**Meeting 19/11/20 Notes**

* Finish literature review over weekend – talk about underlying problem instead of technology, in conclusion highlight research questions, potential solutions and hint at available datasets
* Predict identity of annotator from text and toxicity score, if toxicity score has predictive power over identity then bias present
* Both datasets – see if can generalise to other dataset, find similar words, label subset for verification (could annotate 500 samples if got funds), test model on one dataset and see if can generalise to other using validation set
* Then can manipulate text – gender debiasing/masking race... and see if results is the same (could still predict identity of annotator), use to strengthen argument that identity closely tied to comment, see if directed to race – find correlations e.g. one religion finding comments about that religion offensive but not offensive when religion masked
* Start developing code (begin with gender) – use BERT + contextual embedding to predict gender, think of how to combine toxicity score (get latent representation and multiply/pool or concatenate at last step) otherwise will get mixed in with all the other variables, think about gender balancing (could remove some male annotators), multitasking? predict multiple classes or use separate criteria? 1-class classifier – predict female or other just training on female data and do same for male (if bad prediction then other class/no bias)