

Distribution of Toxicity Scores in Filtered Data

Toxicity Class anti-toxic neutral toxic

