APPENDIX

This appendix contains a detailed explanation of the featurization and hyperparameter tuning experiments, used to determine the best fitting models for the Sick and Multiple tasks.

To select the best performing hyperparameters for each model variation, we ran 500 trials of random search with important hyperparameters sampled from reasonable distributions. We selected the best settings of each model variant using best average bias-adjusted F1-score over 5-fold cross validation on the training data, stratified by class label and biased/complement-sampled label.

Featurization of Documents

Document featurization and text normalization operations were evaluated to determine their impact on system performance. For all trials, we converted tokens to lowercase and filtered stop words (i.e., articles or function words such as "the," "an," "at," etc.). For each trial, we sampled a hyperparameter value at random for the following settings:

- Max document frequency (removing words that occur in more than a threshold percent of documents), sampled uniformly in $[.75, 1.0]$.
- N-gram range (using contiguous word phrases as features with phrases up to length n), sampled uniform categorically from $n \in \{1, 2, 3\}$.
- TF-IDF normalization (how to normalize the TF-IDF vectors), sampled uniform categorically from $\{L_1, L_2, None\}$.
- Whether to use IDF reweighting or not, sampled uniform categorically from $\{Yes, No\}$.

Classifier Hyperparameters

Logistic Regression

- Regularization strength,sampled log-uniformly from $\lambda \in [10^{-3}, ..., 10^4]$.
- Regularization norm type, sampled uniform categorically from $\{L_1, L_2\}$.

Random Forest

- Number of trees in forest, sampled uniform integer in $[10, ..., 200]$.
- Max number of features per tree as a function of the total number of features, D, sampled uniform categorically from $\{\sqrt{D}, log_2 D\}$.

SVM

- Regularization strength, sampled log-uniformly from $L_1, L_2, None$.
- Kernel function always set to linear.