**Meeting 02/03/21 Notes**

* Results below found and discussed. Female model performed slightly better on all test data which may be due to women reading more into context/semantics. Test data with offensive words had 3% improvement in performance compared to the data without offensive words (much smaller difference than when trained on sentences including offensive words). Comparable performance to models trained using offensive words on offensive word test data (female test data exceeds other accuracy) suggesting that training without offensive words forces the model to look for factors other than offensive words and this overall produces a comparable model depending on the test data.
* Results for gender prediction: removing offensive words from training data reduces the % of test data predicted male. When tested using data without offensive words, the female test data produced a roughly 50/50 split of predictions between male and female, while the male data was still slightly skewed towards male predictions. This suggests that the offensive words pushed the prediction of the model towards male, and it becomes a lot less sure of gender when the offensive words are removed, but it still has some way of telling male data apart. When trained on just toxic data (not very toxic) the confusion matrix was also more balanced, indicating again that it was the offensive language skewing the model towards predicting male. (Male model more biased towards toxic words).
* Results continued: Integrated gradients a lot more scattered and less strong attributions when offensive words removed from test and training data. Integrated gradients notice offensive words when they are included in test data but they may not be as strongly attributed as when they were included in training data. Offensive words in test data likely to have been picked up as BERT is pretrained on Reddit data that will include offensive words, which is why they were recognised.
* For next week: measure correlation between number of toxic words and true label/predicted label for all scenarios (toxicity and gender classification). Hoping for a decreased correlation when trained with no offensive words and tested with offensive words. See if high correlation between toxicity classification and toxic words, then if high correlation between male annotators and toxic words before masking. For predicted labels for gender, use attribution scores for words highlighted in sentiment analysis. (Continue writing paper as well).
* For paper: highlight that we’re reducing bias by masking toxic words as part of contribution.

**This week’s results:**

In Male\_BERT\_No\_Offensive.ipynb, Female\_BERT\_No\_Offensive.ipynb, Gender\_BERT\_No\_Offensive.ipynb, Gender\_BERT\_No\_Offensive\_Just\_Toxic.ipynb:

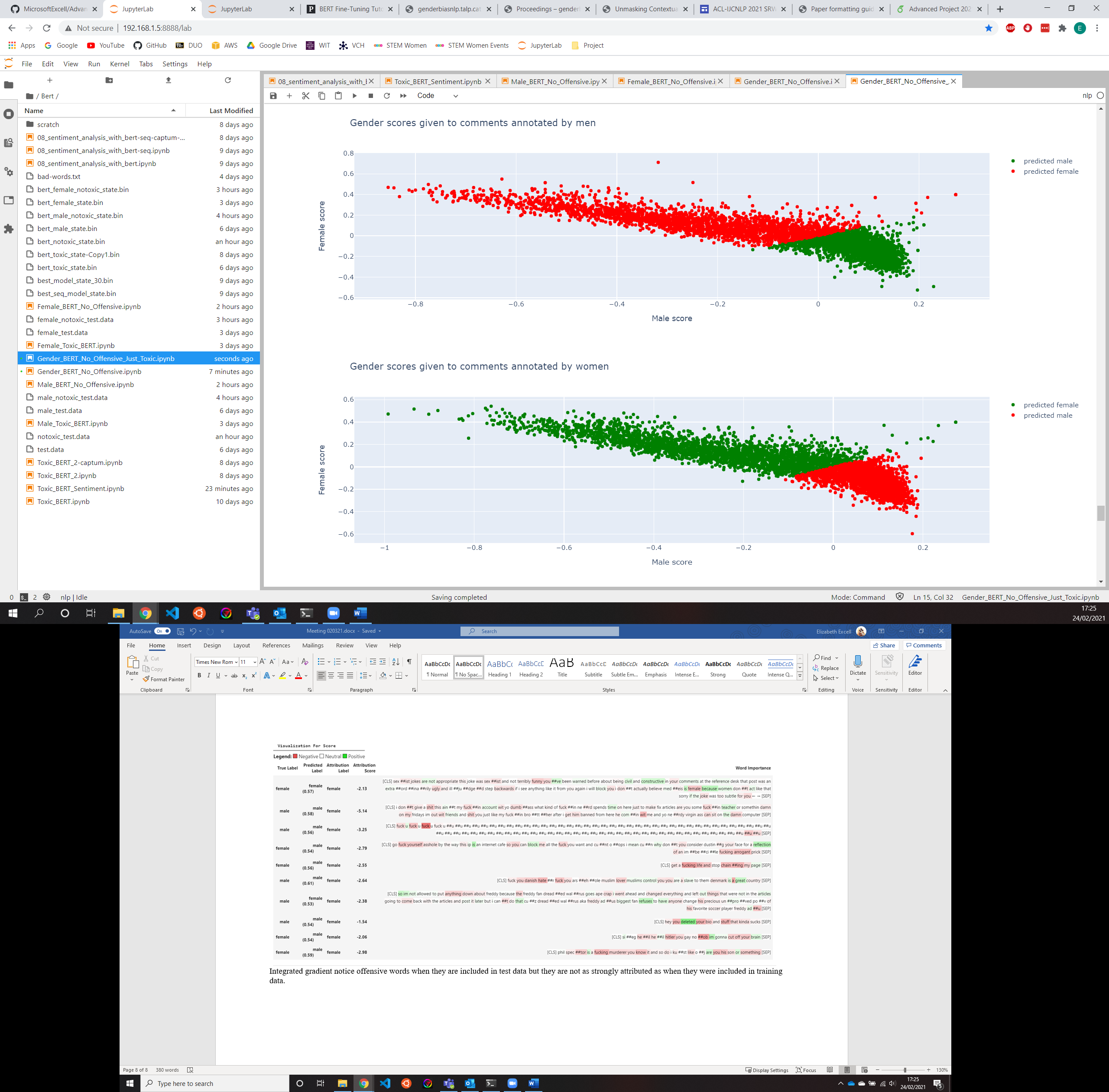
Note: All toxic = toxic + very toxic

**Confusion matrices**

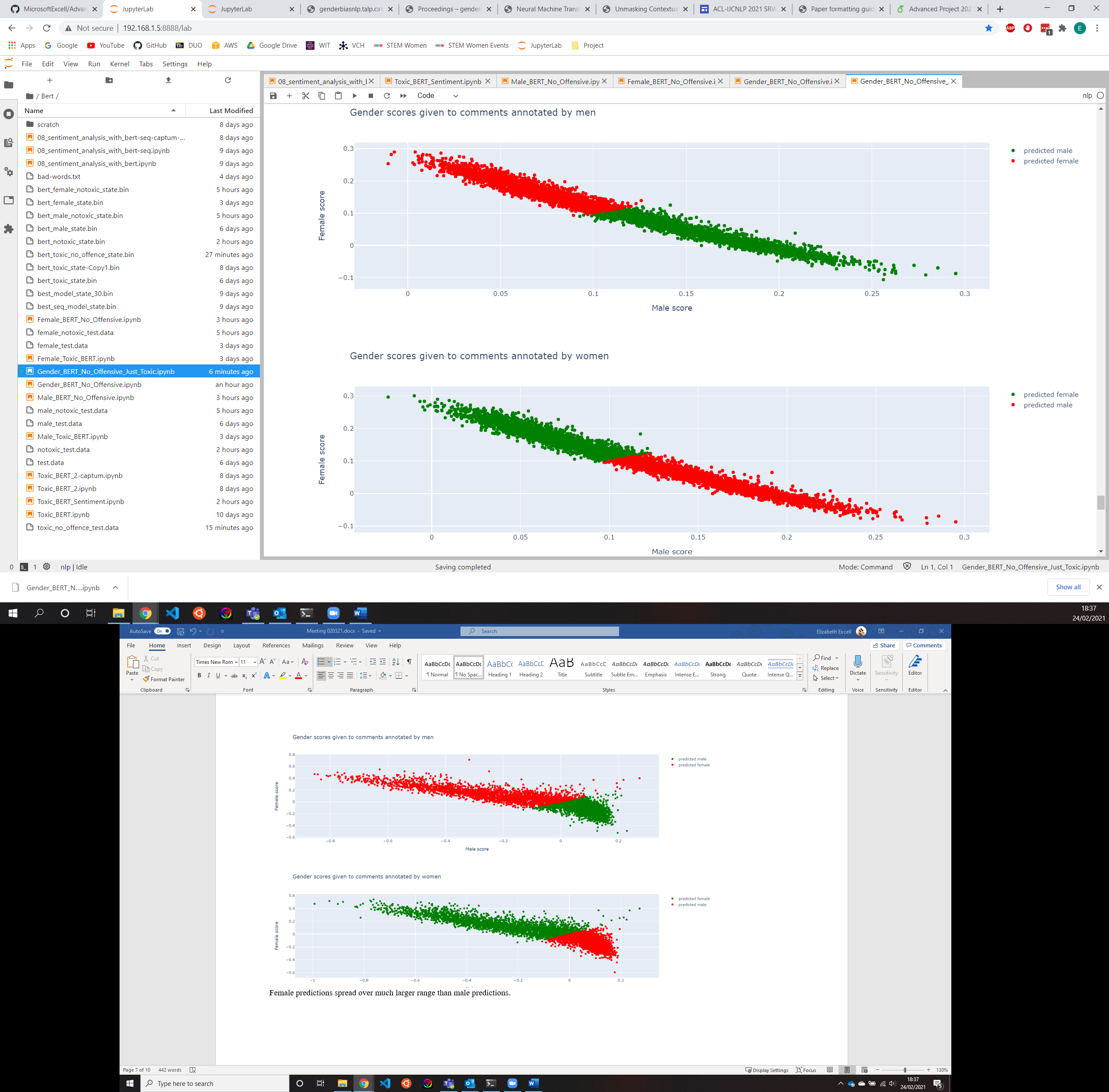
|  |  |  |
| --- | --- | --- |
| Testing Data\Training Data | Male without offensive words | Female without offensive words |
| Male All Toxic with offensive words |  |  |
| Female All Toxic with offensive words |  |  |
| Male All Toxic  without offensive words |  |  |
| Female All Toxic  without offensive words |  |  |

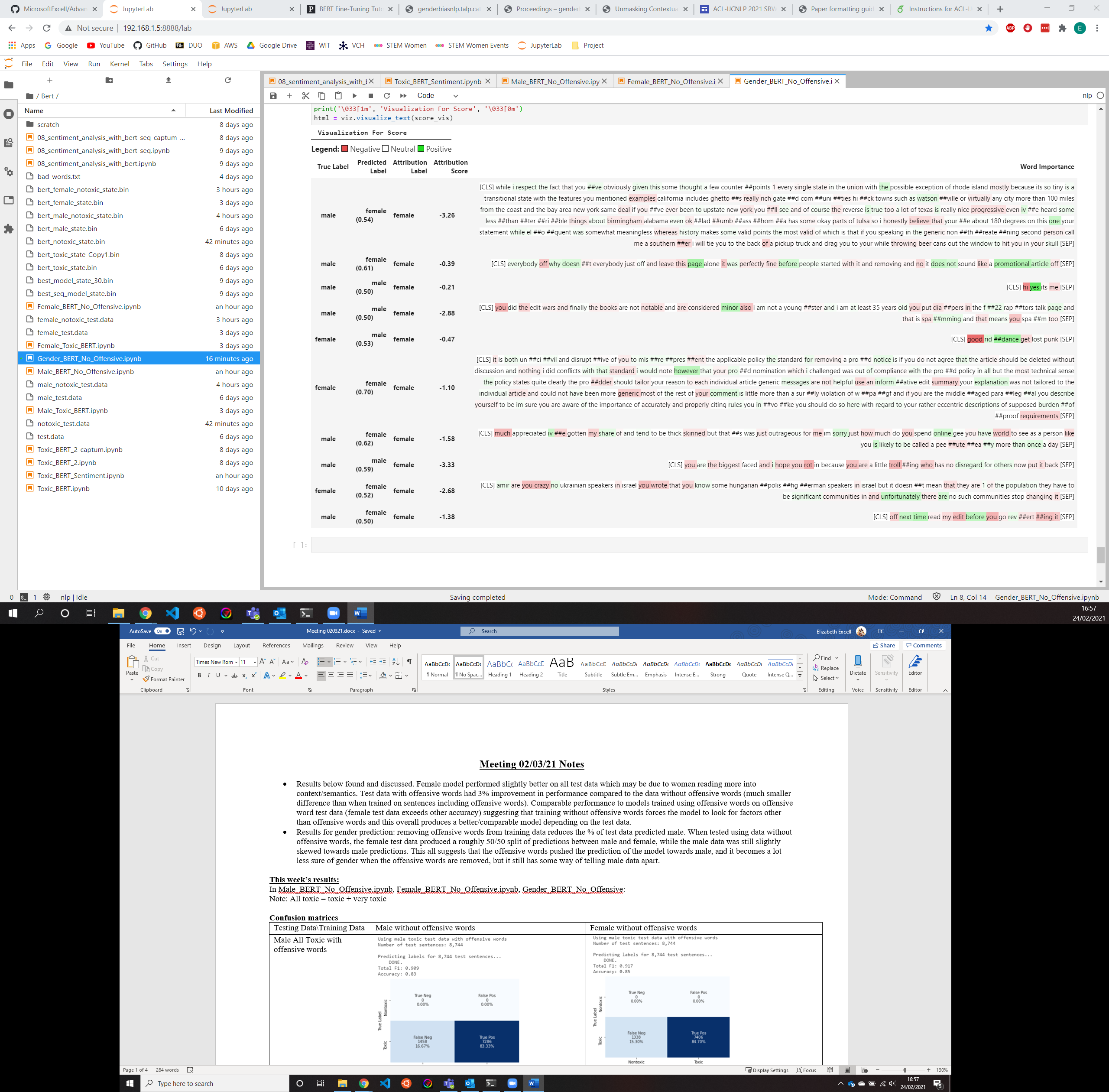
**Gender classification when offensive words removed from training dataset:**

|  |  |
| --- | --- |
| Training/Testing Data | Confusion matrix |
| All Toxic with offensive words/testing with offensive words  (Previous results for reference) |  |
| All Toxic without offensive words/testing with offensive words |  |
| All Toxic without offensive words/testing without offensive words |  |
| Just toxic (not very toxic) without offensive words/testing with offensive words |  |
| Just toxic (not very toxic) without offensive words/testing without offensive words |  |

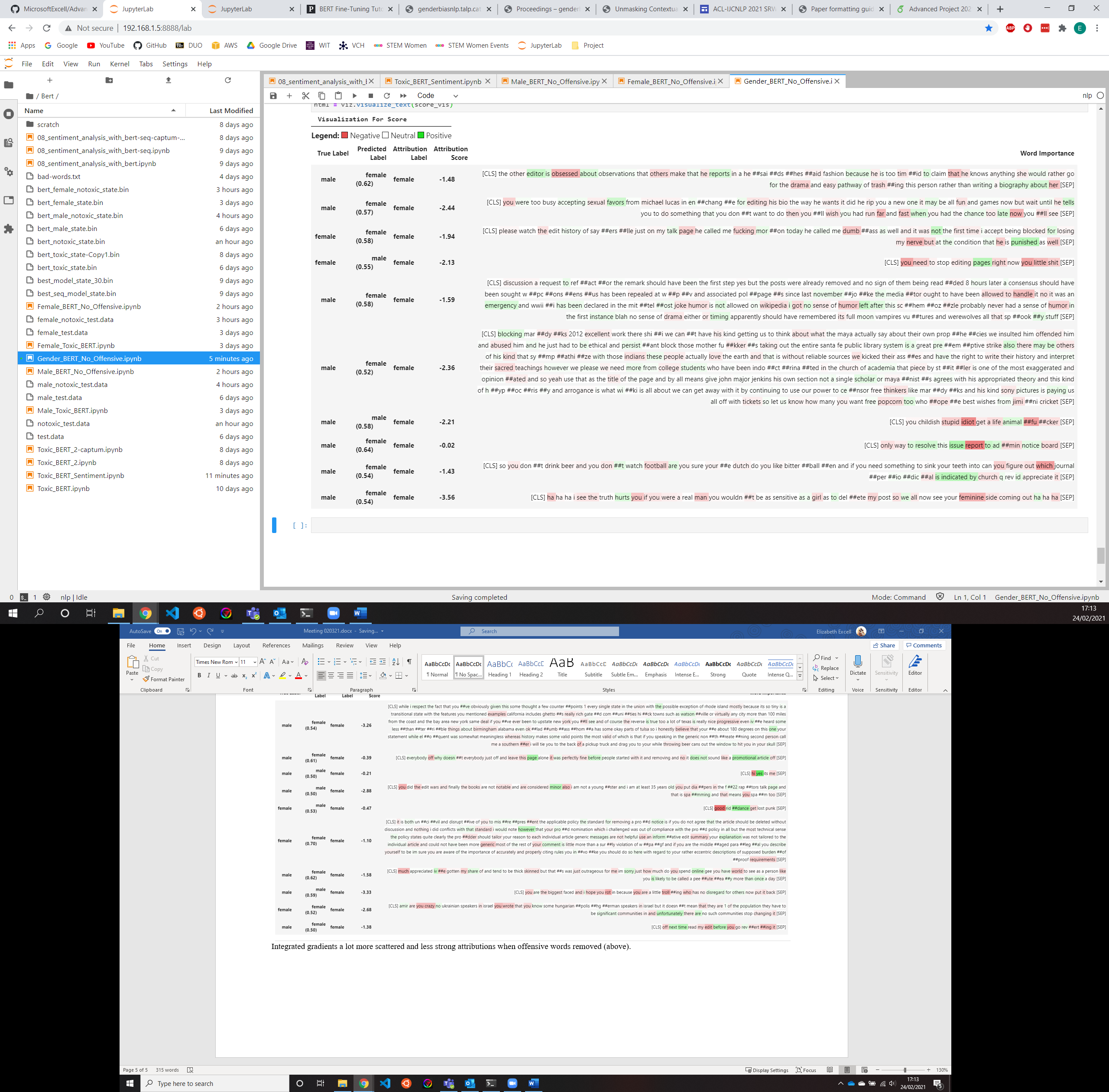


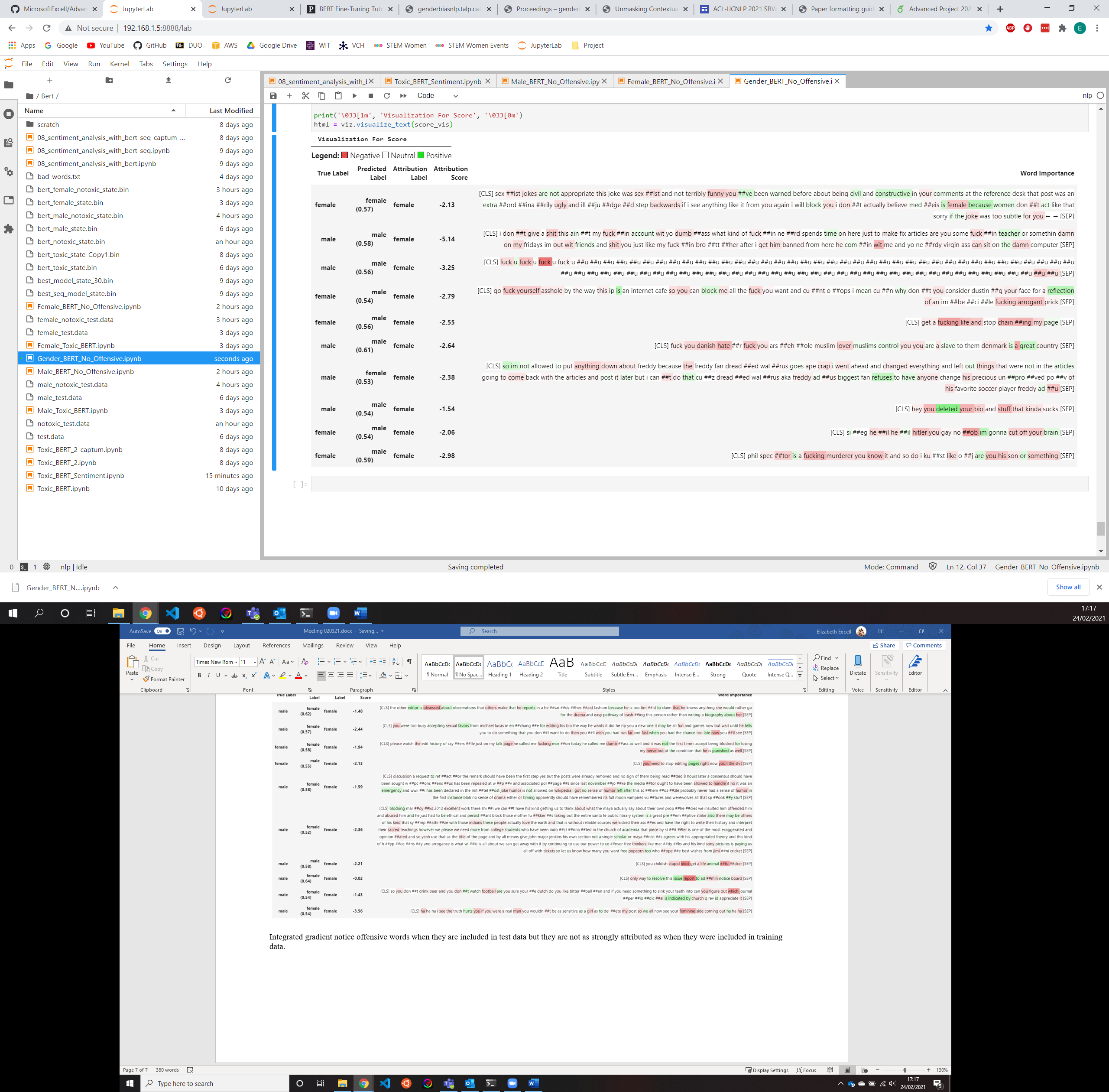
Female predictions spread over much larger range than male predictions for all toxic, not the case for just toxic (below). May just differ between runs.





Integrated gradients a lot more scattered and less strong attributions when offensive words removed from test and training data (above).





Integrated gradient notice offensive words when they are included in test data but they are not as strongly attributed as when they were included in training data.