Visualizing Toxicity of Location Subreddit

J. Schroeder, J.M. Abbess IV, N. Denu, S.U. Rao, S.S. Chakraborty, S.S. Kamble

Summary

Monitoring the prevalence of toxic speech in online communities is a proven strategy in curbing toxicity online. Online discussion in places like Reddit are shown to have real world impacts on people and shape their opinions. Online harassment can also **silence voices** and **drive people away** from the conversation, forcing already marginalized people out of the online commons.

We **collected data** and **analyzed comments** from state-wide subreddits in the USA to provide toxicity scores and break them down into **5 categories** of toxicity. Our dashboard is aimed at easing the moderation of communities and **creating a more welcoming environment** by appealing to the competitive spirit of Reddit users across locations.

Methodology

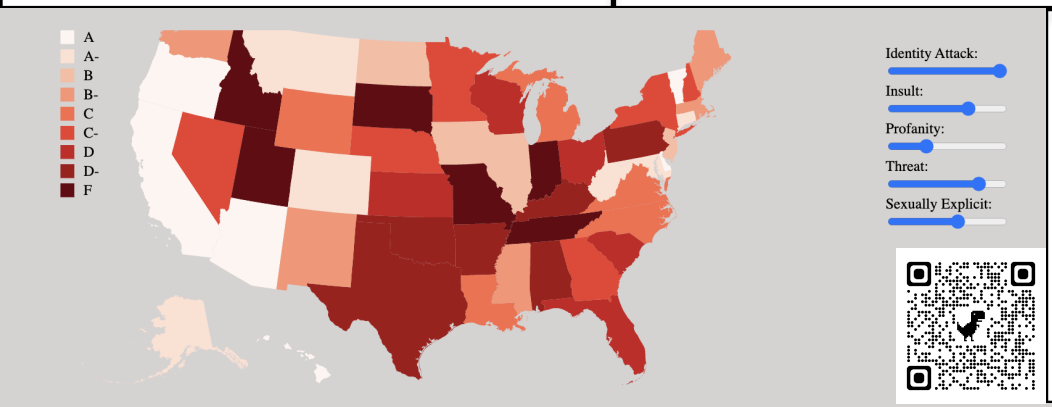
NEW

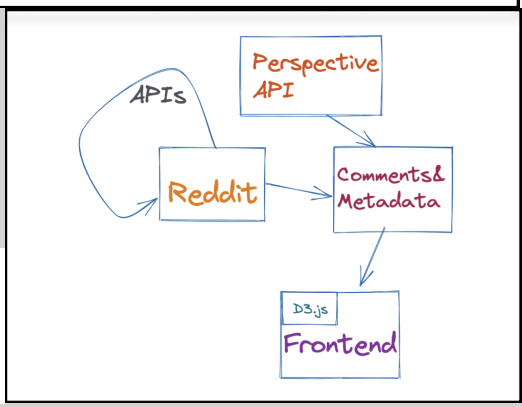
We **scraped** data for location subreddits using the Reddit APIs and experime with toxicity evaluation APIs like hugging face, before choosing Perspective API for its breakdown of its toxicity scores. Though we were contrained by queries per second, after inspecting its output, it was seen to outperform other solutions.

To **measure** the toxicity of a subreddit, we consider all its comments and their toxicities. While we initially considering aggregating their means, we decided to pursue a weighted means strategy where the toxicity of a comment is weighed by its upvote score. This implies, if a toxic comment appears, but the rest of the community heavily downvotes it, that does not count towards the toxicity of the subreddit, and vice versa.

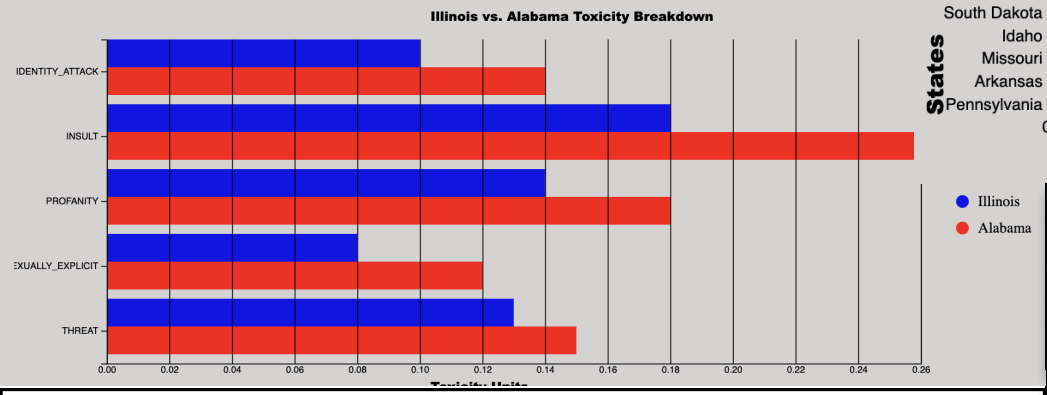
We **innovated** by using a choropleth with shades of red to indicate toxicity gradient and allow the user to assign weights to individual toxicity metrics.

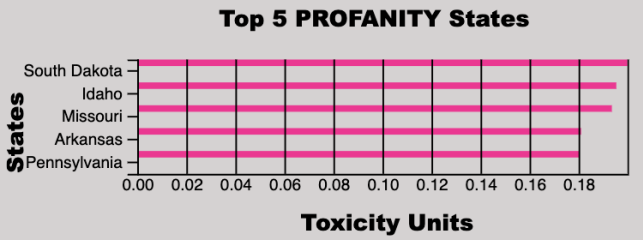
For **evaluation**, we compared Perspective API and huggingface outputs and noticed no significant differences. We also observed that toxic subreddits do vary in metrics (as can be observed by tweaking the weights of the breakdown, and some states show overall more toxic behavior than others.





CLICK ON ANY STATE TO SHOW DETAILED BREAKDOWN OF TOXICITY





What is Toxicity Score?

$$tx(sub) = \frac{\sum_{c \in comments}(tox(comment_c) * score_c)}{\sum_{c \in comments}\sqrt{score_c}}$$

Data

795 location subreddits; each >800 comments+metadata per week; ~400MB on disk per week

Body	Score	TOXICITY
F[redacted] off,liar	1	0.98
What a geriatric f[redacted]	-6	0.96
There's a lot of really stupid c[redacted]s in this state.	10	0.96

Location	TOXICITY	IDENTITY	INSULT	PROFANITY	THREAT	SEXUAL
South Da	0.32	0.18	0.3	0.25	0.19	0.14
Idaho	0.31	0.2	0.29	0.21	0.18	0.13
Missouri	0.29	0.17	0.27	0.18	0.15	0.15
Alabama	0.29	0.15	0.28	0.2	0.16	0.13

Results

We provide the toxicity grades of state subreddits in an easy to consume, interactive format. This will help identify problematic areas and provide insights for moderation of the subreddits. The user can also customize the dashboard by changing weights of categories and compare two states to understand trends. We also provide a series of "Top 5" graphs across different categories for the user to further gain perspective. It is worthwhile to note that the toxicity score of a state is based on the Redditors active on that location's subreddit and may not be indicative of the state's general populace.