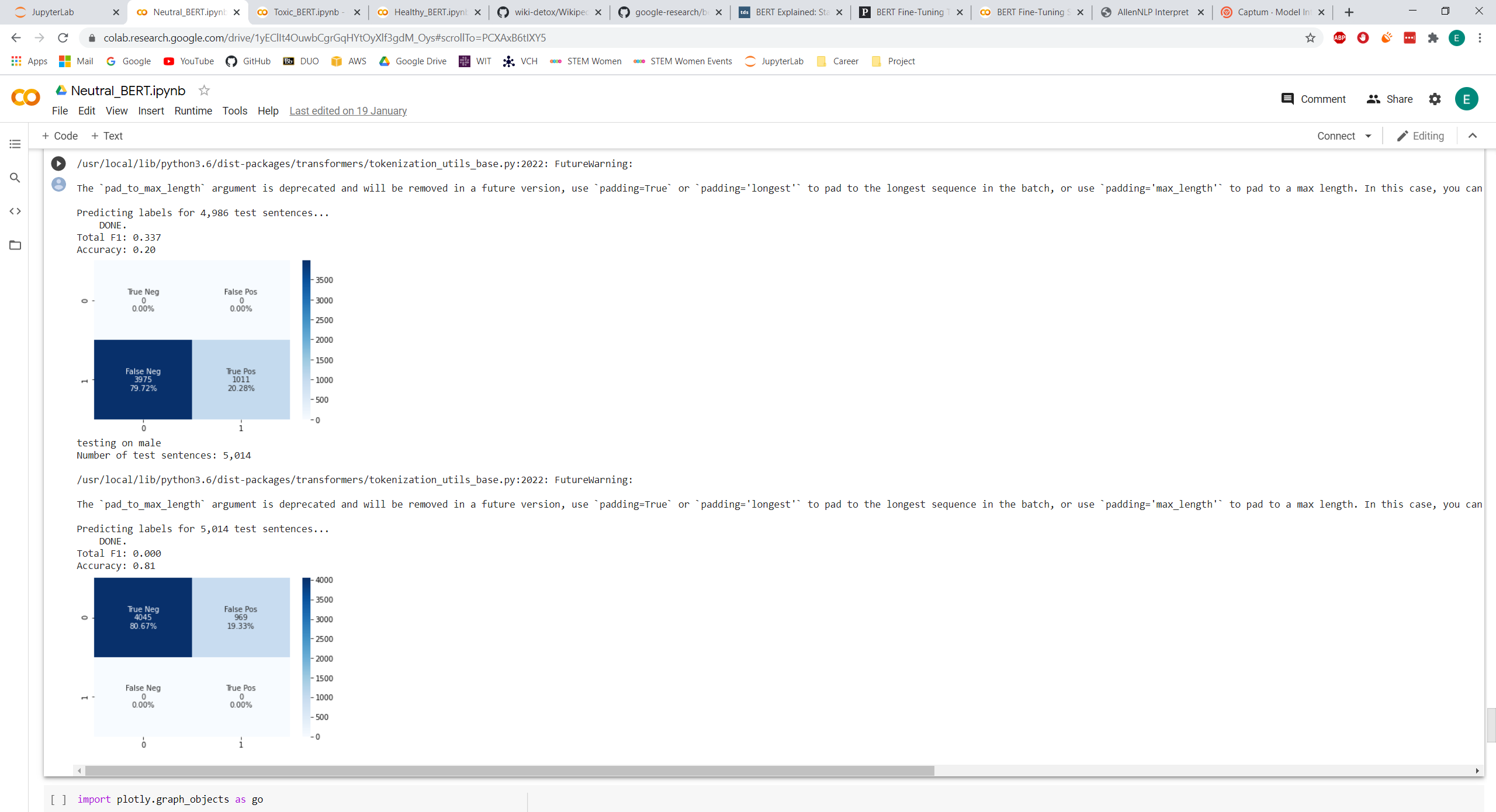
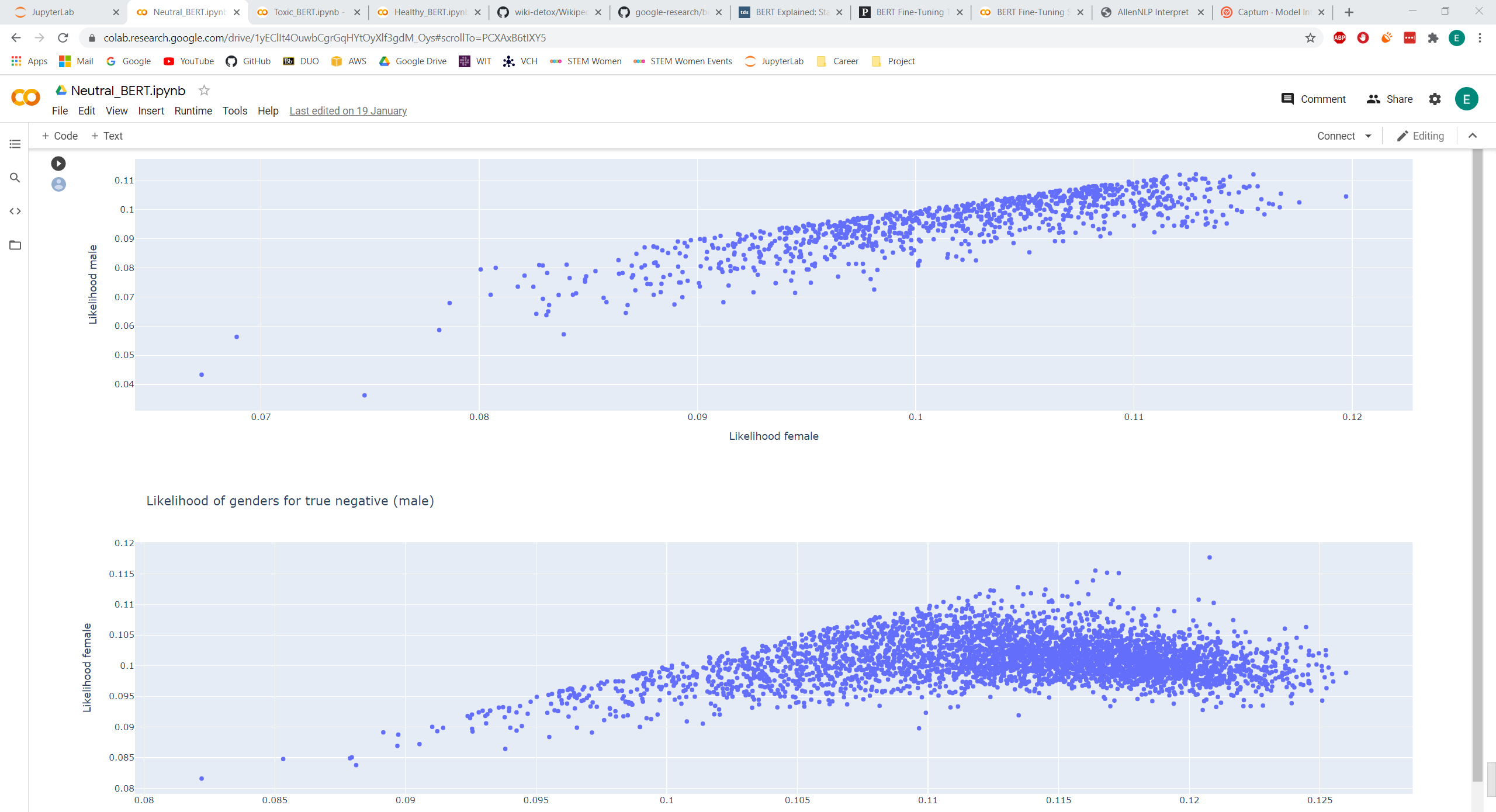
**Meeting 10/12/20 Notes**

* Results below found and discussed – indicative of bias present in data, especially toxic data (already novel discovery showing how important it is to get diverse annotations). This shows that men and women annotate comments (especially toxic ones) differently.
* Add interpretation/explainability algorithms on top of classifier to visualise words with higher gradients that are relied on to make the classification decision (which features aid in gender prediction). Use libraries (<https://allennlp.org/interpret>, <https://captum.ai/tutorials/IMDB_TorchText_Interpret>) to get gradients. Run sentiment tutorial, plug classifier in, replace sentiment with gender as binary sentiment.
* Look at how to change gradients found.
* Create text generation algorithm that given prompt (start of sentence) will complete with language that would be annotated as toxic by annotator of given gender. Could use this to augment dataset.

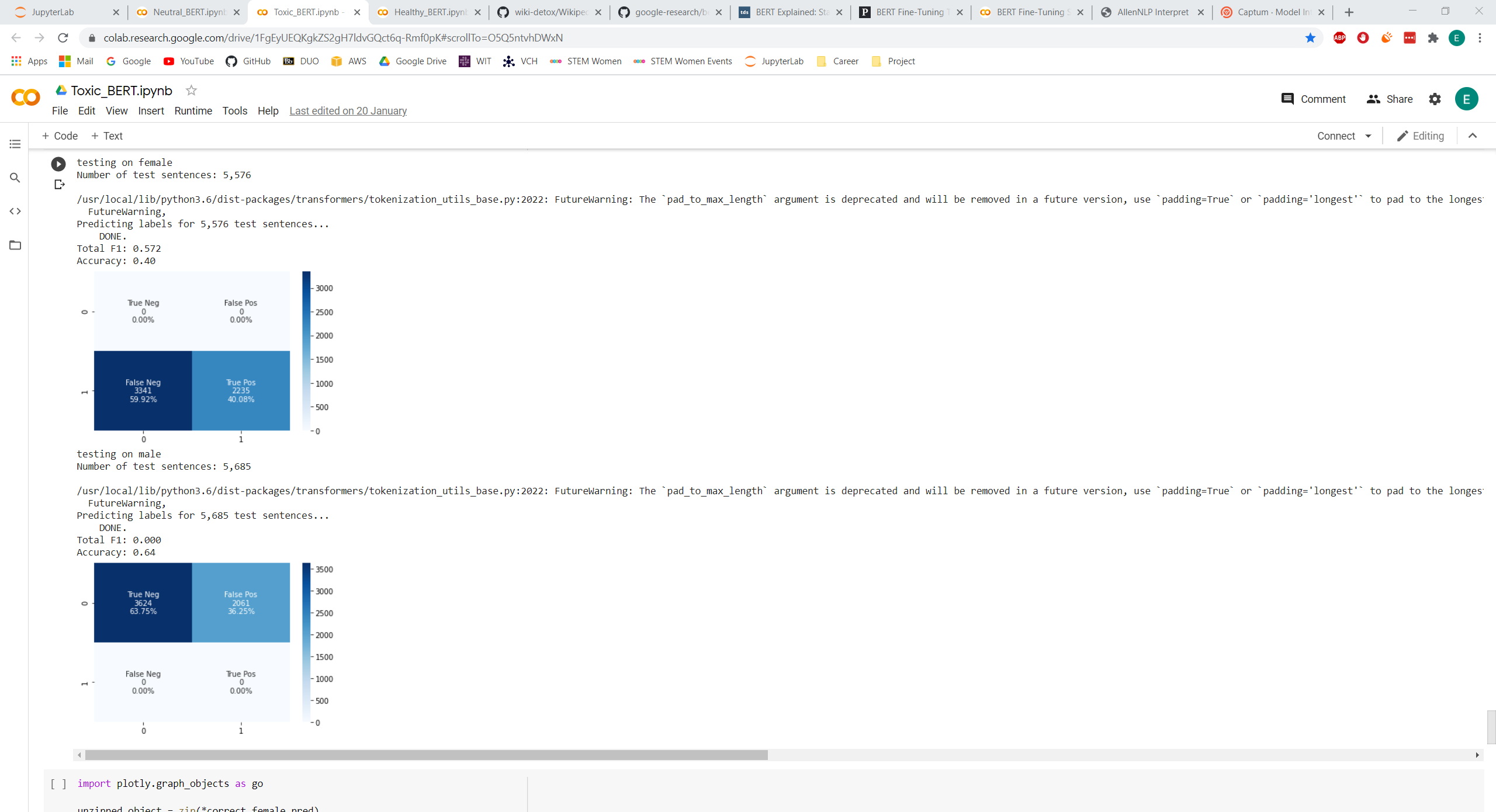
This week’s results:

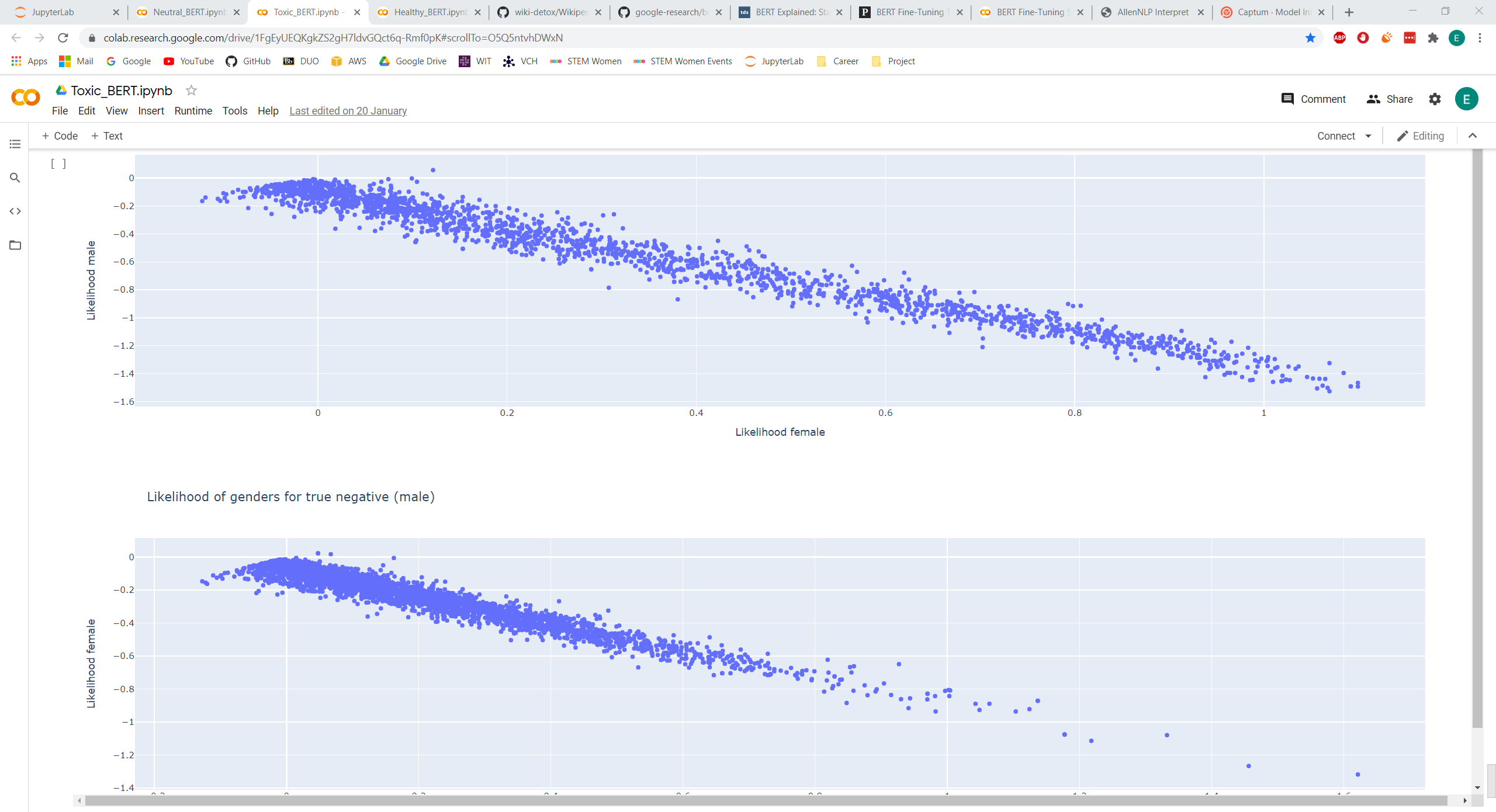
Neutral\_BERT – trained on neutral data balanced for male and female – predicting gender of annotators – predicted roughly 80% male for test data. Likelihoods for true positives and negatives all positive indicating made less of a distinction between male and female but erred on the side of male (both likelihoods increased together).





Toxic\_BERT – toxic balanced data predicting gender of annotators – predicted roughly 60% male for test data. Strong pattern with little variation in likelihoods of predicted genders indicates it learnt a pattern – likelihood of being female increases as likelihood of being male decreases. This all indicates that bias was present (should’ve gained no information and predicted 50/50).





Healthy\_BERT – nontoxic balanced data predicting gender of annotators – predicted roughly 58% male for test data. More variation in likelihood scores of true positives and negative than toxic data so was less confident in scores given. Same pattern of likelihood female increasing as likelihood male decreases. Indicative of bias (should not have found a pattern and returned 50/50 split).

