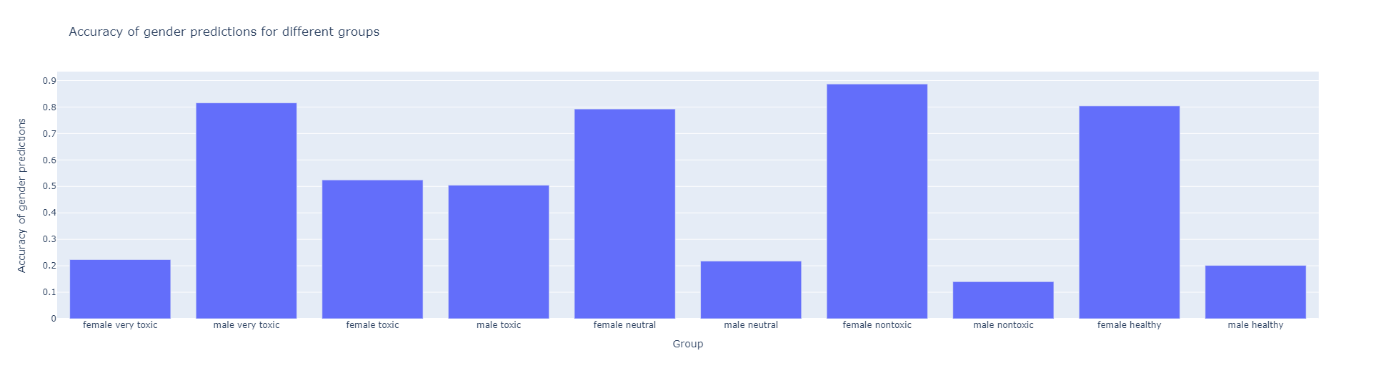
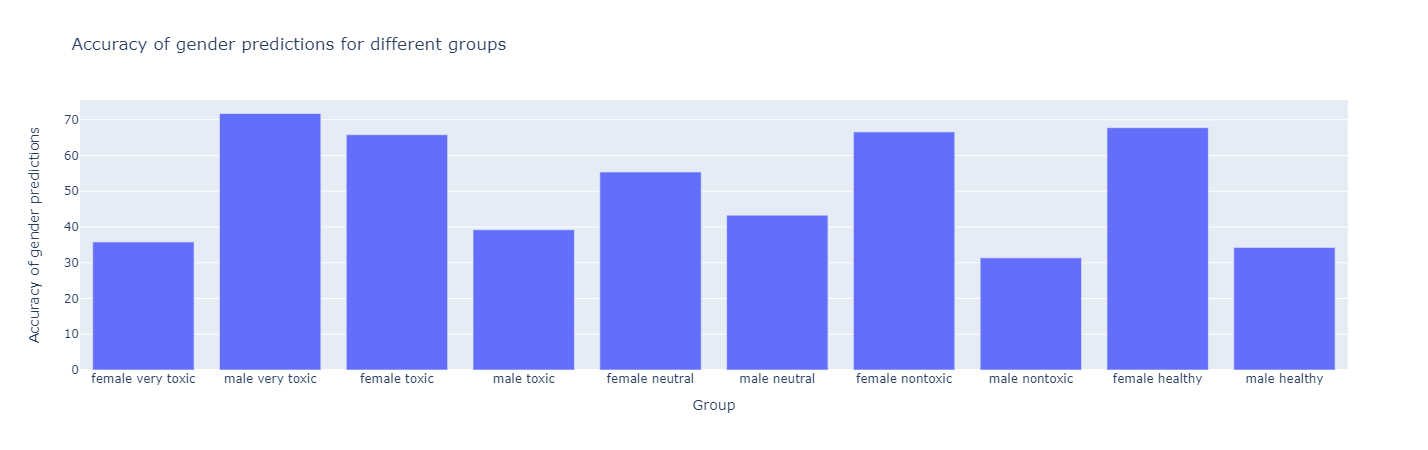
**Meeting 03/12/20 Notes**

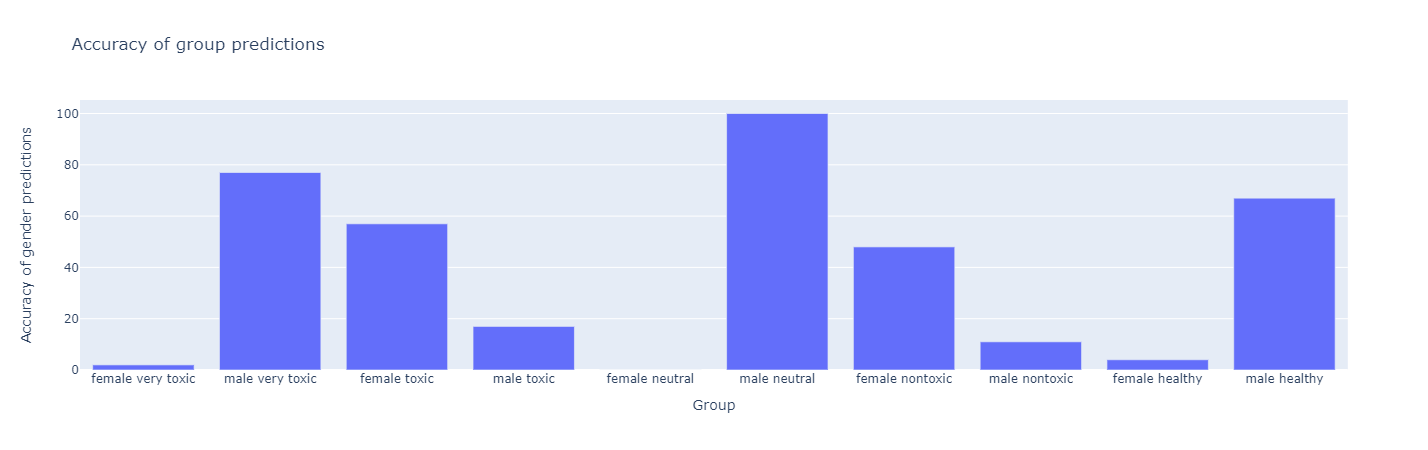
* Break down topic distribution (sklearn, LDA), generate word clouds for male and female annotators (want to be similar). Word clouds are insightful and can be included in final paper
* Results from this week show bias and overfitting
* Can include all results in final paper whether they show what we want or not
* Try 1-class classifiers trained only on 1 group’s data, look at autoencoders, other data should get higher reconstruction error. Can be used as a way to detect bias – biased if it gets classification right as shouldn’t be able to tell the difference between male and female annotators. E.g. train one on female very toxic and one on male very toxic then swap data for testing. Can use all of group’s data for training in this scenario. Start with just 2 groups then expand to rest of bins. Use as proof-of-concept confirming bias results, if bias shown for 1-class classifier as well then problem with data.
* Closely examine a few of the extreme examples (such as very high accuracy)
* Next step is debiasing to try to get 50:50. If current results are 60:40 that makes point as 10% is massive error rate in sensitive applications

This week’s results:

Predicting gender of each bin, bin-balanced data, added toxic/neutral/healthy to comment data



Predicting bin data in (labels = bins)



Distribution of true positive and true negative gender predictions:

