PREVALENCE AND PSYCHOLOGICAL EFFECTS OF
HATEFUL SPEECH IN ONLINE COLLEGE COMMUNITIES

Koustuv Saha, Eshwar Chandrasekharan, Munmun De Choudhury

@kous2v | koustuv.saha@gatech.edu | koustuv.com

Hate Incidents on Campus Still Rising

Results of a new survey show that administrators are still reporting incidents targeting minority students.

By Jeremy Bauer-Wolf // February 25, 2019

27 COMMENTS
HATEFUL SPEECH ON COLLEGE CAMPUSES

“..verbal and written words, and symbolic acts, that convey a grossly negative assessment of particular persons or groups based on their race, gender, ethnicity, religion, sexual orientation, or disability, which is not limited to a face-to-face confrontation or shouts from a crowd, but may also appear on T-shirts, on posters, on classroom blackboards, on student bulletin boards, in flyers and leaflets, in phone calls, etc.” (Kaplin, 1992)
ONLINE HATEFUL SPEECH
ONLINE HATEFUL SPEECH AND COLLEGE CAMPUSES

- Difficult to mitigate negative effects of online hateful speech
- Need to assess the pervasiveness of online hateful speech
- Need to understand the vulnerability in psychological wellbeing to marginalized communities on college campuses
- Present methods to do the above are limited:
  - Anecdotal / covered by popular media / discrete events
  - Social confounds, such as stigma
RESEARCH QUESTIONS

How prevalent is hateful speech in online college communities, across the demographic categories such as gender, religion, race, etc.?

How does exposure to online hate affect an individual’s expression of their psychological state on social media, particularly stress?
Online College Communities on Reddit
DATA

- 174 U.S. college subreddits
- 5,884,905 comments posted by 453,781 unique users
- 4,144,161 comments posted by 425,410 unique users who never cross-post across college subreddits
OPERATIONALIZING ONLINE HATEFUL SPEECH

Hateful speech lexicon*: 157 keywords in hate categories:
- Behavior
- Class
- Disability
- Ethnicity
- Gender
- Physical
- Race
- Religion
- Sexual Orientation
- Other

Gender: *If you disagree with us, then you are an anti-consumerist c*nt. I told you I’d call you a c*nt twice.*

Racial: *Damn n*ggah youze is just a little dude with a litte ole baby d*ck.*

Sexual Ort.: *BIG SHOT, U WANNA FIGHT? U WANNA BOX F*GGOT?*

PREVALENCE OF HATEFUL SPEECH

College Hate Index (CHX)

- Collect 1,436,766 comments from five banned subreddits (r/CoonTown, r/fatpeoplehate, r/KikeTown, r/nazi, r/transf*gs)

- CHX of an online college community to be the ratio of the normalized hate measure in the college subreddit to the same measure in banned subreddits
They admit gifted students with bright futures but produce uneducated hobos who can’t get a job and rely on State alumni for welfare.” [OTHER]
PSYCHOLOGICAL EFFECTS OF HATEFUL SPEECH

Whether and how hateful speech affects the psychological state of the online community members?

Hate exposure per user as an aggregated percentage of hateful words used by others on all the threads the user has participated in

Online stress expression, as measured using transfer learning classification on the user’s posts
ONLINE STRESS EXPRESSION

Transfer learning methodology to classify social media (Reddit) posts with contains expression of stress vs. no expression of stress  

+ r/stress  
- Non-Mental Health subreddits (r/movies, r/music, r/aww, r/Jokes, etc.)

“Feel like I am burning out (again...) Help: what do I do?”

ONLINE STRESS EXPRESSION: CLASSIFICATION

- 5000 n-gram & 3 sentiment Features
- SVM Classifier
- Accuracy: 0.82 k-fold cross-validation (k=5)
- Machine labeled stress in college dataset
- Validated on college subreddit dataset: 82% accuracy

• Conduct Observational Study
• Adopt a Causal Inference Framework based on Matching
• Use propensity score analysis
• Compare the outcomes (stress expression) of similar (matched) individuals, those exposed to a treatment (exposure to hate), and those who were not.
MATCHING FOR CAUSAL INFERENCE

Covariates:
- Social Attributes (#posts, #karma, duration in the community, frequency of posting)
- Offline and local factors (same college subreddit)
- Language and psycholinguistics (Normalized psycholinguistic occurrence of LIWC categories)
- Baseline exposure to hate and baseline stress expression

Propensity score matching for 143K Treatment and 74K Control users
DOES HATE IMPACT STRESS LEVEL?

Stress change of the Tr. users (~139%) is higher than the Ct. users (~106%). Effect size (Cohen’s d=0.40) and paired t-test shows statistically significant difference.

Individuals exposed to hate speech show greater likelihood of stress compared to their matched controls.
PSYCHOLOGICAL ENDURANCE TO HATE

Not all Treatment users are affected equally: *differences in psychological endurance*

Analyze high and low endurance users’ personality traits, and language (SAGE)
**LINGUISTIC MARKERS OF PSYCHOLOGICAL ENDURANCE**

- High endurance individuals show greater agreeableness and extraversion, whereas Low endurance individuals show greater Neuroticism and Openness
- High endurance individuals show relaxed-, leisure-, and non-academic interests
- Low endurance individuals use more classroom- and academic-related topics

Personality and Linguistic Differences (SAGE)

<table>
<thead>
<tr>
<th></th>
<th>Low Endurance</th>
<th>High Endurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>n-gram</td>
<td>SAGE</td>
<td>n-gram</td>
</tr>
<tr>
<td>education code</td>
<td>1.99</td>
<td>-1.19</td>
</tr>
<tr>
<td>prerequisite course</td>
<td>1.68</td>
<td>-1.15</td>
</tr>
<tr>
<td>general education</td>
<td>1.64</td>
<td>-1.15</td>
</tr>
<tr>
<td>classes mentioned</td>
<td>1.26</td>
<td>-1.09</td>
</tr>
<tr>
<td>mathematics</td>
<td>1.07</td>
<td>-1.00</td>
</tr>
<tr>
<td>credit course</td>
<td>1.03</td>
<td>-0.98</td>
</tr>
<tr>
<td>upper division</td>
<td>0.95</td>
<td>-0.97</td>
</tr>
<tr>
<td>prerequisite</td>
<td>0.95</td>
<td>-0.93</td>
</tr>
<tr>
<td>preassessment</td>
<td>0.89</td>
<td>-0.92</td>
</tr>
<tr>
<td>therapy</td>
<td>0.88</td>
<td>-0.90</td>
</tr>
<tr>
<td>senate</td>
<td>0.85</td>
<td>-0.89</td>
</tr>
<tr>
<td>task</td>
<td>0.84</td>
<td>-0.89</td>
</tr>
<tr>
<td>cse</td>
<td>0.83</td>
<td>-0.86</td>
</tr>
<tr>
<td>immigrants</td>
<td>0.82</td>
<td>-0.85</td>
</tr>
<tr>
<td>prereq</td>
<td>0.81</td>
<td>-0.84</td>
</tr>
<tr>
<td>cs1</td>
<td>0.79</td>
<td>-0.84</td>
</tr>
<tr>
<td>mathematical</td>
<td>0.78</td>
<td>-0.83</td>
</tr>
<tr>
<td>irrelevant</td>
<td>0.78</td>
<td>-0.83</td>
</tr>
<tr>
<td>mentor</td>
<td>0.78</td>
<td>-0.80</td>
</tr>
<tr>
<td>tasks</td>
<td>0.76</td>
<td>-0.80</td>
</tr>
<tr>
<td>division</td>
<td>0.75</td>
<td>-0.79</td>
</tr>
<tr>
<td>comptia</td>
<td>0.75</td>
<td>-0.79</td>
</tr>
<tr>
<td>assessment</td>
<td>0.73</td>
<td>-0.78</td>
</tr>
<tr>
<td>anxiety</td>
<td>0.73</td>
<td>-0.78</td>
</tr>
<tr>
<td>sql</td>
<td>0.73</td>
<td>-0.77</td>
</tr>
</tbody>
</table>
DISCUSSION
TAKEAWAYS

- Leverage social media data to understand the prevalence of hateful speech in online college communities.
- Examine the psychological effects of hateful speech in the online community members.
- Analyze the differences in personality traits and language of individuals differing in psychological endurance to hate.
SOCIO-POLITICAL AND POLICY IMPLICATIONS

- Findings help to account for a previously under-explored, but a critical facet of the speech debate, especially in the context of college campuses.
- New dimensions to the college speech debate centering around legal, ownership, and governance issues.
- New policy challenge: How to decipher when online and offline hate reinforce each other.
- Opens up discussions about “counterspeech” to tackle the psychological effects of hateful speech.
TECHNOLOGICAL IMPLICATIONS

- Mental health support provisions on college campuses:
  - Empirical information to guide informed decisions and tailored interventions
  - Proactive detection of the psychological ramifications of online hate

- Moderation efforts in online college communities
  - Call-to-action to moderators to adopt measures beyond blanket banning/censorship
  - Provide assistance and support via peer-matching and redirection to resources
ETHICAL IMPLICATIONS

- Any decision and policy making based on our findings requires careful and in-depth supplemental ethical analysis.

- Caution against surveillance of student speech on college campuses, or as a guideline to censor speech on campus.

- Not intended to be used to inadvertently marginalize or influence prejudice against those groups who are already marginalized.
Prevalence and Psychological Effects of Hateful Speech in Online College Communities


Thank You

@kous2v | koustuv.saha@gatech.edu | koustuv.com