**Meeting 22/10/20 Notes**

* Looking at how demographics of annotators affect classification results: improving accuracy by looking at bias in demographics, predicting demographic identity of annotator
* Cross-dataset validation not very good – reduces accuracy and differences cause bias/overfitting – can use subset of another dataset labelled with identities as a validation set and annotate it with demographics using Amazon Mechanical Turk
* Predicting toxicity of next comment – would need to look more into dataset and label distribution in conversations and after 1st toxic comment to see if conversations too similar (need good distribution of data and variety)
* Mix ideas of masking offensive words and looking at demographics, using subset of CCTK comments with identities – swap gender for sexist comments (see if women find same sexist comment as offensive if target is man), look for biases, could mask demographic info in comments, could flip tokens including names (e.g. gay to man, Mohammed to David)
* Keep an open mind on using extra datasets (use part of CCTK and collect demographic info) to study relationship between identity in comments and demographics of annotators
* Later step – trying to reduce the biases in the model without masking demographic info
* Literature review – no deadline; deliver something that I can benefit from when writing final paper – go through journey, adding papers, what was discussed, natural conclusions drawn, amplify choice of datasets – why chosen
* Modelling and solution design
* Download datasets (CCTK and Wikipedia Abusive Language) – do statistical profile for both (see if already table discussing CCTK stats)
* Could at end annotate larger dataset with demographics (/suggest for future work) - would modelling be better if labelled entire identity dataset? – could put forward to department and even try to get paper released for publication
* Write draft research questions
* Next week – make a start on the code