**Meeting 10/12/20 Notes**

* 3-class classifier using previous BERT classifier – train separately on toxic, neutral, nontoxic data predicting gender – less of an imbalance of gender in same bins compared to whole dataset, see where biggest differences between male and female exist so we can then exploit specific bins or combine bins achieving similar accuracy
* Autoencoder shows performance may not be as good for toxic language without swearing
* Train normal network and replace softmax layer with 1-class SVM or clustering (could train separately – look at number of features at the top)
* We already know there’s a problem with bias – just want to investigate it as much as possible
* With thorough investigation we can then target debiasing algorithms to specific problem/class
* For annotations – they are revising the Amazon mechanical turk policy – could be as early as February but more likely to be March

This week’s results:

Similar word clouds for annotators of different genders

Autoencoder – repeated “fuck you” results for reconstruction of toxic text

