Hyperparameter	Best	Available Settings
Random seed	7439947	integers from 0 to 9999999
# encoder layers	3	3, 5, 7, 10
Embedding size	200	100, 150, 200
Hidden layer size	50	50, 100
# attention heads	1	1, 2, 5, 10
Prop. dropout	0.3	.1,, .2, .3, .4, .5
Learning rate	1e-4	1e-2, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6, 5e-7, 1e-7
# Epochs	3	1, 2, 3
Batch size	150	100, 150, 200, 250, 300, 350

Table 1: Best and hyperparameters and search space

These appendices further specify the second experiment, in which the attitude verbs concern attitudes towards not further specified propositional constants.

1 Appendix: Training

For training the data is split in to a training and a test set. On the training data, 130 hyperparameter settings were explored during a randomised grid-search of the hyperparameter space. The hyperparameter grid explored is documented in table 1. For each explored hyperparameter setting, a 5-fold cross-validation was performed on the training data. The same random validation splits were used for all hyperparameter settings, i.e. independently of the random seed.

The best hyperparameter setting according to this search was then used for evaluating on the hold-out test set. The best hyperparameters can also be found in table 1. The entire training set was used for training the model in this final evaluation.

2 Further Architecture Details

The weights of the model were initialised with a Xavier uniform distribution. The biases were 0 initialised.

The gradients were cliped to a norm of 1.

The classification used the mean of all tokens of the last encoding layer. We used a Binary Cross Entropy Loss and the Adam optimization algorithm.