Prevalence and Psychological Effects of Hateful Speech in Online College Communities

Koustuv Saha
Georgia Tech
koustuv.saha@gatech.edu

Eshwar Chandrasekharan
Georgia Tech
eshwar3@gatech.edu

Munmun De Choudhury
Georgia Tech
munmund@gatech.edu

ABSTRACT

Background. Hateful speech bears negative repercussions and is particularly damaging in college communities. The efforts to regulate hateful speech on college campuses pose vexing socio-political problems, and the interventions to mitigate the effects require evaluating the pervasiveness of the phenomenon on campuses as well the impacts on students' psychological state.

Data and Methods. Given the growing use of social media among college students, we target the above issues by studying the online aspect of hateful speech in a dataset of 6 million Reddit comments shared in 174 college communities. We devise a measure of College Hate Index (CHX) and examine its distribution in college subreddits across the categories of hateful speech, *behavior, class, disability, ethnicity, gender, physical appearance, race, religion*, and *sexual orientation*. We then employ a causal-inference framework to study the psychological effects of hateful speech in these college subreddits, particularly in the form of individuals' online stress expression. Finally, we characterize their psychological endurance to hateful speech by analyzing their language– we examine their discriminatory keyword use with Sparse Additive Generative Model (SAGE), and their personality traits with Watson Personality Insights API.

Results. Our findings suggest that hateful speech is prevalent in college subreddits, and 25% of these subreddits show greater hateful speech than non-college subreddits. We further find that exposure to hate leads to greater stress expression. However, everybody exposed is not equally affected; some show lower psychological endurance than others. Low endurance individuals are more vulnerable to emotional outbursts, and are more neurotic than those with higher endurance to hate.

Discussion. Our work bears implications for policy-making and intervention efforts to tackle the damaging effects of online hateful speech in colleges. From technological perspective, our work caters to mental health support provisions on college campuses, and to moderation efforts in online college communities. In addition, given the charged aspect of speech dilemma, we highlight the ethical implications of our work. Our work lays the foundation for studying the psychological impacts of hateful speech in online communities in general, and situated communities in particular (the ones that have both an offline and an online analog).

KEYWORDS

social media; Reddit; hateful speech; mental health; stress; natural language analysis; college subreddits

WebSci '19, June 30–July 3, 2019, Boston, MA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and Psychological Effects of Hateful Speech in Online College Communities. In *WebSci '19, June 30–July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Colleges are places where intellectual debate is considered as a key aspect of the educational pursuit, and where viewpoint diversity is venerated. Many colleges in the U.S. have been homes of the free speech movement of the 1960s, that catalyzed positive outcomes, ranging from the women's suffrage movement to the civil rights protests [71]. However, the last few decades has also witnessed several instances where minority groups in colleges have been targeted with verbal altercations, slander, defamation, and hateful speech [40]. In fact, between 2015 and 2016, there has been a 25%[1] rise in the number of reported hate crimes on college campuses.

Because colleges are close-knit, diverse, and geographically situated communities of students, the harmful effects of hateful speech are manifold. In addition to being a precursor to potential hate crimes and violence, hateful speech and its exposure can have profound psychological impacts on a campus's reputation, climate, and morale, such as heightened stress, anxiety, depression, and desensitization [52, 86]. Victimization, direct or indirect, has also been associated with increased rates of alcohol and drug use [78]—behaviors often considered risky in the formative college years [64]. Further, hateful speech exposure has negative effects on students' academic lives and performance, with lowered self-esteem, and poorer task quality and goal clarity—disrupting the very educational and vocational foundations that underscore college experience [16, 60].

Given the pervasive adoption of social media technologies in the college student population [68] and as students increasingly appropriate these platforms for academic, personal and social life discussions [7], hateful speech has begun to manifest online as well [18]. This adds a new dimension to the existing issues surrounding college speech. For instance, it has been reported to be a key driver of and an exacerbating factor behind harassment, bullying, and other violent incidents targeting vulnerable students, often making people feel unwelcome in both digital and physical spaces [47, 78], and even causing psychological and emotional upheavals, akin to its offline counterpart [62, 85].

Campus administrators and other stakeholders have therefore been reported to have struggled with mitigating the negative effects of online hateful speech on campuses, while at the same time valuing students' First Amendment rights [9, 48]. An important step towards addressing existing challenges is to first assess the pervasiveness of online hateful speech, and the vulnerability, in terms of psychological well-being, it presents to marginalized communities on college campuses. However, currently methods to make these assessments are heavily limited. Most existing reports are anecdotal, that have been covered in popular media outlets [50], and are based

[1] *chronicle.com/article/After-2016-Election-Campus/242577*

on discrete events. At the same time, there is no empirical way to comprehensively and proactively quantify and characterize the hateful speech that surface online in the student communities. In addition, social confounds such as the stigma of being exposed to, and the psychological ramifications of hate often lead to underestimates of the effects of online hate, further tempering the mitigation efforts that aim to help these very marginalized groups.

To bridge these gaps, this paper leverages an extensive dataset of over 6 million comments from the online communities of 174 U.S. colleges on Reddit to examine the online dimension of hateful speech in college communities, addressing two research questions:

RQ1: *How prevalent is hateful speech in online college communities, across the demographic categories such as gender, religion, race?*
RQ2: *How does exposure to online hate affect an individual's expression of their psychological state on social media, particularly stress?*

Our work operationalizes hateful speech in online college communities on the hateful content posted in these subreddits. We devise College Hate Index (CHX) to quantify the manifestation of hateful speech across various target categories of hate in an online college community. Our findings suggest that, despite several existing moderation policies on college subreddits [44], hateful speech remains prevalent. Adopting a causal inference framework, we then find that an individual's exposure to online hateful speech impacts their online stress expression. In fact, when exposed to hate, these individuals show a wide range of stress levels, which we characterize using a grounded construct of psychological endurance to hate. Individuals with lower endurance tend to show greater emotional vulnerability, and neuroticism, and conscientiousness traits.

Although this work does not capture offline hateful speech on college campuses, it advances the body of research in *online* hateful speech by examining it in a hitherto under-explored community – college campuses, and by surfacing its psychological effects – a hitherto under-explored research direction. We discuss the implications of our work in providing an important empirical dimension to the college speech debate, and for supporting policy-making and well-being support and intervention efforts to tackle the psychological effects of *online* hateful speech in college communities.

Privacy, Ethics, and Disclosure. Given the sensitive nature of our study, despite working with public de-identified data from Reddit, we do not report any information that associates hateful speech and its psychological effects with specific individuals or college campuses. To describe our approach and to ground our research better, this paper includes paraphrased and partially masked excerpts of hateful comments, for which we suggest caution to readers.

2 RELATED WORK

Hateful Speech on College Campuses. Despite being attributed as a form of "words that wound" [33], hate speech lacks a universally accepted definition. In the specific setting of college campuses, we adopt Kaplin's definition as a way to operationalize hateful speech in the online college communities [48]:

> ..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.

College campuses harbor many diverse communities of race, religion, ethnicity, and sexual orientation. Although argued to be "safe spaces" [81], colleges suffer from many problems related to hate speech, some of which have also escalated to hate crime and violence over the years [9]. The situation is not only alarming, but also controversial, because U.S. colleges have been unable to successfully regulate hateful speech on campuses based on the long ongoing debate over the freedom of expression per the First Amendment [48], and hate speech legislation, or the "speech debate" [56]. Therefore, examining hateful speech in colleges remains a subject of interest from the standpoint of legal, political, and social sciences [41].

To measure the pervasiveness of hateful speech in colleges, stakeholders have adopted a handful of methodologies. Most of these are based on discrete and subjective reports of personal experiences [27, 38, 70], whose recollection can be unpleasant and traumatizing to the victims. A significant limitation of this approach, is that they generate 'optimisitic' estimates—many targets of hateful speech refrain from reporting their experiences for the fear of being victimized, and due to social stigma [13, 52].

Researchers have studied hateful speech through crisis reaction model to find that it shows similar three-phase consequences of feelings (affect), thoughts (cognition), and actions (behavior) as other traumatic events [52]. Further, the victims of hateful speech experience psychological symptoms, similar to post-traumatic stress disorder, such as pain, fear, anxiety, nightmares, and intrusive thoughts of intimidation and denigration [57, 86]. Some early work also outlined that prejudice, discrimination, intolerance, hatred, and factors hindering a student's integration into their social and academic environments can lead to stress and lowered self-esteem among minorities in college campuses, even if they are not the direct victims of specific events [16, 60, 78]. However, assessing the psychological impacts of exposure to hateful speech on college campuses is challenging and has so far been unexplored at scale.

As many of students' discussions have moved online and many social media platforms provide open forum of conversation to students [68, 74], these tools have also paved the way for speech that is usually reserved for the edges of society. In fact, many incidents of hateful speech on campuses, that are targeted at marginalized groups, have recently been reported to have been initiated online [78]. Assessing the repercussions of online hateful speech has been challenging, for the same reasons as its offline counterpart. Our work addresses the above noted gaps by utilizing unobtrusively gathered social media data from online college communities to estimate the pervasiveness of online hateful speech, and how it psychologically impacts the exposed individuals.

Online Hateful Speech and Its Effects. Online hateful speech differs from its offline counterpart in various ways, as a consequence of affordances of online platforms, such as anonymity, mobility, ephemerality, size of audience, and the ease of access [14]. Under the veil of (semi)-anonymity, and the ability to exploit limited accountability that comes with anonymous online activity, perpetrators receive reinforcement from like-minded haters, making hatred seem normal and acceptable [11, 79].

However, both online and offline hateful speech are sometimes inter-related with regards to their causes and effects. For instance, Timofeeva studied online hate speech and additional complexities that it brings to the constitutional right to free speech, and Olteanu et al. demonstrated that offline events (e.g., extremist violence) causally stimulate online hateful speech on social media platforms

like Twitter and Reddit [63, 87]. Other work studied the propagation of online hateful speech following terrorist incidents [15].

Over the past few years, a number of studies have focused on detecting and characterizing hateful speech [45, 80], such as distinguishing hateful speech from other offensive language [29], annotating hateful posts on Twitter based on the critical race theory [89], and conducting a measurement study of hateful speech on Twitter and Whisper [58]. In a recent work, ElSherief et al. studied the distinctive characteristics of hate instigators and targets on social media in terms of their profile self-presentation, activities, and online visibility, and Cheng et al. explored the relationship between individual's mood and antisocial behavior on online communities [22, 37]. Other research has also studied moderation of online antisocial behaviors like undesirable posting [17, 23] and online abuse [12, 20, 46].

Apart from understanding online hateful language, some, although limited studies have also examined its *effects* on the online activities of individuals [5]. [47] showed that victims of online abuse leave the platforms, [85] found that the victims feel increased prejudice, and [18] found that the ban of Reddit communities which incited hateful content was effective towards reducing the manifestation of hateful content on the platform. Similarly, other work found that exposure to online hate among young social media users is associated with psychological and emotional upheavals and heightened distancing from family members [62].Further, [90] studied how various minority groups are targeted with hate speech through various modes of media (both online and offline) and how they are affected because of the exposure to hateful content. Our study advances this critical, yet relatively under-explored line of research by examining how the exposure to online hateful speech can psychologically affect the exposed users, or students in our particular setting of online college communities.

Social Media and Psychological Wellbeing. Literature in psychology has established that analyzing language can help us understand the psychological states of an individual [67]. Several studies have showed that social media data can help us infer and understand the psychological and mental health states of individuals and communities,such as related to depression [30] and mood disorders [73], suicidal ideation [31], and post-traumatic stress [26]. Prior work has also used social media to analyze personality traits and their relationship to wellbeing [69, 82]. Social media data has also facilitated psychological assessments in settings where survey-based assessments are difficult, due to the sensitivities of the situations [32, 74].

Pertaining to the population of college students, Ellison et al. in their seminal work, found positive relationship between social media use and maintenance of social capital [36], and Manago et al. found that social media helped college students to satisfy enduring psychosocial needs [54]. Given the ubiquity of social media use among youth [68], and because social media platforms enable them to share and disclose mental health issues [34], researchers have also leveraged social media as an unobtrusive source of data to infer and understand mental health and wellbeing of college students [53, 55]. Of particular relevance are two recent pieces of work: Bagroy et al., who built a collective mental health index of colleges employing social media (Reddit) data [7], and Saha and De Choudhury, who used college subreddit data to study the evolution of stress following gun violence on college campuses [74].

Although these studies provide us with a foundational background, it remains largely understudied how online community dynamics, such as the exposure to hateful speech affects psychological climate of college campuses. Drawing on the recent success of causal analyses in social media research related to both online hateful speech [18, 63], and mental health [31, 75, 77], we focus on a specific online community behavior (hateful speech in online college communities), and examine its psychological impacts on the online expression of stress of community members.

3 DATA

Online College Community Dataset. Reddit, the source of data in this paper, is one of the most popular social media platforms which caters to the age group between 18-29 years: 65% of Reddit users are young adults [68]. We note that this age demography aligns with the typical college student population, making Reddit a suitable choice for our study. Further, Reddit is a social discussion website which consists of diverse communities known as "subreddits" that offer demographical, topical, or interest-specific discussion boards. Many colleges have a dedicated subreddit community, which provides a common forum for the students to share and discuss about a variety of issues related to their personal, social, and academic life (see e.g., [7, 74, 77]). In fact, the college subreddits name themselves after the college communities that they represent and they often customize their pages with college logos and campus images to signal their identity.

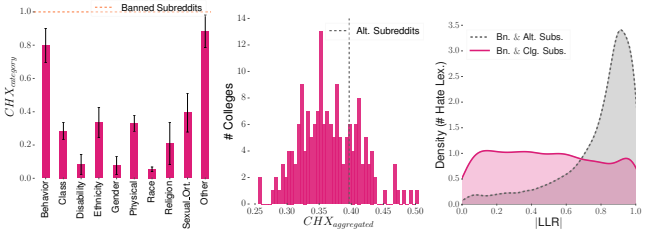
These observations, taken together, indicate that college communities on Reddit can be a source of data to study the research questions posed in this paper. Moreover, such a subreddit dataset has been leveraged in a number of prior work surrounding the study of online college communities [7, 74, 77]. Notably, Bagroy et al. showed that this Reddit data adequately represents the rough demographic distribution of the campus population of over 100 U.S. colleges, is sufficiently widely adopted in these college campuses, and can be employed as a reliable data source to infer the broader college communities' mental wellbeing [7]. While college students likely use other social media platforms as well, such as Facebook, Twitter, Instagram, and Snapchat, obtaining college-specific data from these sources is challenging because many of these platforms restrict public access of data, and they lack defined community structures, precluding gathering sufficiently representative data for specific college campuses. Moreover, these other platforms introduce difficulties in identifying college students users and their college-related discussions on the respective platforms, unless they self-identify themselves, which can limit both scalability and generalizability. In the following subsection we describe how we identify and collect data from college subreddits.

Data Collection. We began by compiling a list of 200 major ranked colleges in the U.S. by crawling the U.S. News (*usnews.com*) website. Next, we crawled the SnoopSnoo (*snoopsnoo.com*) website, which groups subreddits into categories, one of which is "Universities and Colleges". For 174 of these 200 colleges, we found a corresponding subreddit. As of December 2017, these subreddits had 3010 members on an average, and the largest ones were *r/UIUC*, *r/berkeley*, *r/aggies*, *r/gatech*, *r/UTAustin*, *r/OSU*, and *r/ucf* with 13K to 19K members.

Next, we built our dataset by running nested SQL-like queries on the public archives of Reddit dataset hosted on Google BigQuery [1]. Our final dataset for 174 college subreddits included 5,884,905 comments, posted by 453,781 unique users between August 2008 and November 2017. Within this dataset, 4,144,161 comments were posted by 425,410 unique users who never cross-posted across subreddit communities. Students seek and share information and opinion on

Table 1: Excerpts of paraphrased snippets per hate category in the college subreddit dataset.

Category	Example Snippet
Behavior	*Jesus f*cking Christ, hide that and move the f*ck on. Stop whining like a bunch of b*tchy snowflakes.*
Class	*When some hick says some questionable stuff pre 2016 it's just some hick.*
Disability	*I don't want to be called out as a Retarded, dont assume your retarded worldview is correct and try to force such on others.*
Ethnicity	*you are a hispanic? hispanics came from native american p*ssies. this is your land, but you live under shit built by whites!*
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.*
Physical	*That guy is fat and ugly so I didn't read it*
Racial	*Damn n*ggah youze is just a little dude with a litte ole baby dick.*
Religious	*Just like the post about religious fanatics that yell shit on the quad, the same principle is here.*
Sexual Ort.	*BIG SHOT, U WANNA FIGHT? U WANNA BOX F*GGOT?*
Other	*U f*cking uneducated, kid. I ll ruin ur chances of admission in over 700 ways just damn easily.*



(a) (b) (c)

Figure 1: (a) Distribution of category-specific CHX; (b) Histogram of category-aggregated CHX over college subreddits; (c) Kernel Density Estimation of hate lexicon's absolute log-likelihood ratio (LLR) distribution in banned (bn.) against college (clg.) and non-college (alt.) subreddits.

a variety of topics spanning across academics, partying, leisure, relationship, emotional support, and other miscellaneous aspects of college life in particular, and youth life in general.

4 RQ1: PREVALENCE OF HATEFUL SPEECH

4.1 Operationalizing Hateful Speech

A first step in our work revolves around identifying hateful speech in the comments posted on the college subreddits. We adopt a pattern (keyword) matching approach by using a high-precision lexicon from two research studies on hateful speech and social media [29, 58]. This lexicon was curated after multiple iterations of filtering through automated classification, followed by crowdsourced and expert inspection. It consists of 157 phrases that are categorized into: *behavior, class, disability, ethnicity, gender, physical, race, religion, sexual orientation*, and *other*.

Motivation and Validity. Using this lexicon suits our work because we require aggregative assessment of the prevalence of hateful speech— we do not exclusively focus on detecting individual instances of hateful commentary or the specific victims of hate. A lexicon matching approach casts a wider net on all possible manifestations of online hateful speech, compared to supervised learning based detection techniques which are more tuned to keep false positives at a minimum when incorporated in automatic moderation.

Additionally, we frame our reasoning behind the choice of this approach with *validity theory* [28]. First, since we operationalize hate speech by using this validated, crowdsourced, and expert-annotated lexicon, developed and used in prior work, it offers strong *face* and *construct validity*. This lexicon was compiled on hateful words reported by users on the web; thus it offers a better mechanism to capture the subjective interpretation of hate speech, than bag of words based machine learning approaches. From a *convergent validity* perspective, lexicon approaches have performed as good as sophisticated approaches in hate speech detection [29, 58].

This approach is inclusive, using a rich set of cues covering several forms of online hate, it offers rigor in *content validity*, like in prior work [18, 63], [18] used lexicon of as few as 23 phrases to measure hate speech on Reddit. *Content validity* is valuable here because, unlike most work, our goal is not to detect if a post is hateful for moderation, but to get a collective sense of hatefulness in an online community and to support cross-college community comparisons. Finally, we also manually annotated a random sample of 200 college subreddit comments to check *concurrent validity* of the approach. Two researchers familiar with the literature on online hateful speech, independently rated if using the lexicon-based approach, these comments were correctly identified to have hateful content. We found a Cohen's κ of 0.8, suggesting a strong agreement on the comments identified to have evidence of hateful speech and those manually rated.

Approach. Using the above hate lexicon, for every subreddit in our dataset, we obtain a normalized occurrence of hateful speech, given as the fraction of keywords that matched the lexicon, to the total number of words in the subreddit's comments. We obtain both category-specific and category-aggregated measures of hateful speech given in the lexicon.

4.2 College Hate Index (CHX)

Next, we discuss the computation of CHX using the above normalized measure of hate in comments. We first identify five subreddits, which were banned by Reddit primarily due to severe hateful speech usage: *r/CoonTown, r/fatpeoplehate, r/KikeTown, r/nazi, r/transf*gs* [18, 61]. These subreddits glorified hateful speech against certain groups. For example, *r/CoonTown* which grew over 15,000 subscribers self-described itself as "a noxious, racist corner of Reddit" [59]. Our motivation to collect this data stems from the conjecture that hateful speech in these banned subreddits would serve as an upper bound to the amount of hateful speech in any other subreddit (such as the 174 college subreddits, none of which were banned at the time of writing this paper). Accordingly, CHX is a measure to calibrate and standardize the prevalence of hateful speech in a college subreddit, allowing aggregative analysis as well as cross subreddit comparison.

Using the same data collection strategy as explained in the *Data* section, we collect 1,436,766 comments from the five banned subreddits mentioned above. Then, per hate category in our hate lexicon, we compute category-specific and category-aggregated normalized occurrences of hate keywords in the comments of banned subreddits using the method described above. Together with the normalized measures of hate in college subreddits, *we define CHX of an online college community to be the ratio of the normalized hate measure (category-specific or category-aggregated) in the college subreddit to the same measure in banned subreddits*:

$$CHX_T(S) = P_T(S)/P_T(B), \quad (1)$$

where S denotes a specific college subreddit, B denotes all the banned subreddits, T indicates the type of hate speech assessment: category-specific or category-aggregated, $P_T(S)$ and $P_T(B)$ respectively denote the normalized occurrence of

hate keywords for category type T in S and B. For category-aggregated CHX, T includes *all* hate keywords, and for category-specific CHX, it includes *category-specific* ones.

Based on the above equation 1, a college subreddit with no hate shows CHX of 0, whereas if its hateful speech prevalence matches that in banned subreddits, it shows a CHX of 1. Note that, practically speaking, in a college subreddit the normalized occurrence of hate words can exceed that in the banned subreddits. However, it is less likely based on our reasoning above; thus, we cap the maximum value of CHX at 1, allowing us to bound it in the $[0, 1]$ range.

4.3 Measuring the Prevalence

We find that hateful speech in college subreddits is non-uniformly distributed across the different categories of hate (Figure 1a). A Kruskal-Wallis one-way analysis of variance reveals significant differences in the category-specific occurrences of hate ($H = 1507, p < 0.05$). Among the hate categories, *Other* (mean CHX=0.9) and *behavior* (mean CHX=0.8) show the highest occurrence in college subreddits. While hateful speech targeted at ethnicity, race, and religion have been a major concern for many college campuses [48], we observe varied distribution of online hate for these categories. E.g., CHX for *race* ranges between 0.01 and 0.10, for *ethnicity* it ranges between 0 and 0.70, and for *religion* it ranges between 0.01 and 1.00. Hateful speech towards *disability* ranges between 0 and 0.57, and it shows lower average prevalence (mean CHX = 0.08) than all other categories except *race* (mean CHX = 0.05). This observation aligns with a prior finding in the offline context that schools and colleges show comparably lower disability targeted hatefulness compared to non-disability targeted hate [84].

Table 1 reports paraphrased comment excerpts that occur per hate category in the college subreddits. The *Other* category, that demonstrated the highest prevalence, includes keywords like "indecisive", "drunk", and "uneducated". When we examined a random sample of comments, we found that these words are frequently used by the authors to target other members of the community or even the college community in general, e.g., *"They admit gifted students with bright futures but produce uneducated hobos who can't get a job and rely on State alumni for welfare."*.

At an aggregate level, we find that hateful speech in college subreddits is indeed prevalent and ranges between 0.26 and 0.51 (mean=0.37; stdev.=0.05) (see Figure 1b). We find that there are no college subreddits with CHX above 0.51; this reveals reasonable civility in these communities, unlike the banned ones. However, the fact that there are no college subreddits at all with CHX below 0.26 indicates the pervasiveness of the phenomenon.

4.4 Comparison with Non-College Subreddits

Having established the prevalence of hateful speech in online college communities, it raises a natural question: how does this prevalence compare against hateful speech that is manifested elsewhere on Reddit? To answer this, we identify 20 subreddits (alt. subreddits hereon) from the landing page of Reddit, which harbor a diversity of interests and are subscribed by a large number of Reddit users (e.g., *r/AskReddit*, *r/aww*, *r/movies*). From these, we collect a random sample of 2M comments (100K comments per subreddit), and using the same strategy to measure the prevalence of hateful speech (as CHX), we calculate the hate index in these subreddits at an aggregate level, and find it to be 0.40. This shows that although a majority of the online college communities reveal lower CHX (Figure 1b), over 25% of them have *greater* hateful speech than the average prevalence in non-college subreddits.

We further investigate the above distinction in prevalence of hateful speech in college subreddits through a log-likelihood ratio distribution. For every word in our hate lexicon, we calculate their standardized occurrence in the banned, alt., and college subreddits. Then, taking banned subreddits as the common reference, we calculate these keywords' absolute log-likelihood ratios (LLR) in college and alt. subreddits. An absolute LLR of a keyword quantifies its likelihood of presence in either of the two datasets, i.e, lower values of LLR (closer to 0) suggests comparable likelihood of occurrence, whereas higher values of LLR (closer to 1) suggests skewness in the occurrence of a lexicon keyword in either of the two datasets (banned subreddit and college or alt. subreddit).

Figure 1c shows the kernel density estimation of hate keywords' absolute LLR distribution in banned subreddits against college and alt. subreddits. An independent-sample t-test confirms the statistical significance in their differences ($t = -54.95, p < 0.05$). We find that the mean absolute LLR of hate lexicon in banned and college subreddits (mean = 0.49) is lower than that in banned and alt. subreddits (mean = 0.78). This suggests that a greater number of hate keywords show a similar likelihood of occurrence in college subreddits as their occurrence in the banned subreddits.

5 RQ2: PSYCHOLOGICAL EFFECTS OF HATE

Recall that our RQ2 asks whether and how the hatefulness in college subreddits affects the psychological state of the community members. To first operationalize psychological state of these online communities, we refer to prior literature that shows that hateful speech is associated with emotional upheavals and distress [57, 86], with stress being one of the most prominent responses in those exposed to hate both directly and indirectly. We approach RQ2 by first quantifying the extent of hate exposure of an individual in the college subreddits, and then measuring the same individuals' online stress. Eventually, we employ a causal inference framework, drawing from Rubin's causal model [42], to explore the causal link between exposure to hateful speech and stress expression.

5.1 Defining and Quantifying Hate Exposure

We first present a working definition of hate exposure. Without the loss of generality, we assume hate exposure for an individual to be the volume of hateful words shared by others that they are exposed to as a result of participation via commentary in a college subreddit. We calculate this exposure per user as an aggregated percentage of hateful words used by other users on all the threads the user has participated in. We use the same lexicon of hate keywords as described in the previous section.

We note that this is a *conservative definition* of online hate exposure, because individuals can be exposed without commenting on a thread with hateful speech; for instance, by simply browsing such a thread. Exposure may also have offline or spill over effects, such as offline hateful expressions whose effects can get amplified when an individual engages with similar content online. However our definition yields a high precision dataset of exposed users, as commentary explicitly signals that individuals have almost certainly consumed some of the hateful content shared by others in a thread.

Further, through this definition of exposure, we choose to not restrict our analysis only to the intended individual targets of hateful speech, but rather to examine the effects of hateful speech within

college subreddits more broadly, at a community-level. Since college subreddits have an offline analog—the offline community on campus, our choice for this broader definition of "exposure" is also inspired by prior psychology literature which revealed that a toxic (or negative) environment can affect individuals in various forms of presence or relatedness [66].

5.2 Stress Expressed in College Subreddits

Our next objective is to quantify that user's online stress expression, with the psychologically grounded assumption that stress is a manifestation of their psychological state. For this, we appropriate prior work that demonstrated that online stress expression can be measured from content shared in the college subreddits [74, 76].

Specifically, we reproduce a supervised learning based stress detector (classifier) from [74]. This classifier (a support vector machine model with a linear kernel) employs a supervised learning methodology [65] on a Reddit dataset comprising 2000 posts shared on a stress disclosure and help seeking subreddit, *r/stress* (positive ground truth examples or High Stress), and another 2000 posts obtained from Reddit's landing page that were not shared in any mental health related subreddit (negative examples or Low Stress). Using n-grams and sentiment of the posts as features and based on k-fold ($k = 5$) cross-validation, the classifier predicts a binary stress label (High or Low Stress) for each post with a mean accuracy and mean F1-score of 0.82. This classifier was expert-validated using the Perceived Stress Scale [25] (expert validation accuracy = 81%) on college subreddit data like ours [74]. Similar supervised learning approaches have also been recently used in other work to circumvent the challenges of limited ground-truth [7, 77].

In our case, first, applying this stress classifier, we machine label the 4,144,161 comments in our dataset as high and low stress. Then we aggregate the labeled posts per user for the 425,410 users, to assess their online stress expression. Example comments labeled high stress in our dataset include, "*That sounds very challenging for me. I am a CS major*", "*College can be very tough at times like this.*", "*Got denied, but I had to act, I'm very disappointed*".

5.3 Matching For Causal Inference

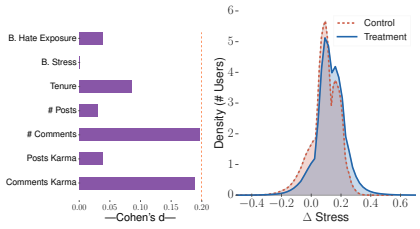
Next, we aim to quantify the effects of exposure to hateful speech with regard to the stress expressed by users in the college subreddits. This examination necessitates testing for causality in order to eliminate (or minimize) the confounding factors that may be associated with an individual's expression of stress. Ideally such a problem is best tackled using Randomized Controlled Trials (RCTs). However, given that our data is observational and an RCT is impractical and unethical in our specific context involving hateful speech exposure and an individual's psychological state, we adopt a causal inference framework based on statistical matching. This approach aims to simulate a randomized control setting by controlling for observed covariates [42]. For our problem setting, we "match" pairs of users using the propensity score matching technique [42], considering covariates that account for online and offline behaviors of users.

5.3.1 Treatment and Control Groups, and Matching Covariates. We define two comparable cohorts of users who are otherwise similar, but one that was exposed to hateful speech (*Treatment* group) whereas the other was not (*Control* group). To obtain statistically matched pairs of *Treatment* and *Control* users, we control for a variety of covariates such that the effect (online stress) is examined between comparable groups of users showing similar offline and online behaviors: 1) First, we control for users within the *same college subreddits*, which accounts for *offline behavioral changes* attributable to seasonal, academic calendar, or local factors [74]. 2) Next, we account for the *user activity* on Reddit with covariates, per prior work [18, 77], which includes the *number of posts and comments*, *karma* (aggregated score on the user's posts and comments), *tenure* (duration of participation) in the community. 3) Finally, to minimize the confounding effects of latent factors of those associated with an individual's stress, we limit our analysis in the period after 2016, and among those 217,109 users who participated in discussion threads *both* before and after 2016. Note that our choice of 2016 hinges on the notion that it enables us roughly 2 years of data for our causal analysis, which is half of the typical period of undergraduate education (4 years). This enables us to obtain a *baseline stress* and a *baseline hate exposure* of every user, which are obtained from the comments posted (shared and encountered) before 2016. This baseline stress measures allows us to account for the fact that the psychological wellbeing of an individual can be impacted by both intrinsic and extrinsic historical factors.

5.3.2 Matching Approach. We use the propensity score matching technique [42] to match 143,075 *Treatment* users with a pool of 74,034 users who were not exposed to any hate on the college subreddits in the period from January 2016 to November 2017. First, we train a logistic regression classifier that predicts the propensity score (p) of each user using the above described covariates as features. Next, for every *Treatment* (T_i) user, we find the most similar *Control* user, conditioning to a maximum caliper distance (c) (with $\alpha = 0.2$), i.e., $\mid T_i(p) - \neg T_i(p) \mid \leq c$, where $c = \alpha * \sigma_{\text{pooled}}$ (σ_{pooled} is the pooled standard deviation, and $\alpha \leq 0.2$ is recommended for "tight matching" [6]). Thereby, we find a matched *Control* user for each of the 143,045 *Treatment* users.

5.3.3 Quality of Matching. To ensure that our matching technique effectively eliminated any imbalance of the covariates, we use the effect size (Cohen's d) metric to quantify the standardized differences in the matched *Treatment* and the *Control* groups across each of the covariates. Lower values of Cohen's d imply better similarity between the groups, and magnitudes lower than 0.2 indicates "small" differences between the groups [24]. We find that the Cohen's d values for our covariates range between 0.001 and 0.197, with a mean magnitude of 0.07 suggesting a good balance in our matching approach (see Figure 2a). Finally, to eliminate any biases in our findings due to the differences in the degree of participation, we also validate whether the matched pairs of users were exposed to similar quantity of keywords in our period of analysis (post 2016). For the number of keywords they were exposed to, the two cohorts of matched users (*Treatment* and *Control*) show a Cohen's d of 0.02, suggesting minimal differences in their exposure to comment threads or their degree of participation in college subreddits.

We further assess the similarity in topical interests between the commenting behavior of *Treatment* and *Control* pairs of users. Here a high value of topical similarity would ascertain minimal confounds introduced due to topical differences (such as high stressed users being more interested in hateful threads). We adopt a word-embedding based similarity approach [8, 74], where for every user, we obtain a word-embedding representation in 300-dimensional vector space of all the subject titles of the discussion threads that they commented on. We choose subject titles because of their prominence on the homepage of a subreddit, and they likely influence users to consume and subsequently comment on the thread. Next, we compute the vector similarity of the subject titles' word-vectors for every pair of *Treatment* and *Control* users, which essentially



(a) (b)

Figure 2: a) Cohen's d for evaluating matching balance of covariates of activity features and baseline (B.) stress and hate exposure; b) Kernel Density Estimation of user distribution with change in stress expression.

Table 2: Regression coefficients for hate categories and change in stress expression (*$p < 0.001$).

Category	Coefficient
Behavior***	2.6×10^{-4}
Class***	5.1×10^{-4}
Disability***	7.3×10^{-3}
Ethnicity***	3.9×10^{-3}
Gender***	8.1×10^{-3}
Physical***	1.5×10^{-3}
Race***	1.7×10^{-3}
Religion***	1.3×10^{-5}
Sexual Ort.***	5.7×10^{-3}
Other***	1.5×10^{-3}

quantifies their topical interests. Across all the pairs of *Treatment* and *Control* users, we find an average cosine similarity of 0.67 (stdev. = 0.17), indicating that our matched users share similar interests in the posts on which they commented on.

5.4 Does Hate Exposure Impact Stress Level?

Following statistical matching, we examine the relationship between the exposure to hate and the expression of stress in college subreddits. Drawing on the widely adopt "Difference in Differences" technique in causal inference research [2], we evaluate the effects of hate exposure on stress by measuring the shifts in online stress for the *Treatment* group and comparing that with the same in the *Control* group. According to Rubin's causal framework, such an evaluation averages the *effect* (online stress expression) caused by the *treatment* (online hate exposure) on the treated individuals by comparing that with what the same individuals would have shown had they not been treated (the individual's matched pair) [42].

We observe that compared to their baseline stress, the stress level of the *Treatment* users (mean=139%) is higher than the *Control* users (mean=106%). An effect size measure (Cohen's d=0.40) and a paired t-test indicates this difference to be statistically significant (t=93.3, $p < 0.05$). Figure 2b shows the changes in stress level for the two user groups *Treatment* and *Control* users subject to hate speech exposure in the college subreddits. Given that these two groups are matched on offline and online factors, such revealing differences in stress between them following online hate exposure suggest that this exposure likely has a causal relationship with the online stress expression of the users.

Now that we demonstrated that *online hate exposure* plausibly influences the *online stress expression* of individuals in college subreddits, we are next interested in how the various categories of hate leads to shifts in online stress expression among the *Treatment* users. For this, we fit a linear regression model with the hate categories as independent variables and the change in stress expression as the dependent variable. Table 2 reports the coefficients of these categories in the regression model where all of them showed statistical significance in their association. These coefficients could be interpreted as—every unit change in online hate exposure from a category leads to an approximate change in online stress expression by the magnitude of the corresponding coefficient. We find that each of the hate categories shows a positive coefficient, further indicating that an increase in exposure to any category of hate increases the stress expression of members of the college subreddits. Among these categories, we find that *gender* (0.81%) and *disability* (0.73%) show the greatest coefficients, and therefore affect most towards the online stress expression of the community members.

5.5 Psychological Endurance to Hate Exposure

Within our *Treatment* group, we observe that users are not *equally* affected in their stress levels. In fact, they show a wide range of online stress (median = 0.05, stdev. = 0.80) at varying magnitudes of online hate exposure (median = 0.68, stdev. = 3.61) (see Figure 3). So, besides observing that hate exposure in these communities bears a causal relationship with online stress expressions, we also find that online hate does not affect everybody's stress expression uniformly. This aligns with the notion that individuals differ in their resilience to the vicissitudes of life [52]. We call this phenomenon of varied tolerance among users as the *psychological endurance* to online hateful speech. Our motivation to examine this endurance construct comes from the psychology literature, which posits that different people have different abilities to deal with specific uncontrollable events, and stress results from the perception that the demands of these situations exceed their capacity to cope [43].

To understand psychological endurance to online hate, we look at two groups of users who express the extremes of online stress at the opposing extremes of online hate exposure. One group comprises those *Treatment* users with low endurance who have lower tolerance to online hate than most other users and show high (higher than median) stress changes when exposed to low (lower than median) online hate (quadrant 4 in Figure 3). The other group consists of those users who have much higher tolerance, and show low (lower than median) stress changes when exposed to high (higher than median) hate (quadrant 2 in Figure 3). We refer to these two groups as *low endurance* and *high endurance* users—we find 38,503 low and 38,478 high endurance users in our data.

5.6 Analyzing Psychological Endurance

Our final findings include an analysis of the attributes of high and low endurance users as manifested in the college subreddits. We focus on two kinds of attributes— users' online linguistic expression, and their personality traits as inferred from their language. Given that we distinguish the psychological behaviors of two cohorts (individuals with low and high endurance to hateful speech), the choice of these attributes stem from prior work that studied psychological *traits* and *states* of individuals as gleaned from their social media activity [21].

Linguistic Expression. To understand in what ways the low and high endurance users differ in their language use, we employ an unsupervised language modeling technique known as the Sparse Additive Generative Model (SAGE) [35], which has been widely applied in computational linguistic problems on social media data [20, 76, 83]. Given any two documents, SAGE selects discriminating keywords by comparing the parameters of two logistically parameterized multinomial models, using a self-tuned regularization parameter to control the tradeoff between frequent and rare terms. We use the SAGE model to identify discriminating n-grams (n=1,2) between the comments of low and high endurance users. The magnitude of SAGE value of a linguistic token signals the degree of its "uniqueness", and in our case a positive SAGE more than 0 indicates that the n-gram is more representative for the low endurance users,

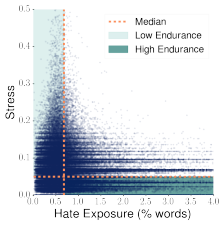


Figure 3: Dist. *Tr.* users with their stress expression and hate exposure.

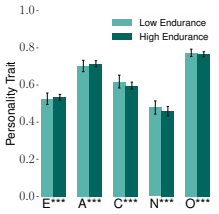


Figure 4: Personality traits in *Tr.* users. Statistical significance reported after Bonferroni correction on independent sample t-tests (*$p < 0.001$).**

Table 3: Top 25 discriminating n-grams ($n = 1, 2$) used by Low and High Endurance users (SAGE [35]).

Low Endurance		High Endurance	
n-gram	SAGE	n-gram	SAGE
education code	1.99	guitar	-1.19
prerequisite course	1.68	discord	-1.15
general education	1.64	baylor	-1.15
classes mentioned	1.26	esports	-1.09
mathematics	1.07	smash	-1.00
credit course	1.03	bath	-0.98
upper division	0.95	bands	-0.97
prerequisite	0.95	tournament	-0.93
preassessment	0.89	phi	-0.92
therapy	0.88	garden	-0.90
senate	0.85	shots	-0.89
task	0.84	pokemon	-0.89
cse	0.83	delicious	-0.86
immigrants	0.82	temple	-0.85
prereq	0.81	wings	-0.84
cs1	0.79	used work	-0.84
mathematical	0.78	players	-0.83
irrelevant	0.78	song	-0.83
mentor	0.78	jazz	-0.80
tasks	0.76	anime	-0.80
division	0.75	basement	-0.79
comptia	0.75	rock	-0.79
assessment	0.73	student section	-0.78
anxiety	0.73	yo	-0.78
sql	0.73	sublease	-0.77

whereas a negative SAGE denotes greater representativeness for high endurance users.

Table 3 reports the top 25 n-grams ($n = 1, 2$) for low and high endurance users. One pattern evident in these n-grams is that low endurance users tend to use more classroom-oriented and academic-related topics, such as "education", "prerequisite", "assessment", "mathematical", etc.. Whereas, the high endurance group demonstrates greater usage of words that relate to a more relaxed and leisure-like context, as well as to diverse non-academic topics/interests, such as "pokemon", "guitar", "delicious", "anime", and "garden". We also find mental health related terms such as "therapy" and "anxiety" for low endurance users, which can be associated with these users self-disclosing their condition or with their help-seeking behaviors around these concerns.

Personality Traits. Our final analysis focuses on understanding the personality differences between individuals showing varied levels of psychological endurance to online stress. Personality refers to the traits and characteristics that uniquely define an individual [82]. Psychology literature posits personality traits as an important aspect to understand the drivers of people's underlying emotional states, trust, emotional stability, and locus of control [3]. For instance, certain personality traits, such as extraversion and neuroticism, represent enduring dispositions that directly lead to subjective wellbeing in individuals, including the dimensions of happiness and negative affect [3]. We study relationship of psychological endurance with personality traits, which can be inferred from social media data of the users [69, 82].

To characterize the personality traits of the users who show low and high psychological endurance, we run their comments' dataset through the Watson Personality Insights API [4], to infer personality in five dimensions of traits: *openness*, *agreeableness*, *extraversion*, *neuroticism*, and *conscientiousness*. Prior research has used this method to extract and characterize several linguistic and psychological constructs from text [21, 37]. Figure 4 shows the distribution of personality traits for low and high endurance users. Paired t-tests revealed statistically significant differences between the two groups. Drawing from seminal work on the big-five factor structure of personality [39], we situate our observations as follows:

We observe that the high endurance group reveals 2% greater *agreeableness* (t=-66.31) and *extraversion* (t=-42.62). *Agreeableness* characterizes an attribute of being well-mannered, and those who show higher values are generally considered to be less reactive to challenges or an attack (here online hateful speech). *Extraversion* indicates greater sociability, energy, and positive emotions, and lower values signal a reflective personality, which suggests that even lower exposure to online hate can negatively impact users with low endurance. The low endurance users also show 4% greater *neuroticism* (t=89.42) and *conscientiousness* (t=109.31). *Neuroticism* indicates the degree of emotional stability and higher values signal increased tendency to experience unpleasant emotions easily. Despite these posthoc conjectures, we do acknowledge that understanding these relationships between endurance and personality would require deeper investigation beyond the scope of this paper.

Based on these observations and what we already found in SAGE analysis of the low and high endurance users (Table 3), we infer that even with comparable hate exposure in the college subreddits, different individuals may respond psychologically differently, and these differences may be observed via attributes such as their language of expression on social media and their underlying personality traits.

6 DISCUSSION

6.1 Socio-Political and Policy Implications

The speech debate has been a part of the American socio-political discussions for many years now [56]. In particular, on college campuses, it presents many complexities in decision or policy making that seeks to combat hateful speech within campuses [48]. While this paper does not provide any resolution to this debate, it makes empirical, objective, and data-driven contributions and draws valuable insights towards an informed discussion on this topic.

First, while the speech debate so far has largely focused in the *offline* context, as our study shows, hateful speech in the *online* domain also bears negative impacts on the exposed population, especially in situated communities like college campuses. Our findings align with prior work on the psychological impacts on hateful speech in the offline context [57, 86]. At the same time, they extend the literature by showing that there are not only pronounced differences in the prevalence of various hate speech types, but also the exposure to hate affects individuals' online stress expression. Here we note that antisocial behaviors like hateful speech continue to be a pressing issue for online communities [18, 22], but the effects of online hateful speech remains the subject of little empirical research. Thus, these findings help to account for a previously under-explored, but a critical facet of the speech debate, especially in the context of college campuses.

Second, our findings add new dimensions to the college speech debate centering around legal, ownership, and governance issues. These issues involve not only those who trigger and those who are exposed to online hateful speech, but also the owners, the moderators, the users, and the creators of social media platforms, who may not necessarily be part of the college community.

Third, our work highlights a new policy challenge: how to decipher when online and offline hateful speech reinforce each other, and to delineate their psychological effects, particularly in a situated community where there is likely overlap between the online and offline social worlds. Our work indicates that the affordances of social media, such as anonymity and low effort information sharing amplify the complexities of the speech debate on college campuses. Typically, colleges can choose to restrict the time, place, and manner of someone's speech. However, when speech is not physically located on campus, these affordances can be exploited to quickly reach large segments of the campus population, posing new threats. Consequently, how should college stakeholders respond, when, based on our approach, a student is found to use online hate, observably outside of the physical setting of the campus, targeting a person of marginalized identity?

Finally, our work opens up discussions about the place of "counter-speech" in these communities to undermine the psychological effects of hate, alongside accounting for the legal concerns and governance challenges that enforcing certain community norms may pose [72]. We do note that any such discussions promoting counter speech would need to factor in the general etiquette of conduct expected from the members of the college community to avoid misethnic or chauvinistic phrasing, and to maintain a vibrant and inclusive environment which is respectful of other members [49].

6.2 Technological Implications

An important contribution of our work is the novel computational framework we developed to assess the pervasiveness and the psychological effects of online hateful speech on the members of online college communities. We believe these methods can lead to two types of technology implications:

6.2.1 Mental Health Support Provisions on College Campuses. The ease of interpretation and the ability to track language changes over time allows our empirical measure of online hateful speech in college campuses to be robust and generalizable across different online college communities, and also accessible to various stakeholders, unlike what is supported by existing hate speech detection techniques [80]. Our methods can thus be leveraged by college authorities to make informed decisions surrounding the speech dilemma on campuses, promote civil online discourse among students, and employ timely interventions when deemed appropriate. While our approach to assess the prevalence of hateful speech is likely to be not perfectly accurate, alongside human involvement in validating these outcomes, timely interventions to reduce the harmful effects of online hateful language can be deployed. As Olteanu et al. [63] recently pointed out that exogenous events can lead to online hatefulness, our framework can assist to proactively detect the psychological ramifications of online hate at their nascent stage to prevent negative outcomes.

Additionally, our work helps us draw insights about the attributes of individuals with higher vulnerability and lower psychological endurance to online hateful speech. This can assist in instrumenting tailored and timely support efforts, and evidence-based decision strategies on campuses. We further note that, any form of hateful speech, whether online or offline, elicits both problem- and emotion- focused coping strategies, and the victims of hateful speech seek support [52]. Many colleges already provide various self, peer, and expert-help resources to cater to vulnerable students. These efforts may be aligned to also consider the effects of *online* hateful speech exposure as revealed in our work.

6.2.2 Moderation Efforts in Online College Communities. Our findings suggest that hateful speech *does* prevail in college subreddits. However, unlike most other online communities, banning or extensively censoring content on college subreddits—a strategy widely adopted today [18, 61] as a measure to counter online antisocial behavior can have counter-effects. Such practices would potentially preclude students from accessing an open discussion board with their peers where not only many helpful information is shared, but also which enables them to socialize and seek support around academic and personal life related topics. Rather, our work can be considered to be a "call-to-action" for the moderators to adopt measures that go beyond blanket banning or censorship. For instance, our approach to assess the stress and hate exposure of users can assist moderators to tune community environment and adapt norms in a way that discourages hateful speech. This could be subreddit guidelines that outline moderation strategies not only discouraging offensive and unwelcoming content, but also around content that *affects* community members. For example, the subreddit *r/lifeprotips* explicitly calls out against *"tips or comments that encourage behavior that can cause injury or harm to others can cause for a (user) ban"*. Other moderation strategies can also be adopted: such as using labels in specific posts which are perturbing, along the lines of *r/AskReddit* which uses "[Serious]" to particularly label very important and relevant discussion threads.

Moderators can also provide assistance and support via peer-matching, and include pointers to external online help resources, especially to members who are vulnerable to the negative psychological impacts of online hateful content. Complementarily, as argued in recent research [11], making the harmful effects of hateful language transparent to the community members in carefully planned and strategic manner, could curb the prevalence of antisocial practices including hateful speech. Specifically in online college communities, where the members are geographically situated and embedded in their offline social ties [10, 36], knowledge of the negative psychological repurcussion of certain online practices could influence them to refrain from or not engage with such behaviors.

In the offline context, the college speech debate has also aroused discussions surrounding *safe spaces*: particular sites of campuses where students join peers, and *trigger warnings*: explicit statements that certain material discussed in an academic environment might upset sensitive students [49, 88]. These measures are advocated to help in minimize hateful speech and its effects. We argue that analogous measures are possible in online communities as well, using the design affordances of the social media platforms (e.g., creating separate subreddits for minority communities in a college, or providing pop-up warnings on certain posts). However, both safe spaces and trigger warnings are critiqued as they are exclusionary and are harmful for open discourse in colleges. So, any such possible consequences should also be carefully evaluated before such measures are adopted in online communities of college students.

6.3 Ethical Implications

Amid the controversy surrounding the freedom of expression, defining (online) hateful speech remains a complex subject of ethical, legal, and administrative interest, especially on college campuses that are known to value inclusiveness in communities, and to facilitate progressive social exchange. While our definition of hateful speech in online college communities may not be universal, our measurement approach provides an objective understanding of the dynamics and impacts of hateful environment within online college communities. Nevertheless, any decision and policy making based

on our findings requires careful and in-depth supplemental ethical analysis, beyond the empirical analysis we present in this paper. For instance, to what extent online hateful speech infringes on the speech provisions on specific campuses remains a topic that needs careful evaluation. Importantly, supported by our analysis, campus stakeholders must navigate a two-prong ethical dilemma: one around engaging with those who use online hateful speech, and two, around treating its extreme manifestations, like hate related threats and altercations directed at campus community members, or its interference with the institution's educational goals.

We finally caution against our work being perceived as a means to facilitate surveillance of student speech on college campuses, or as a guideline to censor speech on campus. Our work is not intended to be used to intentionally or inadvertently marginalize or influence prejudice against those groups who are already marginalized (by gender, race, religion, sexual orientation etc.), or vulnerable, and are often the targets of hateful speech on campuses.

6.4 Limitations and Future Work

Our study includes limitations, and some of these suggest promising directions for future work. Although our work is grounded in prior work [7] that college subreddits are representative of the respective student bodies, we cannot claim that our results extrapolate directly to offline hateful speech on college campuses [14]. Similarly, we cannot claim that these results will be generalizable to general purpose or other online communities on Reddit or beyond, as well as with or without an offline analog like a college campus. Importantly, we did not assess the clinical nature of stress in our data, and focused only on inferred stress expression from social media language [74]; future work can validate the extent to which online hate speech impacts the mental health of students. Like many observational studies, we also cannot establish a *true causality* between an individual's exposure to online hate and their stress expressions. To address these limitations, future work can gather ground truth data about individual stress experiences and clinically validate them with social media derived observations.

We note that our work is sensitive to the uniqueness of the Reddit platform, where the content is already moderated [19, 46, 61]. It is possible that the definition of hateful content qualifying for content removal could vary across the college subreddits, and our work is restricted to only the non-removed comments. Importantly, the norms and strategies to moderate content can vary across different college subreddits. Therefore, our study likely provides a "lower bound estimate" of hateful content on these communities. Additionally, users also use multiple accounts and throwaway accounts on Reddit [51], and we do not identify individual users' experiences of online hate or stress in our dataset. Our findings about the psychological endurance to hate is interesting and inspires further theoretical and empirical investigations—e.g., how can we generalize the relationship between online hate and psychological wellbeing both on campuses and elsewhere, what factors influence the endurance of an individual, and how can we characterize endurance in terms of direct victimization or indirect influence of the ripple effects of online hateful speech on campuses.

7 CONCLUSION

In this paper, we first modeled College Hate Index (CHX) to measure the degree of hateful speech in college subreddits. We found that hateful speech does prevail in college subreddits. Then, we employed a causal inference framework to find that the exposure to hateful speech in college subreddits impacted greater stress expression of the community members. We also found that the exposed users showed varying psychological endurance to hate exposure, i.e, all users exposed to similar levels of hate reacted differently. We analyzed the language and personality of these low and high endurance users to find that, low endurance users are vulnerable to more emotional outbursts, and are more conscientious and neurotic than those showing higher endurance to hate.

8 ACKNOWLEDGEMENT

We thank Eric Gilbert, Stevie Chancellor, Sindhu Ernala, Shagun Jhaver, and Benjamin Sugar for their feedback. Saha and De Choudhury were partly supported by research grants from Mozilla (RK677) and the NIH (R01GM112697).

REFERENCES

[1] Bigquery. Google BigQuery. bigquery.cloud.google.com. Accessed: 2017-10-27.
[2] Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *Rev. Econ. Stud.* 72, 1 (2005), 1–19.
[3] Gordon Willard Allport. 1937. Personality: A psychological interpretation. (1937).
[4] IBM Watson Personality API. 2018. personality-insights-demo.ng.bluemix.net. Acc: 2018-04-18.
[5] Pinar Arslan, Michele Corazza, Elena Cabrio, and Serena Villata. 2019. Overwhelmed by Negative Emotions? Maybe You Are Being Cyber-bullied!. In *The 34th ACM/SIGAPP Symposium On Applied Computing (ACM SAC 2019)*.
[6] Peter C Austin. 2011. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical statistics* 10, 2 (2011), 150–161.
[7] Shrey Bagroy, Ponnurangam Kumaraguru, and Munmun De Choudhury. 2017. A Social Media Based Index of Mental Well-Being in College Campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.
[8] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?. In *IJCNLP*.
[9] Katharine T Bartlett and Jean O'Barr. 1990. The Chilly Climate on College Campuses: An Expansion of the" Hate Speech" Debate. *Duke Law J.* (1990).
[10] Victor Battistich and Allen Hom. 1997. The relationship between students' sense of their school as a community and their involvement in problem behaviors. *American journal of public health* 87, 12 (1997), 1997–2001.
[11] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When Online Harassment is Perceived as Justified. In *ICWSM*.
[12] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *Proc. ACM Hum.-Comput. Interact.* CSCW (2017), 24:1–24:19.
[13] Robert Boeckmann and Jeffrey Liew. 2002. Hate speech: Asian American students' justice judgments and psychological responses. *J. Soc. Issues* (2002).
[14] Alexander Brown. 2017. What is so special about online (as compared to offline) hate speech? *Ethnicities* (2017).
[15] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7, 2 (2015).
[16] Alberto F Cabrera, Amaury Nora, Patrick T Terenzini, Ernest Pascarella, and Linda Serra Hagedorn. 1999. Campus racial climate and the adjustment of students to college: A comparison between White students and African-American students. *The Journal of Higher Education* (1999).
[17] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proc. CSCW*. 1201–1213.
[18] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *PACM HCI (CSCW)* (2017).
[19] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *PACM HCI* CSCW (2018).
[20] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proc. CHI*.
[21] Jilin Chen, Gary Hsieh, Jalal U Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *CSCW*.
[22] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proc. CSCW*.
[23] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *International AAAI Conference on Web and Social Media*.
[24] Jacob Cohen. 1992. Statistical power analysis. *Curr. Dir. Psychol. Sci.* (1992).

[25] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[26] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *ICWSM*.

[27] Gloria Cowan and Cyndi Hodge. 1996. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology* 26, 4 (1996), 355–374.

[28] Linda Crocker and James Algina. 1986. *Introduction to classical and modern test theory*. ERIC.

[29] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

[30] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.

[31] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. CHI*.

[32] Munmun De Choudhury, Andres Monroy-Hernandez, and Gloria Mark. 2014. Narco emotions: affect and desensitization in social media during the mexican drug war. In *CHI*. ACM, 3563–3572.

[33] Richard Delgado and Jean Stefancic. 2004. *Understanding words that wound*. Westview Press.

[34] Daniel Eisenberg, Justin Hunt, and Nicole Speer. 2012. Help seeking for mental health on college campuses: Review of evidence and next steps for research and practice. *Harvard review of psychiatry* (2012).

[35] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. (2011).

[36] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends": Social capital and college studentsúse of online social network sites. *Journal of Computer-Mediated Communication* (2007).

[37] Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *International AAAI Conference on Web and Social Media (ICWSM)* (2018).

[38] Steve France. 1990. Hate goes to college. *ABA Journal* 76 (1990), 44.

[39] Lewis R Goldberg. 1990. An Alternative "Description of Personality": The Big-Five Factor Structure. *J. Pers. Soc. Psychol.* (1990).

[40] Daniel Goleman. 1990. As bias crime seems to rise, scientists study roots of racism. *New York Times* (1990), C1.

[41] Jon B Gould. 2001. The precedent that wasn't: College hate speech codes and the two faces of legal compliance. *Law Soc. Rev.* (2001).

[42] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge.

[43] Rick E Ingram and David D Luxton. 2005. Vulnerability-stress models. *Dev. Psychopathol.: A vulnerability-stress perspective* (2005).

[44] Business Insider. 2018. businessinsider.com/what-is-a-reddit-moderator-2016-1. Acc: 2018-04-18.

[45] Shagun Jhaver, Larry Chan, and Amy Bruckman. 2018. The view from the other side: The border between controversial speech and harassment on Kotaku in Action. *First Monday* 23, 2 (2018).

[46] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018).

[47] Ruogu Kang, Laura Dabbish, and Katherine Sutton. 2016. Strangers on your phone: Why people use anonymous communication applications. In *CSCW*.

[48] William A Kaplin. 1992. A proposed process for managing the First Amendment aspects of campus hate speech. *J. High. Educ.* (1992).

[49] Mae Kuykendall and Charles Adside III. 2013. Unmuting the Volume: Fisher, Affirmative Action Jurisprudence, and the Legacy of Racial Silence. *Wm. & mary bill rts. J.* 22 (2013), 1011.

[50] L.A.Times. 2017. latimes.com/local/lanow/la-me-berkeley-free-speech-20170605-story.html. Acc: 2018-01-17.

[51] Alex Leavitt. 2015. This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *Proc. CSCW*.

[52] Laura Leets. 2002. Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of social issues* 58, 2 (2002), 341–361.

[53] Sam Liu, Miaoqi Zhu, and Sean D Young. 2018. Monitoring freshman college experience through content analysis of tweets: observational study. *JMIR public health and surveillance* 4, 1 (2018), e5.

[54] Adriana M Manago, Tamara Taylor, and Patricia M Greenfield. 2012. Me and my 400 friends: the anatomy of college students' Facebook networks, their communication patterns, and well-being. *Developmental psychology* (2012).

[55] Gloria Mark, Yiran Wang, Melissa Niiya, and Stephanie Reich. 2016. Sleep Debt in Student Life: Online Attention Focus, Facebook, and Mood. In *Proc. 2016 CHI Conference on Human Factors in Computing Systems*. 5517–5528.

[56] Toni M Massaro. 1990. Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.* (1990).

[57] Mari J Matsuda. 1993. *Words that wound: Critical race theory, assaultive speech, and the first amendment*. Westview.

[58] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proc. ACM HT*.

[59] Justin Wm Moyer. 2015. wapo.st/1JmimPm?tid=ss_tw-bottom&utm_term=.75ee6968b36d. Accessed: 2018-07-05.

[60] Daniel G Muñoz. 1986. Identifying areas of stress for Chicano undergraduates. *Latino college students* (1986), 131–156.

[61] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest.. In *ICWSM*.

[62] Atte Oksanen, James Hawdon, Emma Holkeri, Matti Näsi, and Pekka Räsänen. 2014. Exposure to online hate among young social media users. *Soul of society: a focus on the lives of children & youth* (2014).

[63] Alexandra Olteanu, Carlos Castillo, Jeremy Boy, and Kush Varshney. 2018. The Effect of Extremist Violence on Hateful Speech Online. In *International AAAI Conference on Web and Social Media (ICWSM)*.

[64] Patrick M O'Malley and Lloyd D Johnston. 2002. Epidemiology of alcohol and other drug use among American college students. *Journal of Studies on Alcohol, Supplement* 14 (2002), 23–39.

[65] Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2010), 1345–1359.

[66] Ellen E Pastorino and Susann M Doyle-Portillo. 2012. *What is psychology? Essentials*. Cengage Learning.

[67] James W Pennebaker and Cindy K Chung. 2007. Expressive writing, emotional upheavals, and health. *Handbook of health psychology* (2007), 263–284.

[68] Pew. 2018. pewinternet.org/fact-sheet/social-media. Accessed: 2018-04-18.

[69] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *SocialCom*.

[70] Nadine Recker Rayburn, Mitchell Earleywine, and Gerald C Davison. 2003. Base rates of hate crime victimization among college students. *Journal of Interpersonal Violence* 18, 10 (2003), 1209–1221.

[71] Robert A Rhoads. 1998. *Freedom's web: Student activism in an age of cultural diversity*. ERIC.

[72] Robert D Richards and Clay Calvert. 2000. Counterspeech 2000: A New Look at the Old Remedy for Bad Speech. *BYU L. Rev.* (2000), 553.

[73] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proc. ACM IMWUT* (2017).

[74] Koustuv Saha and Munmun De Choudhury. 2017. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. *PACM HCI (CSCW)* (2017).

[75] Koustuv Saha, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kiciman, and Munmun De Choudhury. 2019. A Social Media Study on The Effects of Psychiatric Medication Use. In *ICWSM*.

[76] Koustuv Saha, John Torous, Sindhu Kiranmai Ernala, Conor Rizuto, Amanda Stafford, and Munmun De Choudhury. 2019. A computational study of mental health awareness campaigns on social media. *TBM* (2019).

[77] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. In *ICWSM*.

[78] Allison M Schenk and William J Fremouw. 2012. Prevalence, psychological impact, and coping of cyberbully victims among college students. *Journal of School Violence* 11, 1 (2012), 21–37.

[79] Ari Schlesinger, Eshwar Chandrasekharan, Christina A Masden, Amy S Bruckman, W Keith Edwards, and Rebecca E Grinter. 2017. Situated anonymity: Impacts of anonymity, ephemerality, and hyper-locality on social media. In *Proc. CHI*.

[80] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *SocialNLP*.

[81] Nicholas A Schroeder. 2017. Avoiding Deliberation: Why the Safe Space Campus Cannot Comport with Deliberative Democracy. *BYU Educ. & LJ* (2017), 325.

[82] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* (2013).

[83] Eva Sharma, Koustuv Saha, Sindhu Kiranmai Ernala, Sucheta Ghoshal, and Munmun De Choudhury. 2017. Analyzing ideological discourse on social media: A case study of the abortion debate. In *Proc. CSS*. ACM.

[84] Mark Sherry. 2016. *Disability hate crimes: Does anyone really hate disabled people?* Routledge.

[85] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. 2017. Exposure to hate speech increases prejudice through desensitization. *Aggress. Behav.* (2017).

[86] Megan Sullaway. 2004. Psychological perspectives on hate crime laws. *Psychology, Public Policy, and Law* 10 (2004).

[87] Yulia A Timofeeva. 2002. Hate Speech Online: Restricted or Protected-Comparison of Regulations in the United States and Germany. *J. Transnat'l L. & Pol'y* 12 (2002), 253.

[88] Alexander Tsesis. 2016. Campus speech and harassment. *Minn. L. Rev.* (2016).

[89] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL-HLT*.

[90] Mikolaj Winiewski, K. Hansen, M. Bilewicz, W. Soral, A. Świderska, and D. Bulska. 2017. Hate speech, contempt speech. Report on verbal violence against minority groups. *Warsaw: Stefan Batory Foundation*. (2017).