**Project Diary**

**Completed:**

* (9/10/20) Started GitHub repository – containing project diary, meeting notes and summary of literature
* (15/10/20) Read 11 papers and made notes
* Constructed table of datasets and their strengths/weaknesses
* Started to come up with ideas for project focus
* (22/10/20) Read a few more papers
* Created table of ideas examining feasibility/if addressed in literature/potential datasets
* Completed ethics form
* (29/10/20) Investigated initial dataset statistics of demographics dataset
* Drew up project plan with background, aims, deliverables, Gantt chart
* Drafted some research questions
* Read more papers related specifically to area of project
* Began writing literature review
* (05/11/20) Continued writing literature review
* Looked at overlap of toxicity dataset with personal attack/aggression datasets and assessed viability of combining them
* Assessed current research using chosen dataset to see if there’s any useful research and look at how novel the idea is
* Created proposal on new dataset, looking at number of annotators, what questions to ask, how dataset composed and why chosen
* (12/11/20) Continued writing literature review
* Analysed demographics of more reliable subgroup of data (>=10 annotators, high agreement)
* Changed crowdsourcing proposal to focus on gender bias
* Identified groups overlapping with gender, choosing race as an additional demographic group to include in annotations
* Looked into gender/race swapping/debiasing
* (26/11/20) Finished writing literature review
* Built toxicity classifier (BERT) and adapted to predict annotator gender based on comment and toxicity score
* (3/12/20) Examined detoxify library and Roberta-base model
* Adapted classifier – balanced data by bins (female very toxic, male very toxic, female toxic…), supplemented comments with “toxic”/”neutral”/”healthy”, predicted gender for each bin (labels = gender) and predicted bins (labels = bin numbers)
* Examined true negative and true positive distributions in terms of gender likelihoods
* Results showed some overfitting and bias (in meeting notes for this week)
* (10/12/20) Word clouds for annotators of different genders
* Built autoencoder for reconstructing text to see accuracy for when autoencoder trained on different groups

**To Do:**

* Extend literature review into mini version of paper
* Train classifier on 3 classes separately and examine how gender predictions differ
* Replace softmax layer of normal network with 1-class SVM/clustering