Orange: Capstone Project

Research questions and Hypotheses

Data:
Reddit comments: 206,606 Reddit comments that include at least one of the hate terms (from our hate term lexicon based on the Davidson research and amended with terms from other research) and was posted during January 2022. Each comment includes 35 data points. We removed comments posted by bots, duplicate comments, non-English comments, and frequent promotional

Reddit reference comments: A selection of 46,274 comments posted in January 2022 that was not filtered to include the hate terms. This is used as a comparison for the filtered Reddit comments dataset.

Reddit users: 135,708 unique users each with 12 data pints.

Reddit submissions: 187,874 submissions (the original post to which the comments were made) each with 23 data points .

Classification:
Classifiers categorized comments, submission text, and submission titles as having either hate speech, offensive language, or neither. These classifications may be compared to scores from the unbiased Detoxify model, which has measures related to toxic language (threat, toxicity, severe toxicity, obscenity, sexually

Gender:
Comments, submission text, and submission title are identified as having either male identifiers (e.g., "he", "guy"), female identifiers (e.g., "she", "gal"), both gender identifiers, or neither gender identifiers. The labels are based on a set of gender-related terms we developed.

Toxic language:
Toxic language affects people, which confirms why it is important to understand its usage on social media. This raises the issue of how to handle this language in the visualizations and text. I propose that we use the original text in the visualizations. The text will be small and people may choose not to explore the interactive visualizations which will allow them to avoid the toxic language if they choose. In the written story of our exploration, I propose that we use asterisks to fill in some letters in offensive words. This is primarily an acknowledgement that the language is toxic. An asterisk or two will not obscure the word and people will still understand its meaning and have a visceral and intellectual response to it. However, completely masking a word would render it

meaningless, which is not a good solution since the project is about the words and the power they have.

note: numbering is solely for referencing; it does not represent the importance of the questions.

Toxic speech is defined as speech with a hate and/or offensive classification or with qualifying Detoxify scores
statistics: see item 5 for t-test method for unequal samples

Area I: Hate Speech and Gendered Language

	research question	*operationalized question*	*hypothesis*
1	How is toxic speech related to gendered speech? toxic speech: * determined by Detoxify score * hate or offensive classification	What percentage of Reddit comments that are predicted to be toxic include: • male pronouns or identifiers (e.g., "guy")? • female pronouns or identifiers (e.g., "gal")? • both • not gendered speech? note: compare frequency in hate, offensive, toxic, and neutral speech for female, male, and no identifiers. method: chi-square with post-hoc chi-square comparisons with Bonferroni Adjustment compare to gendered speech in sample that is not filtered to include toxic language	A higher percentage of comments categorized as toxic speech will have female identifiers than male identifiers or no gendered identifiers. A higher percentage of comments with higher Detoxify scores will have female identifiers.

2	Do the Detoxify scores of comments differ based on the gender identified in the comments?	Do the median Detoxify scores (toxicity, severe toxicity, obscenity, threat, insult, identity attack, and sexually explicit) differ based on the gender in the comment? • male pronouns or identifiers (e.g., "guy")? • female pronouns or identifiers (e.g., "gal")? • both? • not gendered speech? method: data is non-parametric so use Kruskal-Wallis and Mann-Whitney post-hoc test	Comments with female identifiers will have higher Detoxify (unbiased model) scores.
3	How does language differ based on gendered speech?	How do the words associated with Reddit comments differ based on the presence of male or female pronouns?	Words associated with females are more likely to be passive, patronizing, or related to domesticity or appearance.
4	Is gendered toxic speech more prevalent on some subreddits than on others? compare with item 7 below ~ similar question without the gender component	Do some subreddits have a higher frequency of toxic speech mentioning females? Also examine which subreddits note: identify the subreddits with the most toxic speech with female identifiers and compare with the frequency of toxic speech with male identifiers Consider examining by comparing hate, offensive, and neither classifications as well as exploring the Detoxify scores.	Reddit has a wide variety of subreddits (micro communities). Toxic speech that includes female identifiers is more prevalent on some subreddits than on others.

5	Is toxic speech with female identifiers perceived differently than hate or offensive speech with male identifiers?
	
	similar to 8 and 9
	
	note: The controversiality score is 1.0 if the number of upvotes is similar to the number of downvotes, demonstrating a divided opinion on the comment. The controversiality score is 0.00 means that the number of upvotes is not close to the number of downvotes, indicating more agreement about the comment.
		Toxic language and controversiality
		
		Is the mean controversiality score of toxic comments with female identifiers higher (closer to 1.00 -- more controversial) than the mean controversiality score of toxic comments with male identifiers?
		
		method:
		t-test
		use Welch's t-test (aka. Satterwaite test; Smith-Welch-Satterwaite test, Aspin-Welch test, or unequal variances t-test) test for independent samples (robust for unequal sample sizes -- which we predict -- or variances)
		
		import:
		scipy stats as stats
		numpy
		
		# check variances
		total_var = np.var(all_data)
		grp1_var = np.var(data_grp1)
		grp2_var = np.var(data_grp2)
		
		# change equal_var to True if necessary based on variance results and size of samples
		
		ttest_ind(data_grp1, data_grp2, equal_var= False)
			The mean controversiality score of toxic speech with female identifiers will be higher (closer to 1.00 -- more controversial) than the mean controversiality score of toxic speech with male identifiers. Toxic speech related to females is more likely to be accepted. Also, some Redditors may not want to engage with the post at all by downvoting, even though the comment may be hidden if it receives enough downvotes.

6	How likely is it that gendered speech is related to violence?	What percentage of Reddit comments with high scores on threat (or other measures ex. sexual_explicit + threat) also have: • male pronouns or identifiers (e.g., "guy")? • female pronouns or identifiers (e.g., "gal")? • both genders? • no gendered speech?	A higher percentage of comments with female identifiers will include terms related to violence than will comments with male identifiers.
6.1	How often do the gender identifiers appear in the corpus?	For each gender identifier, what is its frequency in the corpus and its proportion of the separate totals female or male identifiers? note: could also compare to the proportion of gender identifiers included in the unfiltered sample.	A higher percentage of female identifiers will be toxic words.

Area II: Hate Speech

	research question	*operationalized question*	*hypothesis*
7	Is toxic speech more prevalent on some subreddits than on others? see 5 above ~ similar question with gender component added	Do some subreddits have a higher frequency of * hate or offensive speech? * toxic language (specific measures from Detoxify)? note: compare frequency of hate, offensive, and neutral speech in the subreddits with the most hate speech. Also compare subreddits with the most hate speech with the subreddits that have the most gendered hate speech	Reddit has a wide variety of subreddits (micro communities). Toxic speech is more prevalent on some subreddits than on others.

8	Is hate speech more controversial than offensive speech? Also examine controversiality and the various Detoxify measures. similar to 5 and 9	Is the mean controversiality score of hate speech comments higher (closer to 1.0 -- more controversial) than the mean controversiality score of offensive comments with male identifiers? method: see item 5 note: if there is not a significant difference it could help support the idea that there is not a significant difference between the hate speech and offensive speech	The mean controversiality score of hate speech will not be significantly higher (closer to 1.00 -- more controversial) than the mean controversiality score of offensive comments. The challenge of separating hate speech and offensive speech means that Redditors will not respond to them significantly differently.
9	Is toxic speech more controversial than speech that is neither hateful or offensive? note: about 17% of the Davidson corpus was identified as neither hate nor offensive speech similar to 5 and 8	Is the mean controversiality score of toxic comments higher (closer to 1.0 -- more controversial) than the mean controversiality score of non-toxic comments? Also examine Detoxify scores and controversiality. method: see item 5 note: The controversiality score is the proportion of upvotes to downvotes. A score close to 1.00 means that the number of upvotes is close to the number of downvotes, which indicates controversy about the comment.	The mean controversiality score of toxic language will be significantly higher (closer to 1.00 -- more controversial) than the mean controversiality score of non-toxic comments.
10	Which words are most common in toxic speech?	Which words appear most frequently in toxic comments?	A high percentage of words will be related to sex and sexual violence.

11	When do Redditors respond to hate speech?	What is the time gap between the original submission and the comments with toxic speech What is the time gap between the first comment with toxic speech and the next comment with toxic speech?	The time gap between the first hateful or offensive comment and the next hateful or offensive comment will be shorter than the gap between the original submission and the first hateful or offensive comment.
12	Why do Redditors respond with toxic speech?	What characteristics elicit comments with toxic speech? Examine: active users topic (title, submission, previous comments)	Topics in the title of the submission and the submission, previous comments with hate speech, and the number of active users in the subreddit will affect when toxic speech begins in the thread of comments.
13	Do Redditors use language to avoid detection by moderators or build in-group lingo?	Explore terms used in toxic speech looking for: • words with no vowels • words with non-alphabetic characters • words that are the reverse of words in a lexicon Examine the most common words in comments with hateful or offensive language.	Hate speech authors do use text anomalies (e.g., unexpected or nonstandard spelling or novel words) to attempt to avoid detection, however, it may be difficult to identify these efforts. This "secret" language may also contribute to group identity since only those within the group will at first understand the meaning and intent of the language.
14	How do particular individuals spread hate across Reddit?	How does a person identified as making comments with toxic speech post hate speech across Reddit?	Hate speech may be more concentrated in particular subreddits, but a person is likely to spread toxic speech to numerous subreddits.

15	How often and when are toxic comments removed or deleted?	What percentage of toxic comments were removed by Reddit? What percentage of toxic comments were deleted by the comments' authors? What is the time gap between when the comment was created and when it was deleted? note: compare % hateful removed to % offensive removed. (see method in 4) If there is not a significant difference, this is support for the idea that there is not a considerable difference between hateful and offensive speech If there is a meaningful number of deleted comments, compare the % hateful deleted to the % offensive deleted.	There will not be a significant difference between the percentage of comments classified as hateful that are removed by Reddit moderators and the percentage of comments classified as offensive that are removed by Reddit moderators. The percentage of comments that are deleted by the original authors will be very low.
16	Are comments or submissions more toxic?	Compare the % comments identified as hateful or offensive to the % submissions selftext classified as hateful or offensive. Also compare Detoxify scores for comments and submission selftest	Classification: Comments will have more items identified as hateful or offensive than will the text of submissions Detoxify: Comments will have higher scores on Detoxify measures than will the text of submissions

17	Do the Detoxify scores differ based on the hate/ offensive/neither classifications of the comments?	What are the relationships between Detoxify scores and hate, offensive, toxic (hate + offensive), and neither classifications? revised method: data is non-parametric so use Kruskal-Wallis then Mann-Whitney post-hoc tests.	The Detoxify scores will be lower for the neither classification than for the hate or offensive classifications. Detoxify scores will not be significantly different between the hate and offensive comments.
18	Does toxic language win some approval on Reddit?	Are submissions and comments with toxic language upvoted at a higher rate than submissions or comments without toxic language?	Comments that are more toxic will receive more upvotes.
19	Are the hate/offensive/ neither classifications and Detoxify scores related to sentiment?	Is there a relationship between the hate/offensive/neither classifications and sentiment or between the Detoxify scores and sentiment?	Sentiment scores will not be related to most of the Detoxify scores or the Davidson classification categories. Sentiment will not be sensitive enough to differentiate the foul comments

Future: network analysis, analyses examining submission text and titles