**Meeting 05/11/20 Notes**

* Few new comments can be added from aggression/attack dataset
* Few papers using same dataset and looking at bias so large space for new contributions; also have lots of baselines for model accuracy as a lot of literature on dataset, but be critical of “better datasets” as they may have chosen the samples that worked best for their models
* Can’t follow same trend of annotators in CCTK dataset as number is so varied (large standard deviation)
* Can we safely assume original data applies? (e.g. if similar toxicity scores to newly annotated then are toxicity subcategories likely to remain the same as well?). Don’t have subcategories such as racism so use current scheme of toxicity + same scheme as original annotators
* Look at reliable subgroup – remove comments with <10 annotations and only keep those with high agreement (e.g. >70%) and see how much size of set reduced by
* Cost estimation – maybe 2 cents per annotation? May get £1000-£2000 funding (will get more accurate estimation next week and insight from PhD student)
* Change crowdsourcing proposal to focus just on gender bias (include all reliable gender comments) – want really strong reliable study so ignore other subgroups (but mention them and why chose not to use them in paper). Can balance gender well, has the most comments of any demographic, less impacted with regional data, gender bias very well studied. Can look at algorithms for gender debiasing and also annotate same set of comments after gender debiasing and evaluate difference.
* Want feasible questions and to choose groups with the most impact and best use of time and money – can still look at other questions using demographics e.g. do religious people find comments more offensive?
* Can try to include other subgroup – look at overlap between gender and gay/lesbian or race, but that subgroup won’t be debiased
* Will still be first study to link gender with subjectivity of annotations and add debiasing element; also creating new dataset that will be helpful to field.
* Look at paper sent on gender debiasing – “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods”
* Next week: find specific dataset and start annotations before debiasing