

| | | | | |
|-------|-----|-------|------|------|
| Round | R1 | 54.7 | 64.6 | 49.2 |
| | R2 | 34.3 | 38.9 | 29.7 |
| | R3 | 27.8 | 20.5 | 35.1 |
| | R4 | 27.7 | 23.7 | 31.7 |
| | | Total | Not | Hate |
| Round | All | 36.6 | 35.4 | 37.7 |
| | | Total | Not | Hate |

(a) Original results

| | | | | |
|-------|-----|-------|------|------|
| Round | R1 | 64.7 | 75.1 | 58.9 |
| | R2 | 37.3 | 31.7 | 42.8 |
| | R3 | 26.1 | 19.5 | 32.8 |
| | R4 | 26.9 | 29.6 | 24.2 |
| | | Total | Not | Hate |
| Round | All | 39.4 | 37 | 41.5 |
| | | Total | Not | Hate |

(b) Replicated results

Figure 2: Experiment 2 - Model Error rate of each model on its respective dataset. The first column shows the aggregated error rate while the other two columns show the error rates for each class on its own. The last row shows the mean error rates of the four models.