Comparative analysis of the Perspective API and community-based classification system on Reddit comments

Haji Mohammad Saleem & Derek Ruths
Network Dynamics Lab
School of Computer Science
McGill University

Abstract

In our previous work, we developed a community-based classification system that used the comments in self-identifying hateful communities to train target-specific hate speech classifiers. In this analysis, we compare the performance of our classification system against that of the Perspective API, on a set of Reddit comments. We find that of toxicity (as operationalized in the Perspective API) and hate speech, are similar yet distinct concepts. The two systems surface different behaviors and combined results of the two can increase the quality of our community-based *silver-standard* hate speech dataset.

Data

For this analysis, we used Reddit comments from three communities that self-identify as hateful towards a group of people. Specifically:

- `r/coontown`: a racist subreddit [now banned]
- `r/fatpeoplehate`: a body-shaming subreddit [now banned]
- `r/theredpill`: a misogynistic subreddit

We collected large number of user comments from each subreddit. Further details in Table 1.

Methods

Perspective API

Post cleanup of deleted comments, we ran the Perspective API on about 50,000 random comments from each subreddit (exact numbers in the Table 2). We had to take a subset since the API is rate-limited to 1,000 comments per 100 seconds.

Subreddit	# of Comments
`coontown`	350,851
`fatpeoplehate`	1,577,681
`theredpill`	51,504

Table 1: Number of Reddit comments collected from each subreddit.

Subreddi	# of Comments	Avg Score
`coontown`	50,000	0.391
`fatpeoplehate`	50,000	0.316
`theredpill`	43,000	0.300

Table 2: Number of Reddit comments analyzed using the Perspective API.

While using the API, we came across two major issues and had to discard a few comments:

- The API is limited to the analysis of comments with at most 3000 characters. Reddit comments can exceed that.
- Comments using slangs or malformed words were identified as belonging to languages other than English.

Community-based classifier

We performed supervised binary classification with logistic regression. Comments from the three subreddits were automatically labelled as positive. An equal number of random Reddit comments were used as the negative sample. Since we do not manually label the data and use the subreddits as proxy, we call it a *silver-standard* dataset. The learning task included 10-fold cross-validation on a shuffled *silver-standard* dataset. The dataset size (positive & negative) for each subreddit is provided in Table 3, and the results of the classifier across the 10 folds is provided in Table 4.

Subreddit	# of Comments
`coontown`	615,682
`fatpeoplehate`	795,197
`theredpill`	87,900

Table 3: Reddit comments in the silver-standard dataset, for the community-based classifier.

`coontown`					`fatpeoplehate`					`theredpill`				
Acc	Pre	Rec	F1	K	Acc	Pre	Rec	F1	K	Acc	Pre	Rec	F1	K
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.78	0.58	0.82	0.85	0.76	0.8	0.63
0.82	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.78	0.59	0.81	0.83	0.77	0.8	0.62
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.78	0.58	0.81	0.84	0.74	0.79	0.61
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.79	0.59	0.81	0.84	0.76	0.8	0.63
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.78	0.58	0.81	0.85	0.75	0.8	0.63
0.81	0.87	0.74	0.8	0.62	0.79	0.82	0.75	0.78	0.58	0.82	0.86	0.75	0.8	0.64
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.76	0.79	0.59	0.81	0.85	0.75	0.8	0.62
0.81	0.87	0.74	0.8	0.62	0.79	0.82	0.75	0.79	0.59	0.81	0.84	0.76	0.8	0.63
0.81	0.87	0.74	0.8	0.63	0.79	0.82	0.75	0.78	0.58	0.81	0.84	0.76	0.8	0.62
0.81	0.87	0.74	0.8	0.62	0.79	0.82	0.74	0.78	0.58	0.82	0.85	0.76	0.8	0.63
0.81	**0.87**	**0.74**	**0.8**	**0.63**	**0.79**	**0.82**	**0.75**	**0.78**	**0.58**	**0.81**	**0.85**	**0.76**	**0.8**	**0.63**

Table 4: Performance measures of the community-based classifier over 10 folds. It is worth noting the consistency among the folds, which signifies that the classifier is stable across the folds in identifying hate speech.

Quantitative Analysis

After running the two methods, we binned the results into six categories based on the combined results. The binning schematics are provided in Figure 1.

	Toxicity Score		
	Low	Mid	High
Hate Speech	Bin 1	Bin 2	Bin 3
Not Hate Speech	Bin 4	Bin 5	Bin 6

Figure 1: The binning schematics for the combined results of the community-based classifier and the Perspective API. For the Perspective scores: low toxicity: 0 - 0.33, mid toxicity: 0.33 - 0.66, high toxicity: 0.66 – 1.

Qualitative Analysis

For qualitative analysis, we manually analyzed the contents of each bin, to get an idea of where the two methods converged and diverged. We also manually labelled 50 random comments from each bin to further ascertain each bins composition.

What is the size of individual bins?

Results of Quantitative Analysis

In Table 5 we present the size of each bin, relative to the total number of analyzed comments. Figure 2 presents these results as a heat map. Across all three subreddits we find that bin 1 constitutes the largest portion of analyzed comments while bin 5 and 6 together represent less than 6% of the comments.

	Toxicity Score		
	Low	Mid	High
coontown		47,260	comments
Hate Speech	30.02%	18.58%	25.68%
Not Hate Speech	20.10%	3.14%	2.48%
fatpeoplehate		13,576	comments
Hate Speech	41.38%	17.02%	17.04%
Not Hate Speech	18.70%	3.46%	2.39%
theredpill		42,465	comments
Hate Speech	45.16%	15.81%	14.35%
Not Hate Speech	19.18%	3.20%	2.30%

Table 5: The size of individual bins for each set of Reddit comments. Note the values represent the percentage of overall comments that constitute a particular bin.

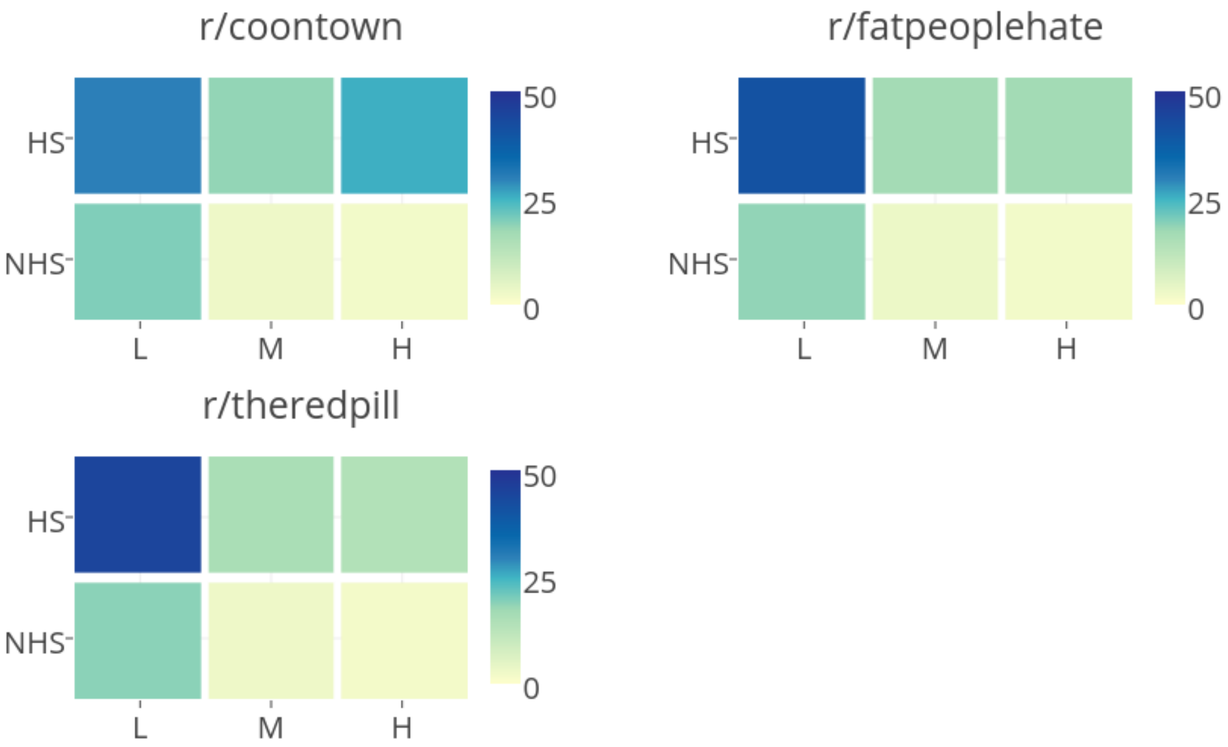

Figure 2: Heat map representing the percentage portion of individual bins for each set of Reddit comments.

What do the individual bins represent?

Results of Qualitative Analysis | WARNING: Graphic Language

For a better understanding of how the two methods work together, we performed a detailed analysis of the contents of the individual bins for `r/coontown`. Furthermore, to get a view of the overall composition, we hand-labelled 50 random comments from each bin as hate speech or not.

Bin 1: hate speech, low toxicity

Manual analysis revealed that bin 1 did contain hateful comments. However, the comments were subtler and would require context/expert knowledge to understand their hateful nature. A few examples:

- I got her pegged as a coon lover.
- Previous chimp free zone was being run by a pole smoker.
- St Dindu, dying on the cross... blessed be his name for he dwelleth at the right have of the (potential) father along with the The Skittles Kid.
- Ebola was a god send.
- The court needed a translator; I can't speak coon
- They are animals. No humans can compare.
- We don't tolerate nigging in Arizona

- Abortion in the black coonmunity has been a wonderful thing. We need to take it to the next level and start getting them fixed at an early age.

Overall, while low in *toxicity*, Bin 1 did contain a high amount of hate speech with 33 of the 50 comments labelled as such.

Bin 2: hate speech, mid toxicity

In bin 2, the hateful comments seemed to have more keywords present.

A few examples:

- To justify nigger on human being rapes. What kind of parallel universe does a woman wear GLOVES + carry a handbag like that in an area where it's socially acceptable for a nigger to be outside with no shirt on?
- Well, empirical evidence plus my own observations (I live in a chimp heavy town) have led me to reach certain conclusions about niggers. Other groups are otherwise alright by me. (even Jews, who are a pretty hated minority on this sub.)
- What are you doing back in coon town? I thought you died 3 months ago. Are you spreading the word of Jesus again?
- Got accepted into NU! Good for that nigger and his family.
- I'd go on that ride. I've never seen nigs in their natural habitat.
- Are negro populations growing in Colombia?

Overall, bin 2 had an enormous presence of hateful content. 40 of the 50 comments were labelled as hate speech. The presence of the term ‘nigger’ became more prevalent with 14 of the comments mentioning it.

Bin 3: hate speech, high toxicity

In bin 3, we see some of the most disturbing content.

A few examples:

- Most blacks are usually niggers and religious. So double trouble.
- I live in a nigger infested city. Sadly, I'm bilingual at this point. I can understand about 3/4 of what dribbles out of their mouths.
- Uh. Niggers. Duh.
- Got served some lead too. Eat up nigger
- If this happened to all white liberal dumbasses who still think chimps are human, the coddling of these fucking apes would come to a screeching halt.
- Goddam kikes.
- U mad faggot?
- When niggers run a city, you can be sure a corruption scandal will follow.

Bin 3 contained an overwhelming amount of vile and evident hate speech. 46 of the 50 comments were labelled as hateful, with 30 of them containing the term ‘nigger’.

Bin 4 & 5: not hate speech, low and mid toxicity

Bin 4 and 5 were the bins for non-hateful comments.

A few examples:

- I was waiting for this to show up.
- This is one very, very brave man. Assuming he's real.
- Thanks man. Creating a new subreddit, just as forking an opensource project, is seldom a best idea :-)

- whats this blog about? got any recommendation/intro posts?
- It's sad that this looks like such a crazy statement, because it really isn't if you've dealt with these girls. Big part of what made me leave atheism, I eventually realized the secular religion was incomprehensibly more grotesque than some of the theological ones.
- Thank you for finding and reposting here.
- Uh... Forks up, I guess? Terrible call, ASU
- It's the end of the day at work, nothing to do so I'm killing my last 30 minutes.

Bin 4 and 5 comes across as random community chatter.

While very infrequent, they did contain a few hateful comments.

- whatever jew lover
- 2 nig r na 2 nig dat b da qwesten
- I'll bring the nig nog!
- Not to mention how much they already kill each other

A vast majority of the comments were benign with only 7 of the 100 comments were labelled as hateful.

Bin 6: not hate speech, high toxicity

While bin 6 did not seem to contain hate speech, it did contain toxic language.

A few examples:

- Yeah they fuckin line up for laptops at downtown library every morning. The lowest of the low.
- Well fuck the EU. But they should join NATO still..
- She's looking at its butt.
- Will you please show us your boobs?
- She gon' kill me!
- This cunt is finally realizing that people only cared about her when she was giving out free shit. Now, she is just another talking head getting paid too much money to wear too much make-up. She needs to disappear already.

Out of the 50 comments, only 4 were labelled as hateful.

Conclusions / Discussion

1. *Toxicity and hate speech are similar yet distinct concepts.*

As evident from the content of the six bins, we can say that hate speech and comment toxicity, while similar, are concepts that are not exactly alike. Content analysis of the bins provided us with examples of hate speech deemed not toxic by the Perspective API (in bin 1) and examples of toxic / objectionable speech that is not necessarily hate speech (bin 6). For our research, the former, that includes more subtle forms of hate speech, is an important and challenging form of hate speech to capture. In the future we would like to study and analyze is much of this content is actually toxic.

2. *The Perspective API is better at finding content that is overall objectionable content.*

The community-based classifier is trained to identify hate-speech towards a specific community, and therefore will miss out on obviously hateful content that target communities other than the one the classifier was trained on. However, the Perspective API, being target independent, labels such comments as toxic. For example, the misogynistic speech examples in Bin 6:

This cunt is finally realizing that people only cared about her when she was giving out free shit. Now, she is just another talking head getting paid too much money to wear too much make-up. She needs to disappear already.

3. *Bin 2 and 3 are strong examples of hateful speech.*

Bin 2 and 3 contain a significant amount of hateful content (40/50 & 46/50 respectively in manual annotation) and constitute high confidence examples of hate speech. However this confidence is the result of prevalence of common slurs, in this case ‘nigger’, in these two bins. While being the source of high-quality training data, just using the two bins can lead to a training bias towards the common slurs. In contrast, bin 1, the largest bin, contains hate speech with fewer slurs / keywords. Going forward, it would be interesting to see how the classification system perform when using just bin 1 as the training source.

4. *Bin 4 and 5 are examples of non-hateful and non-offensive content in a hateful subreddit.*

Bin 4 and 5 contain comments that come across as mostly as community chatter with very few examples of actual hate speech. Removing the two bins can increase the quality of our silver-standard dataset.