of Hyperparameter Tuning in Random Forests with Extreme Gradient Boosting

Task	Learning Rate	Tree Subsample	Node Subsample
Disagree / Agree	0.4	0.4	0.8
Emotion / Fact	0.4	0.4	0.8
Attacking / Respectful	0.6	0.6	0.6
Nasty / Nice	0.8	0.4	0.6
Individual / Audience	0.2	0.6	0.6
Defeater / Undercutter	0.2	0.2	0.8
Counterargue / Attack	0.2	0.4	0.4
Question / Assert	0.8	0.8	0.6
Argument Quality	0.8	0.8	0.6

Table S2: Results of Hyperparameter Tuning in Fine-Tuned BERT Neural Networks

Task	Initial Learning Rate	Weight Decay Rate	Warm-Up Partition	Dropout Proportion	Batch Size
Disagree / Agree	0.00015	0.05	0.05	0.4	32
Emotion / Fact	0.00005	0.1	0.2	0.4	64
Attacking / Respectful	0.00015	0.1	0.2	0.3	128
Nasty / Nice	0.0001	0.01	0.05	0.3	128
Individual / Audience	0.00005	0.05	0.2	0.4	128
Defeater / Undercutter	0.00005	0.05	0.2	0.4	128
Counterargue / Attack	0.00015	0.1	0.05	0.4	32
Question / Assert	0.00015	0.01	0.2	0.2	128
Argument Quality	0.0001	0.1	0.1	0.2	32

S2 Classifier Performance Plots

Figure S1 displays precision-recall curves for all classifiers. Black curves represent baselines and colored curves represent deep learning models. In theory, the best-performing model is the one with the curve closest to the upper-right corner of the plot, but this is often difficult to determine, especially when some classifiers' precision is relatively flat (like the baselines). This is why an optimal threshold is determined for each classifier separately and the results compared using a summary measure like the AUC or F1 score in the main text. Figure S2 displays receiver-operating characteristic curves for all classifiers, with the same color scheme as in Figure S1. The best-performing classifier will be the one closest to the upper-left corner of the plot. The AUC metrics reported in the main text are calculated from these curves.

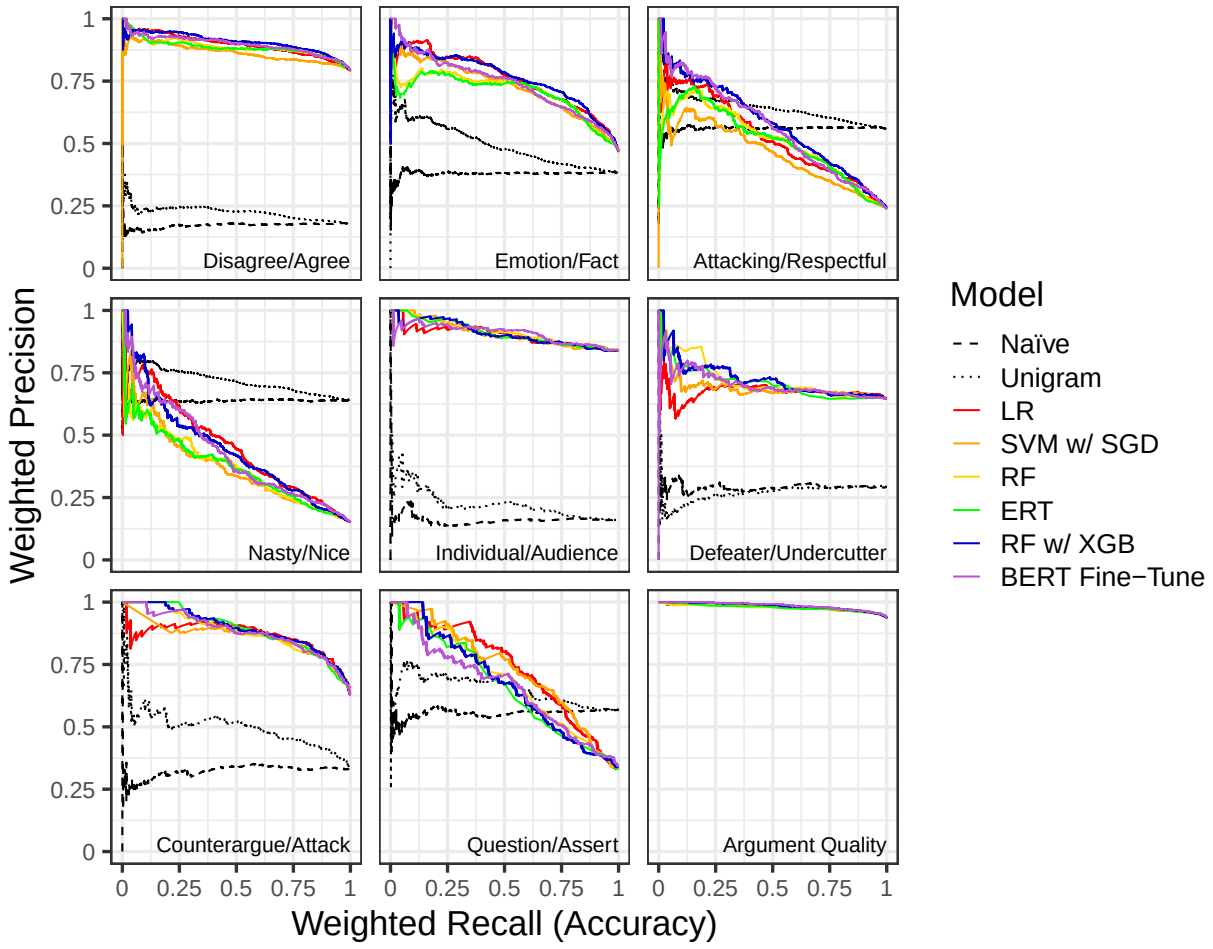

Figure S1: Weighted Precision and Recall Curves. Weighted recall is mathematically equivalent to accuracy. Black curves represent baselines, colored curves represent deep learning models.

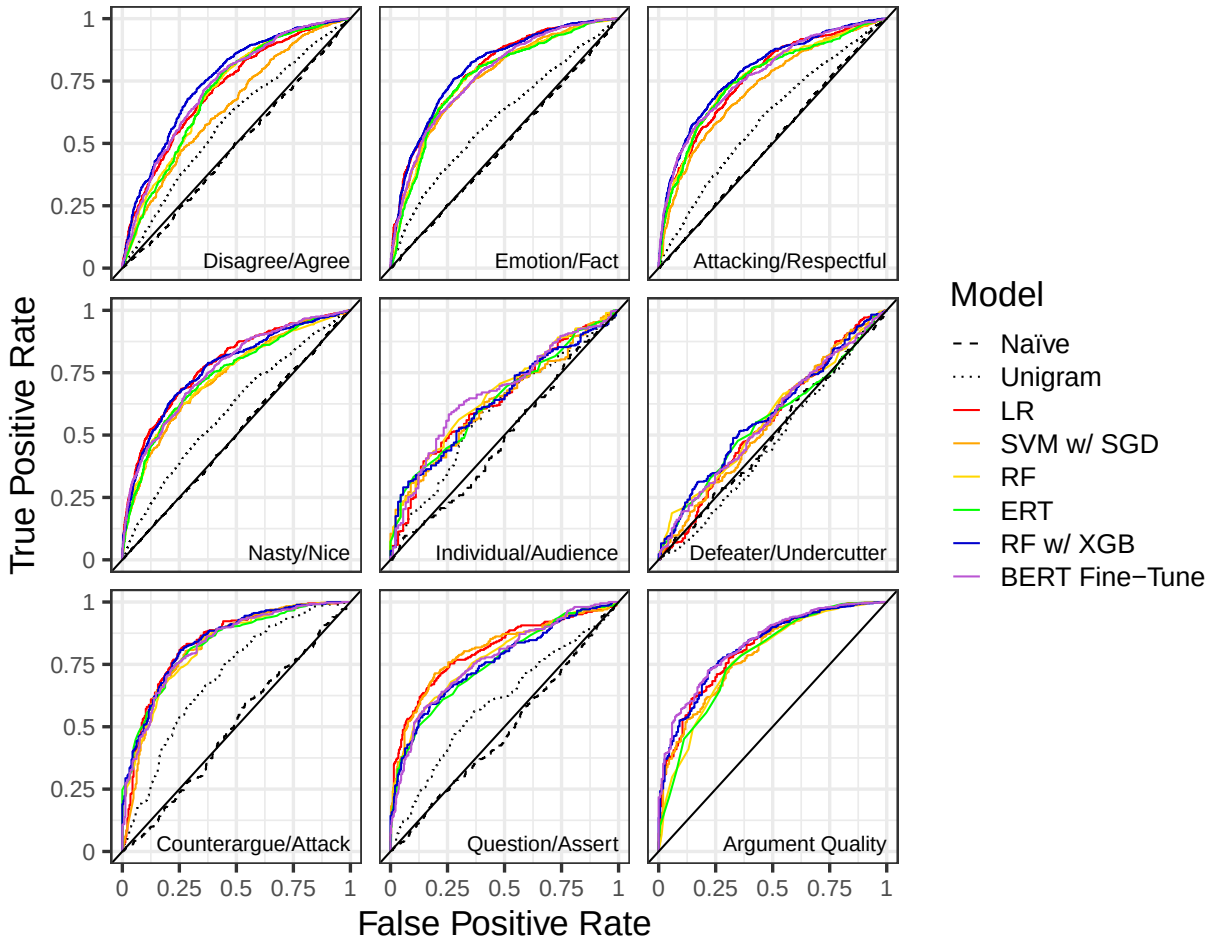

Figure S2: Receiver-Operating Characteristic Curves. Black curves represent baselines, colored curves represent deep learning models.