Distribution of Toxicity Scores in Filtered Data 1 0 -**Toxicity Score Toxicity Class** anti-toxic neutral toxic -2 anti-toxic neutral toxic **Toxicity Class**