**(Some) Results**

Sensitivity for toxicity classification across different runs:

seed\_value = 42

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.838 | 0.685 | 0.769 | 0.647 |
| MaleNoProfanity | 0.847 | 0.818 | 0.787 | 0.759 |
| Female | 0.855 | 0.706 | 0.800 | 0.667 |
| FemaleNoProfanity | **0.860** | **0.832** | **0.801** | **0.775** |

seed\_value = 9743

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.837 | 0.674 | 0.779 | 0.625 |
| MaleNoProfanity | 0.837 | 0.814 | 0.777 | 0.753 |
| Female | **0.856** | 0.681 | **0.809** | 0.638 |
| FemaleNoProfanity | 0.849 | **0.815** | 0.791 | **0.760** |

seed\_value = 5936

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | **0.866** | 0.718 | **0.793** | 0.661 |
| MaleNoProfanity | 0.841 | **0.818** | 0.774 | 0.748 |
| Female | 0.846 | 0.673 | 0.776 | 0.621 |
| FemaleNoProfanity | 0.847 | 0.816 | 0.784 | **0.753** |

seed\_value = 14280

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.806 | 0.637 | 0.737 | 0.588 |
| MaleNoProfanity | 0.835 | 0.810 | 0.774 | 0.752 |
| Female | 0.856 | 0.716 | **0.814** | 0.678 |
| FemaleNoProfanity | **0.859** | **0.829** | 0.810 | **0.785** |

seed\_value = 29988

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.838 | 0.683 | 0.763 | 0.623 |
| MaleNoProfanity | 0.836 | 0.811 | 0.762 | 0.739 |
| Female | **0.854** | 0.700 | **0.794** | 0.646 |
| FemaleNoProfanity | 0.849 | **0.820** | 0.786 | **0.758** |

**Means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.837 | 0.6794 | 0.7682 | 0.6288 |
| MaleNoProfanity | 0.8392 | 0.8142 | 0.7748 | 0.7502 |
| Female | **0.8534** | 0.6952 | **0.7986** | 0.65 |
| FemaleNoProfanity | 0.8528 | **0.8224** | 0.7944 | **0.7662** |

**Standard Deviations**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model\Test Data | Male | MaleNoProfanity | Female | FemaleNoProfanity |
| Male | 0.018995 | 0.025928 | 0.018616 | 0.024823 |
| MaleNoProfanity | 0.004400 | 0.003370 | 0.007985 | 0.006615 |
| Female | 0.003774 | 0.015917 | 0.013260 | 0.020367 |
| FemaleNoProfanity | 0.005528 | 0.006888 | 0.009769 | 0.011923 |

Sensitivity and specificity for gender classification across different runs:

seed\_value = 42

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.385 | 0.695 |
| NoProfanity | 0.380 | 0.688 |
| NotVeryToxic | **0.490** | **0.560** |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | **0.493** | 0.565 |
| NoProfanity | 0.429 | 0.628 |
| NotVeryToxic | 0.515 | **0.532** |

seed\_value = 9743

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.358 | 0.721 |
| NoProfanity | 0.413 | 0.647 |
| NotVeryToxic | **0.507** | **0.554** |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.457 | 0.612 |
| NoProfanity | 0.424 | 0.600 |
| NotVeryToxic | **0.518** | **0.538** |

seed\_value = 5936

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.324 | 0.748 |
| NoProfanity | 0.448 | 0.605 |
| NotVeryToxic | **0.454** | **0.581** |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.457 | 0.666 |
| NoProfanity | **0.496** | 0.550 |
| NotVeryToxic | 0.510 | **0.521** |

seed\_value = 14280

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.355 | 0.728 |
| NoProfanity | 0.519 | **0.535** |
| NotVeryToxic | **0.508** | 0.536 |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | **0.494** | 0.595 |
| NoProfanity | 0.564 | **0.484** |
| NotVeryToxic | 0.519 | 0.518 |

seed\_value = 29988

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.379 | 0.679 |
| NoProfanity | **0.492** | 0.564 |
| NotVeryToxic | 0.512 | **0.530** |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.427 | 0.622 |
| NoProfanity | **0.535** | 0**.513** |
| NotVeryToxic | 0.552 | 0.479 |

**Means**

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.356 | 0.714 |
| NoProfanity | 0.450 | 0.608 |
| NotVeryToxic | **0.494** | **0.552** |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.466 | 0.612 |
| NoProfanity | **0.490** | 0.555 |
| NotVeryToxic | 0.523 | **0.516** |

**Standard Deviations**

Testing with offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.025 | 0.024 |
| NoProfanity | 0.051 | 0.055 |
| NotVeryToxic | 0.021 | 0.018 |

Testing without offensive words

|  |  |  |
| --- | --- | --- |
| Model | Sensitivity | Specificity |
| Original | 0.025 | 0.033 |
| NoProfanity | 0.056 | 0.053 |
| NotVeryToxic | 0.015 | 0.027 |