**Meeting 09/03/21 Notes**

* Results below found and discussed.
* Toxicity Results: Definitely trend that comments with more offensive words are given higher toxicity probabilities by model and are truly more toxic. Non-toxic sentences with a high number of offensive words were often misclassified. Male test data had a stronger positive correlation between the number of offensive words and the predicted toxicity probability than female test data. Predicted labels showed slightly stronger positive correlation than true labels (indicates model was equating toxicity with number of offensive words). Models trained without offensive words had stronger positive correlations.
* Gender Results: True labels all had very slight negative correlation showing women slightly less prone to annotating comments with lots of offensive words as toxic than men. Reasonably large difference between correlations for true and predicted labels meaning model learnt that toxic comments with lots of offensive words were annotated by men and propagated this bias. This is also confirmed by the graphs where the true labels look like a fairly even distribution between men and women and the predicted labels are heavily skewed towards the men for comments with higher numbers of offensive words. Predicted probabilities show negative correlation between female annotators and lots of offensive words in comments (strongest for training without offensive words, weakest for without offensive words on only toxic data).
* Suggested next step would be training new model only using words that were given high gradients by the explainability addition. (Unlikely to accomplish due to time constraints).
* For next week: Focus on writing paper. For conference paper – cut down, cut out results and add as supplementary material. For project paper – extend based on mark scheme, justify everything I’ve done and show understanding. (Can include paper as supplementary material for project?). Supervisor will suggest small measures to improve paper/add to models/improve validation of hypotheses. In the experiments section of the paper include standard deviation with figures, explain steps taken to ensure not just lucky run such as including random sampling of test set and training set. Add a footnote that the code and the data will be available on github after the paper has been accepted. Can include fact that talk given in NLP lecture in project paper. If needed, can have extra meetings next week to go over talk in lecture.

**This week’s results:**

Added correlation information between the number of toxic words and the true/predicted labels to various notebooks to gain extra insight.

(Removed largest outliers with most toxic words in sentence).

**Toxicity Prediction - Graphs**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Toxicity Probability** | **Predicted Label** |
| Male with offensive words\Male (toxic and nontoxic) |  |  |  |
| Male with offensive words\Female (toxic and nontoxic) |  |  |  |
| Female with offensive words\Male (toxic and nontoxic) |  |  |  |
| Female with offensive words\Female (toxic and nontoxic) |  |  |  |
| Male without offensive words\Male (toxic and nontoxic) |  |  |  |
| Male without offensive words\Female (toxic and nontoxic) |  |  |  |
| Female without offensive words\Male (toxic and nontoxic) |  |  |  |
| Female without offensive words\Female (toxic and nontoxic) |  |  |  |

**Pearson Correlation – Can largely ignore as nonlinear relationship.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Toxicity Probability** | **Predicted Label** |
| Male with offensive words\Male (toxic and nontoxic) | 0.232 | 0.266 | 0.25 |
| Male with offensive words\Female (toxic and nontoxic) | 0.216 | 0.268 | 0.251 |
| Female with offensive words\Male (toxic and nontoxic) | 0.23 | 0.275 | 0.25 |
| Female with offensive words\Female (toxic and nontoxic) | 0.227 | 0.289 | 0.263 |
| Male without offensive words\Male (toxic and nontoxic) | 0.253 | 0.281 | 0.267 |
| Male without offensive words\Female (toxic and nontoxic) | 0.252 | 0.291 | 0.274 |
| Female without offensive words\Male (toxic and nontoxic) | 0.248 | 0.275 | 0.255 |
| Female without offensive words\Female (toxic and nontoxic) | 0.251 | 0.292 | 0.271 |

**Spearman Correlation**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Toxicity Probability** | **Predicted Label** |
| Male with offensive words\Male (toxic and nontoxic) | 0.464 | 0.564 | 0.511 |
| Male with offensive words\Female (toxic and nontoxic) | 0.433 | 0.545 | 0.496 |
| Female with offensive words\Male (toxic and nontoxic) | 0.464 | 0.565 | 0.509 |
| Female with offensive words\Female (toxic and nontoxic) | 0.447 | 0.557 | 0.512 |
| Male without offensive words\Male (toxic and nontoxic) | 0.583 | 0.674 | 0.614 |
| Male without offensive words\Female (toxic and nontoxic) | 0.531 | 0.651 | 0.588 |
| Female without offensive words\Male (toxic and nontoxic) | 0.578 | 0.673 | 0.593 |
| Female without offensive words\Female (toxic and nontoxic) | 0.533 | 0.653 | 0.586 |

**Gender Prediction – Graphs**

**Note: 0 is male, 1 is female**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Gender Probability** | **Predicted Label** |
| Male and Female, toxic and very toxic, **with** offensive words\Male and Female, toxic and very toxic |  |  |  |
| Male and Female, toxic and very toxic, **without** offensive words\ Male and Female, toxic and very toxic |  |  |  |
| Male and Female**, only toxic, without** offensive words\Male and female, only toxic |  |  |  |

**Pearson**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Gender Probability** | **Predicted Label** |
| Male and Female, toxic and very toxic, **with** offensive words\Male and Female, toxic and very toxic | -0.015 | -0.177 | -0.169 |
| Male and Female, toxic and very toxic, **without** offensive words\ Male and Female, toxic and very toxic | -0.024 | -0.19 | -0.17 |
| Male and Female**, only toxic, without** offensive words\Male and female, only toxic | -0.01 | -0.13 | -0.123 |

**Spearman**

|  |  |  |  |
| --- | --- | --- | --- |
| **Training\Test Data** | **True Label** | **Predicted Gender Probability** | **Predicted Label** |
| Male and Female, toxic and very toxic, **with** offensive words\Male and Female, toxic and very toxic | -0.041 | -0.378 | -0.35 |
| Male and Female, toxic and very toxic, **without** offensive words\ Male and Female, toxic and very toxic | -0.047 | -0.415 | -0.36 |
| Male and Female**, only toxic, without** offensive words\Male and female, only toxic | -0.024 | -0.236 | -0.207 |