**Ideas**

* (From meeting) – see what happens if mask offensive words from dataset – can we still accurately tell toxic language based on context? (might be innocent if between friends)
* Current findings that performance of classifier doesn’t improve when context-aware due to small number of context-sensitive comments in dataset – use new dataset? Look at how to efficiently annotate larger corpora of comments in context? See if can find additional context beyond title/parent comment? – comments randomly sampled – look into particular topic/tones/frequently target communities?
* Evaluating relative benefit of CCTK dataset compared to sub-string matching of terms that reference an identity (effectiveness of annotator’s recognition of identities)
* Developing full taxonomy of different possible biases + systematic approach for metrics used in diagnosis
* Investigating how relationships between annotators (demographics/ if agreed with comment) /way annotated (multiple questions/asked independently) affects the annotations and classification results
* Whether some terms are predictive of constructive comments / how the context of a conversation affects the constructiveness
* Modelling contributions to discussions other than constructiveness – diverse points of view / healthy levels of disagreement
* Impact of personal attacks/toxic comments on target user’s future contributions?
* Based on conversation, predict likelihood of next comment being an attack