**Meeting 23/02/21 Notes**

* Results below found and discussed. Similar performance between models trained on male and female data (female had less training and testing data) – strong degree of agreement. Both performed better on male toxic data, female slightly outperformed male on female data (male data generalises better to female model than other way around by very slim margin?). Both models have decreased performance when offensive words omitted. Female model less resilient to omission of toxic words. Very toxic comments have more offensive words. Similar “very toxic” word clouds for both genders, “toxic” word clouds slightly different – “die” and “vagina” more prevalent in women’s. False negative data had far fewer swear words but quite a few negative words – toned down language is reason behind misclassification? Classifier needs to take context and semantics into account more to correctly classify false negatives? True positive data full of swear words. Very toxic data had more offensive words per comment than toxic data. False negative data had fewer offensive words per comment.
* Thoughts: When offensive words are removed, the sentence may no longer be toxic so female model may be better at reading neutrality of sentences?
* Main difference between male and female word clouds in toxic class (not very toxic class) so may be more interesting to investigate differences in just toxic class? (For very toxic class, just a search algorithm looking for offensive words could perform well?)
* Results very informative and should be included in paper, shows some assumptions were wrong.
* For next week: train male and female models again, excluding offensive words from toxic comments, conclude which is then more accurate and reads into meaning/context, conclude if toxic words were key factor (see if maintains high accuracy without). Repeat gender classification.
* Paper: upload on overleaf and share, follow conference structure. At beginning, write everything (can be used later in project paper). Sign up using durham account for all overleaf perks. To begin, add template and literature review, start revising and expanding.

**This week’s results:**

In Male\_Toxic\_BERT.ipynb and Female\_Toxic\_BERT.ipynb:

Note: All toxic = toxic + very toxic

**Confusion matrices**

|  |  |  |
| --- | --- | --- |
| Testing Data\Training Data | Male (Train size: 69,964) | Female (Train size: 45,044) |
| Male All Toxic  (Test size: 8,744) |  |  |
| Female All Toxic  (Test size: 5,630) |  |  |
| Male All Toxic without offensive words |  |  |
| Female All Toxic without offensive words |  |  |

**Word Clouds and counts of top offensive words in Test Data**

|  |  |  |
| --- | --- | --- |
| Test Data\Train Data | Male | Female |
| Male Toxic |  | (same test data as left so same word cloud) |
| Male Very Toxic |  | (same test data as left so same word cloud) |
| Female Toxic |  | (same test data as left so same word cloud) |
| Female Very Toxic |  | (same test data as left so same word cloud) |
| Male All Toxic - True Positive (Predicted Toxic) |  |  |
| Male All Toxic – False Negative (Predicted Not Toxic) |  |  |
| Female All Toxic - True Positive (Predicted Toxic) |  |  |
| Female All Toxic – False Negative (Predicted Not Toxic) |  |  |

**Number of Offensive words per comment in Test Data (+ normalised by length of comment)**

|  |  |  |
| --- | --- | --- |
| Test Data\Train Data | Male | Female |
| Male Toxic |  | (same test data as left so same results) |
| Male Very Toxic |  | (same test data as left so same results) |
| Female Toxic |  | (same test data as left so same results) |
| Female Very Toxic |  | (same test data as left so same results) |
| Male All Toxic - True Positive (Predicted Toxic) |  |  |
| Male All Toxic – False Negative (Predicted Not Toxic) |  |  |
| Female All Toxic - True Positive (Predicted Toxic) |  |  |
| Female All Toxic – False Negative (Predicted Not Toxic) |  |  |