A Data Annotation

Annotators were asked to judge the toxicity of each comment, given the following definitions:

- VERY TOXIC: A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
- TOXIC: A rude, disrespectful, unreasonable comment or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.
- UNSURE: Due to polysemy, lack of context or other reasons.
- NOT TOXIC: Not containing any toxicity.

For annotation, we used the 'Figure Eight' platform and we invested 5 cents per row.[1] For the CAT-SMALL we employed high accuracy annotators (i.e., from zone 3), selected from 7 English speaking countries (i.e., UK, Ireland, USA, Canada, New Zealand, South Africa, Australia), and only ones allowing explicit content (we also warned about the explicit content in the title). 62 quiz questions were used. For the CAT-LARGE, we invested the same amount of money but all the annotators were able to participate (again, they ware warned for the explicit content). Inter annotator agreement was measured on the quiz questions with Krippendorff's alpha and was found to be 70% and 72% for the C and N sets.

GC annotators had one more question, which was asking them to compare the toxicity of the target comment to that of the parent comment. The main scope of that question was to make it less easy for annotators to ignore the parent comment.

B Hyper parameters

All systems were trained for 100 epochs with patience of 3 epochs. We performed early stopping by monitoring the validation ROC AUC.

BILSTM

The hidden size of the LSTM cells had size 128. We used batch size 128, max length 512, and we concatenated the forward and backward last hidden states before the FFNN. We used binary cross entropy for loss and Adam optimizer was used with default parameters (learning rate 1e-03).

CA-BILSTM-BILSTM

We used the same hyper-parameters with BILSTM but included one more bidirectional LSTM to encode the parent text. The parent biLSTM had 64 hidden nodes and we concatenated the forward and backward last hidden states. The parent and the target embeddings (the ones generated by the two biLSTMS) were concatenated before being passed to the FFNN.

BERT

We used a learning rate of 2e-05 for BERT and only unfroze the top three layers during training to our data. On top of the BERT [CLS] representation, we added a FFNN of 128 hidden nodes and a sigmoid to yield the toxicity probability. 128 tokens were used as maximum sequence length.

CA-SEP-BERT

A [SEP] token separated the two texts and the [CLS] token was used as with BERT. Same parameters with BERT were used.

CA-BILSTM-BERT

We used a bidirectional LSTM to encode the parent comment, similarly to CA-BILSTM-BILSTM. The biLSTM representation was concatenated with the [CLS] representation before the FFNN. All other parameters were set to the same values as BERT.

[1] https://www.figure-eight.com/