**Project Diary**

**Completed (updates on past week given on date of each supervisor meeting):**

* (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.
* (11/01/21) Focused on coursework over Christmas holidays – nothing on project done.
* (19/01/21) Father in hospital – no work done this week.
* (26/01/21) Trained BERT classifier on toxic, neutral, nontoxic data separately on balanced male and female classes, tried to predicted gender of annotator given comment.
* Found bias as classifier showed pattern assuming more comments were annotated by men for all classes (should’ve been 50/50 and not picked up on anything if no bias existed) so this clearly shows that men and women annotate comments differently.
* (02/02/21) Grandmother died, dealing with bereavement – no work done this week.
* (09/02/21) Working on using Allen NLP/Captum libraries to add explainability to results.
* Ran Captum tutorial and started to adapt to implemented classifier.
* (17/02/21) Got sentiment analysis working and found male annotators appeared to react more strongly to offensive words.
* (23/02/21) Ran models separately for male and female data, found both models had comparable performance and performed worse on female test data.
* Created words clouds and graphs showing offensive words in various categories of test data to show differences.
* Tested models on data with offensive words removed; female model performed worse than male model.
* (02/03/21) Began writing conference paper (introduction, related work, and data).
* Trained male and female models without offensive words; female model performed slightly better than male model and performance was comparable to models trained on offensive data on test data containing offensive words.
* Retrained gender classification with offensive words removed; found male predictions strongly reliant on offensive words.
* (09/03/21) Found correlations between true/predicted labels for toxicity/gender and number of offensive words in a comment.
* Continued writing paper, writing results and solution.
* (18/03/21) Ran models 5 times with different seeds to get means and standard deviation.
* Created slides and prepared talk for NLP lecture.

**To Do:**

* Give NLP talk – arrive 20 mins early to sort out technology.
* Continue writing paper – **change tables into box plots and compare which is better.**
* Separate paper into conference and project papers and adapt each – **organise meeting** after writing each one.
* (Conference) **Add a footnote** that the code and the data will be available on github after the conference paper has been accepted.
* (Project) Include fact that gave talk in NLP lecture.
* (Oral) Can mention that paper under review.