# The Strange Case of Jekyll and Hyde: Analysis of r/ToastMe and r/RoastMe Users on Reddit

**Wooyong Jung[1], Nishant Asati[1], Phuong (Lucy) Doan[2], Thai Le[3], Aiping Xiong[1], Dongwon Lee[1]**

[1]The Pennsylvania State University, University Park, PA
[2]Amazon, New York, NY
[3]University of Mississippi, Oxford, MS
[1]{wjung, nxa5283, axx29, dongwon}@psu.edu, [2]lucydoa@amazon.com, [3]thaile@olemiss.edu

## Abstract

This study, focusing on two Reddit subcommunities of r/ToastMe and r/RoastMe, aims to (1) characterize and understand users (named Jekyll and Hyde) who simultaneously participate in two subreddits with opposing tones and purposes, (2) build predictive models detecting those Jekyll and Hyde users to assess how unique and idiosyncratic their characteristics are, and (3) investigate their motivations of participation and potential interaction between the two contrasting activities through a survey and one-on-one interviews. Our results reveal that the Jekyll and Hyde users are generally more active and popular than ordinary users. Also, they use assimilated language customized to each community's tone. Combining these findings with their motivations unveiled through the survey and interviews, we conclude that the Jekyll and Hyde users are digitally culture-savvy, who know how to utilize online community benefits and enjoy each community's culture by assimilating themselves into the community and observing its rules. Moreover, the users' duality observed in this process underscores the dynamic and multifaceted nature of online personas. These findings highlight the need for a nuanced approach to understanding online behaviors and provide insights for designing healthier online environments, emphasizing the importance of clear community norms and the potential interplay of users' activities across different communities.

## Introduction

Online communities where users actively communicate and interact with each other provide a fertile research ground for human behavior studies and cyberpsychology. *Reddit*, one of the most extensively studied online communities, is a social platform where registered users post contents about any topics, which are then voted up or down and discussed by other users. It has 52 million daily active users and addresses a wide variety of topics via 2.8 million subcommunities (called subreddit and denoted as "*r/topic*" ), and provides a wide selection of information on users and communities through API calls (Lin 2021). On this platform, we focus on two polar-opposite subreddits: *r/ToastMe* and *r/RoastMe*. In *r/ToastMe*, users post their photos and plead for random toastings, often followed by other users' genuine, heartwarming, and empathetic compliments. In contrast, in *r/RoastMe*, users give out mocking, humiliating, or sometimes hurtful comments toward users who posted their photos and pleaded for random roastings.

In particular, we stumbled upon users who actively and simultaneously participate in both subreddits. For example, the same user made a thread of comments below consecutively within a day in response to other users' postings. This user oscillated between two subreddits and joined conversations with conflicting and contrasting attitudes.

12:34 pm, r/RoastMe
*"If death, depression, and anxiety had a baby, it would be you"*

11:05 pm, r/ToastMe
*"I just want to encourage you that things may be looking down now, but you will overcome it. Depression sucks and I'm praying for you"*

2:26 am, r/RoastMe
*"Good news! The Taliban accepted your application! ... Just put on this vest, run to that crowd, push the trigger, and pay with your life"*

10:39 am, r/ToastMe
*"You got a heart of gold ... You are enough and your value is not defined by what other people say or think about you."*

In other words, they repeatedly communicate with strangers along the same line, but their tone and sentiment of language used in the communication are entirely opposite. These perplexing behaviors made us wonder: why do some users actively participate in two polar-opposite communities, and who are they? Because their behaviors are redolent of *Dr. Jekyll and Mr. Hyde* from Robert Louis Stevenson's Gothic novella, *Strange Case of Dr. Jekyll and Mr. Hyde*, we named them Jekyll and Hyde users (or **JH** in short). Although we use the term throughout this study for its convenient reference, it does "not" necessarily indicate their authentic characteristics or personalities but merely describes their participation in opposite communities and our first impression from it. Driven by our curiosities about JHs, we investigate the following research questions:

- **RQ1.** Who are JHs, and what are their characteristics?
- **RQ2.** How unique and distinguishable are JHs' identified characteristics? Can machine learning models detect JHs from ordinary users?
- **RQ3.** Why do JHs participate in both subreddits and how do their activities affect each other?

To answer these questions, we take a three-fold methodological approach. In the first stage, we scrutinize JHs' demographic and linguistic traits and activity patterns through exploratory data analysis and text analysis on collected user data through the Reddit API. In the second stage, based on our observations and findings from Stage 1, we build several predictive models using machine learning and evaluate how accurately we can distinguish JHs from other ordinary users. By separating predictors into user-level (e.g., demographic characteristics, activity pattern, and overall popularity) and content-level features (e.g., content's readability, length, tone, and language use), we try to see how well the languages used by JHs in each subreddit acculturate to the subcommunity's own tone and atmosphere. Lastly, in the third stage, we conduct user studies, including an online survey ($N = 39$) and one-on-one online interviews ($N = 8$) with the recruited JHs to delve into their motivation and ascertain the interaction between their opposing activities.

Our exploratory data analysis results show that JHs are generally more active and popular than ordinary users in terms of the number of postings and comments and their activity score (often referred to as *Karma Scores* on Reddit). However, they show different linguistic characteristics when participating in the two subreddits. Their linguistic tone and traits are akin to ordinary users of each subreddit (**RQ1**). It implies that JHs tend to follow each community's tone and rules instead of sticking to their unique linguistic habit.

Using predictive models, we assessed the validity of our findings about JHs' characteristics. We identified that JHs' user-level characteristics are distinguishable from ordinary users, but their content-level characteristics are not (**RQ2**). These results empirically support our findings to **RQ1**.

Finally, in connection with JHs' motivations for participating in both subreddits, encouragement and entertainment were JHs' most significant motives for r/ToastMe and r/RoastMe, respectively. They also recognized each community's rules well and appreciated their importance (**RQ3**). Regarding the interaction between the two subreddits, we identified a couple of possible interconnections. The antecedent r/ToastMe activities can make JHs use milder language in the ensuing r/RoastMe activities. Also, in a reverse way, it is observed that JHs use r/ToastMe to assuage feelings of guilt after they leave caustic and hurtful remarks on other users' r/RoastMe posts (**RQ3**). Based on this observation, we present that participation motivation in one community can be shaped by activities in another, especially to redeem or offset their prior actions.

Taken together, we conclude that JHs are "not" real Jekyll and Hyde, who cannot control the dark personality at his discretion. Instead, JHs are digital culture-savvy users who present online personalities in various ways suitable for each online community. Because these users follow the community rules and know how to enjoy each community's culture and tone, any type of self-presentation, even from Hyde's side, can be accepted and regarded as socially valued humor.

## Related Work

Cyberpsychology, examining the psychological aspects of "technologically interconnected human behavior," has grown with the rise of information and internet technologies (Ancis 2020; Attrill-Smith et al. 2019). Most of all, having penetrated our daily lives and transformed traditional interaction and communication patterns among individuals, online communities fuel this newborn scholarship further. In this section, we draw insights from three key research areas in cyberpsychology to elucidate JHs' behaviors.

### Multiple Selves in Online World

Online communities enable users to meticulously craft their self-images, offering a window into their ideal and multiple selves in contrast with their real identities (Shen, Brdiczka, and Liu 2015). Researchers have found that individuals employ varied strategies, like obfuscation and using multiple accounts, to adjust their online personas dynamically (Marwick 2013). Studies further validate the link between personal traits and online self-presentation (Fullwood, James, and Chen-Wilson 2016; Strimbu and O'Connell 2019). Adolescents with a stable self-concept tend to portray an online self in line with their offline personas. Factors like more time on Facebook and fewer friends lead to the display of multiple online identities. Similar tendencies were identified among young adults aged 18-35, where one's higher self-concept correlated with fewer online personas (Strimbu and O'Connell 2019).

The phenomenon of people presenting multiple selves in cyberspace can also be attributed to the online disinhibition effect. (Joinson 2007; Suler 2004). This effect, where users manage their images more liberally and communicate with fewer reservations, is primarily driven by factors like anonymity (or pseudonymity) (Christopherson 2007; Dumont and Candler 2005; Hollenbaugh and Everett 2013; Kabay 1998; Suler 2004) and further magnified by invisibility, lack of eye contact, and synchronicity (Lapidot-Lefler and Barak 2012, 2015; Suler 2004). This disinhibition bifurcates into benign and toxic effects (Suler 2004). While benign disinhibition encourages sharing personal experiences for understanding and healing, the toxic effect prompts individuals to express negativity, use coarse language, and explore internet's darker alleys. The subtle difference in the interaction of the core disinhibition factors can induce toxic (Lapidot-Lefler and Barak 2012) or benign disinhibition effect (Lapidot-Lefler and Barak 2015).

### Community Norms and Tone

Community norms and tone are pivotal in influencing users' perceptions of interactions as either benign or toxic. For instance, in some online contexts like r/RoastMe, personal attacks, which are generally seen as antisocial, can be reframed as harmless jokes if the community's ambiance supports such behavior (Allison, Bussey, and Sweller 2019). This contrasts starkly with typical real-world interactions. This phenomenon can be understood through two lenses: the social cognitive theory of morality (Bandura 2014) and the benign violation theory of humor (McGraw and Warren 2010). According to the former, individuals develop their moral standards through ongoing interactions with others and the environments they belong to (Bandura 2014). In spaces like r/RoastMe, moderators reinforce explicit norms

and rules, and their consistent interventions guide members towards the community's unique moral standards.

The benign violation theory, meanwhile, posits that certain threats can be interpreted as humorous if they are perceived as sufficiently benign (McGraw and Warren 2010). Three key mechanisms determine this benignity: the weakness of the violation, the existence of an alternative norm explaining the violation, and the psychological distance from the violation (McGraw and Warren 2010). Within the r/RoastMe community, these conditions are mostly met. In particular, their principle of "comedy, not hate" frees users from guilt when making hurting comments about others. In this case, the second mechanism of the benign violation theory reinterprets their moral disengagement as a benign and safe one (Kasunic and Kaufman 2018).

### Motivations for Multi-Community Engagement

The prior studies introduced above offer insights about individual users, such as JH, who actively engage in multiple online communities, showcasing diverse personalities across these communities. However, the motivations for the participation in several communities concurrently remain elusive. The researchers pinpointed three primary needs that users seek from online communities. These include the desire to access specific information and engage in discussions, the intent to connect with others harboring similar interests, and the pursuit of attention from a broader audience through content posting (TeBlunthuis et al. 2022). Because a single community often fails to meet all these needs, it prompts users to diversify their participation in multiple communities (Hwang and Foote 2021; TeBlunthuis et al. 2022; Zhu, Kraut, and Kittur 2014). For instance, niche communities that center around specialized purposes or interests can sometimes resonate more profoundly with users seeking specified knowledge or a more supportive community ambiance (Hwang and Foote 2021).

Moreover, understanding users' motivations and engagement patterns can provide important implications for community managers and designers. Some studies have identified specific behavior patterns among online community users, such as "wandering" behavior (Tan and Lee 2015) and "reshaping" behavior (Butler and Wang 2012). With the former behavior, users exhibit a propensity to continuously seek out and engage with new online communities (Tan and Lee 2015). Meanwhile, "reshaping" behavior refers to users behavior cross-posting identical content across various communities. Such behavior can have a dichotomous impact on a community (Butler and Wang 2012). On the positive side, reshaping tends to correlate with an increase in the number of newcomers attracted to the community. However, it appears to inversely correlate with the retention of existing members. This phenomenon could be attributed to the dilution of a community's distinctiveness when content is ubiquitously shared across various communities. Such dilution, in turn, discourages current members from staying (Butler and Wang 2012).

To summarize the previous studies' findings, which could be initial clues for explaining JH's behavior, (i) people feel more uninhibited in cyberspace than in the real world due to anonymity, lack of eye contact, invisibility, and asynchronicity; (ii) these causes interact with each other and lead to benign or toxic disinhibition effect; (iii) the more unstable the real-world self-concept is, the more divergent online multiple identities appear; (iv) online community's norms and tone promote the disinhibition effects; (v) users of online communities frequently seek out and engage with new communities in an ongoing effort to fulfill their needs and desires; (vi) the actions and behaviors of users play a pivotal role in determining the evolution and content boundaries of a community. Based on these findings, our study dives into tracing JH's online behavior to understand why they reveal starkly opposing personalities simultaneously and what systematic and mental mechanisms are behind their behavior.

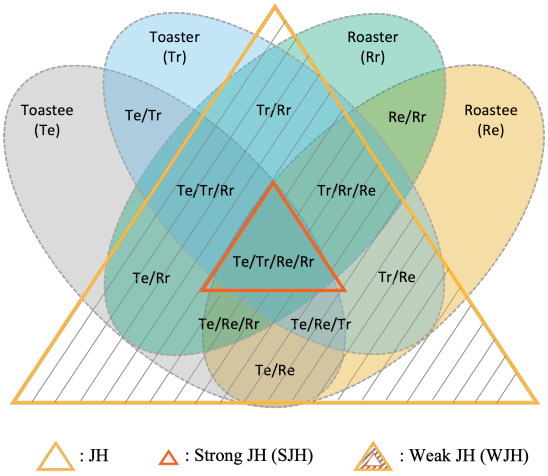


Figure 1: ToastMe and RoastMe User Types. There are four user types in terms of activities a user can partake in: Toastee (Te), Toaster (Tr), Roaster (Rr), and Roastee (Re). From these categories, a total of 15 unique user types can be derived based on a user's activity combination. The JHs denote a group of users who engage in at least one user type from each subreddit (users inside the yellow triangle).

## Definition and Data

### Definition

Before diving into the three research questions, we define several key terminologies for convenient reference.

**Toastee/er, Roastee/er, and Jekyll and Hyde user (JH)** Following the convention of previous literature which studies r/ToastMe and r/RoastMe users (Dynel and Poppi 2020; Kasunic and Kaufman 2018), we call r/ToastMe and r/RoastMe users Toaster, Toastee and Roaster and Roastee. A Toastee and Roastee indicate a user who posts their own content on each r/ToastMe and r/RoastMe, and a Toaster and Roaster indicate a commenter who comments on Toastee's and Roastee's posts, respectively. JHs are Reddit users participating in both subreddits. Thus, as shown in Figure 1, JHs can be defined as every possible combination of nine subgroups that include at least one user type from each subreddit. In Figure 1, users inside the yellow triangle are gen-

eral JHs (the union of the nine subgroups). To compare JHs with ordinary users, we use the terms Toaster, Toastee, or Roaster and Roastee. These ordinary users[1] participate only in r/ToastMe or r/RoastMe and belong to one of the six subgroups outside the yellow triangle, depending on their role in each community.

**Strong and Weak JHs** To better understand JHs by their participation degree, we distinguish strong JHs (SJHs) from the general JHs. SJH refers to a JH who plays every role–*roastee/er* and *toastee/er*–in both subreddits. The red triangle area in Figure 1 indicates SJH. We call the remaining JHs Weak JHs (WJHs), who make either posts or comments in both subreddits besides SJHs. The grey shaded area in the figure indicates WJHs.

## Data

We collected all user data of r/ToastMe and r/RoastMe through the Reddit API since each community was created (*r/ToastMe*: Jul 1, 2015 - July 1, 2020, *r/RoastMe*: Aug 3, 2015 - July 1, 2020). Considering users' acquaintance with the community cultures, we truncated users who made posts or comments less than three times. In addition, we removed three types of outliers to reduce any potential information bias. Those outliers embrace (i) each community's moderators, whose comments are mostly warning messages to rule violators, (ii) users who canceled their accounts during our research period since their information is no longer available, and (iii) Reddit AutoModerators, which are bots that help moderate the communities. After sorting out these observations, we have a total of 3,985,686 posts or comments made by 265,112 unique users, and out of them, 17,337 users are JHs: 1,356 SJHs and 15,981 WJHs as summarized in Table 1.

| | # of Users | % |
|---|---|---|
| Ordinary Roastee/er | 221,543 | 83.57 |
| Ordinary Toastee/er | 26,232 | 9.89 |
| JHs | 17,337 | 6.54 |
| SJHs | 1,356 | 0.51 |
| WJHs | 15,981 | 6.03 |
| Total | 265,112 | 100 |

Table 1: The Number and Proportion of Users

# RQ 1: Who are JHs?

## Demographic Traits

First, we examined JHs' demographic traits and compared them with ordinary users. Because users upload their photos when posting, not commenting, only toastees' and roastees' photos are available. We identified 38,264 users with photos, which is about 14.4% of all users in our dataset.

[1]In our context, the "ordinary users" group refers to those who are not affiliated with JH, distinguishing them from the general notion of average users.

To identify the age and gender, we analyzed their frontal photos using Microsoft Azure's facial recognition API. [2] Given this reduced accuracy, we limited the application of facial recognition results to roughly fathom and compare each user group's age (or generation). As a result, the average age of JHs is 25.2 years old, and the average age of ordinary Roastees and Toastees are 25.0 and 25.1 years old, respectively. As shown in Table 2, the standard deviations indicate there were no significant age differences among user types. Regarding the gender ratio, there are only 19.2% of female users in ordinary Roastees, while 53.5% in ordinary Toastees. Among JHs, 33.4% are female users, and more female users are identified (38.6%) for SJHs. From these demographic traits, we infer that female users tend to participate more in the r/ToastMe community than in r/RoastMe. JHs are more gender-balanced than ordinary RoastMe users but less balanced than ordinary ToastMe users. Also, we make a guess that most users, regardless of their related subreddits, are the digital generation who have grown up with the internet. Despite the low accuracy of the age inference technique, this age estimate is broadly consistent with the survey respondents' age groups, where 56.4% of them are 18 to 24, and 38.5% are 25 to 34. The chi-squared test also shows that there is no statistically significant difference in the age proportions.

| | Average Age (SD) | Female User Ratio (%) |
|---|---|---|
| Ord. Roastees | 25.0 (6.8) | 19.2 |
| Ord. Toastees | 25.1 (6.4) | 53.5 |
| JHs | 25.2 (6.6) | 33.4 |
| SJHs | 24.6 (5.8) | 38.6 |

Table 2: Average Age and Female Ratio (%)

## Activeness and Popularity

Next, we investigated JHs' subreddit activity by comparing their average posts and comments with ordinary RoastMe and ToastMe users. Figure 2(a)-(d) reveals SJHs as the most active, averaging 17.38 comments and 2.15 posts in r/RoastMe, and 12.79 comments and 1.57 posts in r/ToastMe. Notably, SJHs posted three times more than ordinary users in both communities (Figure 2(c) and (d)). While SJHs consistently outperform regular users in activity, WJHs primarily focus on commenting in r/RoastMe. The key difference between SJHs and WJHs is posting frequency, with most WJHs lacking posts in either subreddit.

JHs are not just more active, they are also more popular

[2]Azure Face API exhibited a 97% accuracy for gender inference and 45% for age inference (Jung et al. 2018). To scrutinize the limited precision in age prediction, we examined the discrepancies in estimated ages for a cohort of 25,487 users, each of whom had uploaded a minimum of two distinct frontal images within the same calendar year. Our findings indicated that the mean standard deviation stood at 2.03 (median of 1.41), with a 95% confidence interval ranging from 0 to 7.54. This implies that the Azure Face API may produce varying age estimations for the same user with an average difference approximating two years.

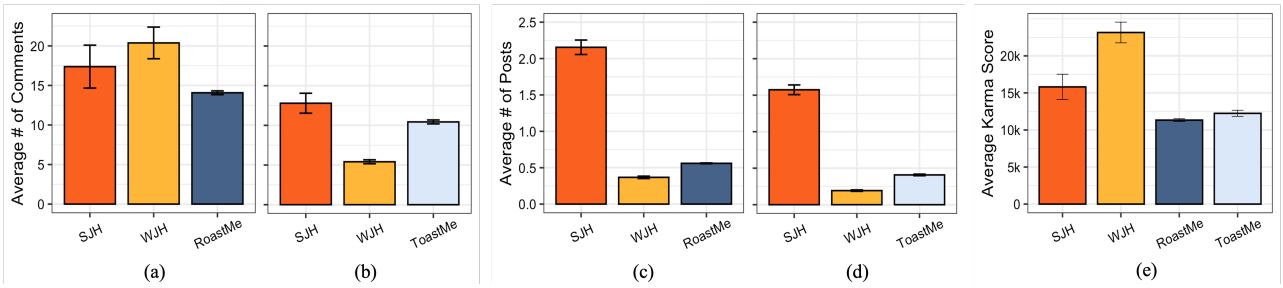


Figure 2: Average Number of Comments in (a) r/RoastMe and (b) r/ToastMe; Average Number of Posts in (c) r/RoastMe and (d) r/ToastMe; (e) Average Karma Score. SJHs are represented in orange, WJHs in yellow, and ordinary RoastMe and ToastMe users in navy and light blue, respectively. It is evident from the figure that SJHs exhibit a markedly higher level of activity and popularity compared to ordinary users.

than regular users. By examining the *Karma Score* (KS)[3], Figure 2(e) indicates JHs are more popular in both communities. Interestingly, despite fewer community contributions in terms of posting, WJHs have a higher KS than SJHs. Although the Reddit API doesn't offer KS breakdowns by subreddit, a plausible explanation is that Comment Karma—relatively easier to accumulate through commenting on r/RoastMe—constitutes a major portion of the KS. Given that WJHs are most active in commenting on r/RoastMe, this could elevate their overall KS above that of SJHs. However, what must be highlighted here is that both SJHs and WJHs tend to receive more upvotes and interact with other users more actively by exchanging awards. This suggests they immerse themselves in the community culture and enjoy it more than ordinary users.

Since SJHs are more distinguishable from ordinary users in terms of involved activity types (both posting and commenting), we will primarily focus on SJHs over WJHs for the rest of this study.

## Linguistic Characteristics

To identify SJH's linguistic characteristics, we applied the LIWC-22 dictionary (Boyd et al. 2022) to both subreddit users' posts and comments. The LIWC-22 dictionary has a hierarchical structure of various features representing one's linguistic traits. It provides eight composite variables each of which summarizes the following aspects of individual's linguistic characteristic: the total word count, metric of logical thinking, the language of leadership, perceived genuineness, degree of positive and negative tone, percentage of 7-letter-or-longer words, and percentage of words captured by LIWC (Boyd et al. 2022).

Figure 3 illustrates the results of the eight summary features about SJHs and ordinary user groups. The solid red line represents SJHs' linguistic traits when they participate in r/RoastMe, and the solid blue line represents the same features when they are in r/ToastMe. The red and blue dotted lines refer to ordinary users in each subreddit. 3(a) and (b) describe the subredditors' linguistic traits when they post content and comment on others' posts, respectively.

The most recognizable point from the figure is that SJHs' linguistic characteristics closely mirror those of ordinary users in each subreddit as seen by the well-aligned solid and dashed lines of both colors in Figure 3(a) and (b). However, regardless of SJHs or ordinary users, there are noticeable differences within themselves comparing when they are in r/RoastMe and when in r/ToastMe, highlighted by the different shapes of colored lines between Figure 3(a) and (b). In other words, they assimilate into each subreddit's environment and tone instead of sticking to their own linguistic habits. This observation well aligns with the findings of a prior study, which explored the "situation" vs. "personality" debate. The study concluded that online community members' language is predominantly influenced by the community's ambiance, leading them to adopt varying linguistic patterns across different communities (Tan and Lee 2015). Our study not only reconfirms this pattern but also highlights its presence even in two diametrically opposed communities.

Other findings worth noting are the linguistic differences between the r/RoastMe and r/ToastMe communities. The r/ToastMe users tend to make more lengthy posts and comments and use a far more positive emotional tone when commenting on others' posts. When they post content, the r/ToastMe users use less language of leadership (Clout). That is, their language is less confident but humble and even anxious (Pennebaker et al. 2015). In contrast, the r/RoastMe users' comments reflect more logical and formal thinking (Analytic) than r/ToastMe users. This makes sense because people usually use more informal and personal language when encouraging other users instead of having an analytical and logical conversation. Conversely, when roasting other people in r/RoastMe, users tend to analyze the Roastee's appearance, gestures, clothes, and even the background of their photo and show more hierarchical thinking in their conversation.

[3]The KS on Reddit comprises four sub-Karmas: Link, Comment, Awardee, and Awarder. Users gain or lose Link and Comment Karma based on upvotes or downvotes on their posts and comments. Introduced in July 2020, Awardee Karma is earned when a user receives an award, with costlier awards granting more points. Conversely, Awarder Karma rewards those who give out awards to promote quality content.

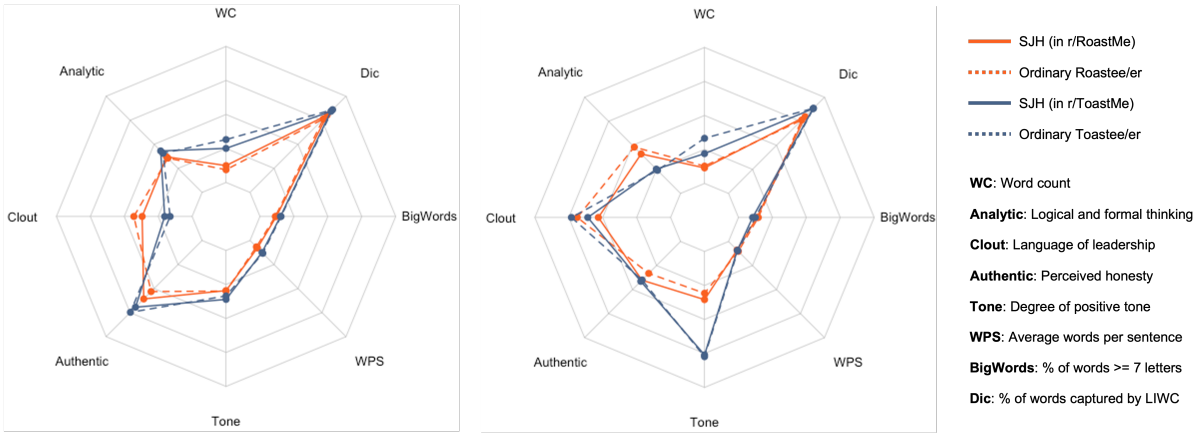


Figure 3: Linguistic Characteristics: (a) posting and (b) commenting. Regarding both posting and commenting, the style, tone, and sentiment of the language employed by SJHs align closely with that of ordinary users within each community, rather than adhering to a distinct language of their own.

### Temporal Behavior Patterns

In this subsection, we investigate SJHs' activity log, such as how frequently they return to the community and what they do (posting or commenting) when they return. By separating their activity patterns into macro- and micro-temporal patterns, we tried to look deeper into their behavior changes in flow over time. The macro-temporal resolution aims to find SJHs' overall activity patterns by ignoring their daily (within 24 hours) consecutive activities. That is, it only keeps track of SJHs' first activity each day. Conversely, through the micro-temporal resolution, we intend to uncover the patterns of their daily activities on a specific day.

Figure 4(a) and (b) are alluvial diagrams correspondingly describing SJHs' macro- and micro-temporal behavior patterns over time. On the horizontal axis, DA stands for daily activity, and $DA_{np}$ refers to the p-th activity during the user's n-th instance of participation. Each bar comprises four blocks representing different activities: commenting and posting in r/RoastMe and r/ToastMe, respectively. Figure 4(a) shows how SJHs start their daily activity by tracking the first daily activity they are involved in during their first five times of participation. From the figure, it is evident that over half of SJHs (52.2%) initiate their JH activity by posting their photo on r/RoastMe and pleading for random roasting ($DA_{11}$). Following this, they have traveled back and forth between r/RoastMe and r/ToastMe. In contrast, as shown in Figure 4(b), most SJHs are inclined to stay in the same subreddit during their consecutive daily activities on a specific day ($DA_{k1}$ to $DA_{k5}$).

These temporal patterns of SJHs, in conjunction with the results of SJHs' linguistic traits, show that they, instead of flitting between the two subreddits for a short time like the example we have seen in the introduction, tend to settle down to one subreddit and join the conversation with assimilated languages to each community. They move to the other subreddit another day. Combining these findings with their demographic and activeness traits from previous sections, we infer that a JH, especially SJH, is a digital-savvy who has grown up with digital technology and knows how to enjoy a different kind of online communities by adapting themselves to the community's unique tone. Thus, they are generally well-recognized and popular in the communities.

## RQ 2: Detecting SJHs

So far, we have found JHs' distinct characteristics, with a particular focus on SJHs, as follows: They frequently visit both subreddits and actively interact with other users by posting content and commenting on others' posts. They are well recognized in the communities by other users, and their contents, both postings and comments, are usually more popular than others. They tend to encourage other users to produce better content and also enjoy finding quality content earlier than others. Their language is similar to ordinary users of each community. They often move from one community to another over a long time interval but tend to stay in the same community within a shorter time interval.

Based on these findings and observations, in this section, we attempt to develop predictive models detecting SJHs from ordinary users. These models aim to provide a general assessment of the identified SJHs' traits and their probabilistic robustness in terms of ROC-AUC, or the area under the ROC curve. This single metric summarizes test performance across every possible threshold value, and a higher ROC-AUC means the model distinguishes positive classes (SJHs in our case) better (Halligan, Altman, and Mallett 2015). Most of all, ROC-AUC considers both sensitivity and specificity, and it is a standard metric to evaluate predictive models on imbalanced datasets. To examine SJHs' distinctive characteristics in a more multifaceted manner, we designed a $4\times2\times3$ prediction framework (confrontation $\times$ feature level $\times$ algorithm) as illustrated in Figure 5.

Each of the four confrontations (*C1* to *C4*) represents

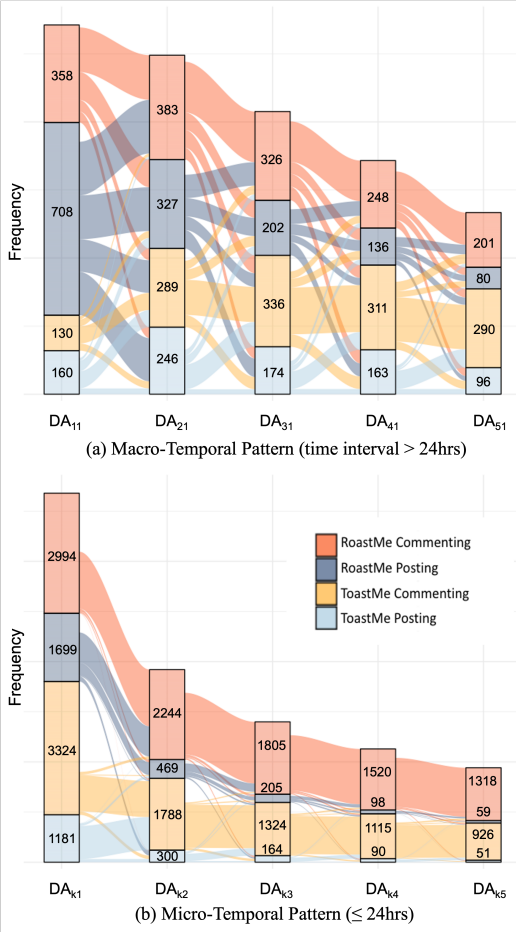


Figure 4: SJHs' Temporal Behavior Patterns. $DA_{np}$ denotes the p-th daily activity (DA) during the user's n-th instance of participation. For instance, $DA_{21}$ represents the first activity during the user's second participation. Observations from the figure make it clear that SJHs mostly gravitate towards one subreddit, engaging in consecutive conversations as their daily activity, as in (b). They typically transition to the other subreddit on subsequent days, as in (a).

two opposing user groups depending on their subreddit (r/RoastMe or r/ToastMe). *C1* and *C2* compare SJHs with ordinary users within each subreddit, while *C3* and *C4* compare user groups across the two subreddits. We separated all identified characteristics into two levels: user level and content level. The user-level features are derived from individual user's Reddit accounts and provide basic activity information, such as frequency of visits, four kinds of *Karma Scores*, user titles (whether or not the user is a moderator in other subreddits; the user is Reddit premium member; and the user is Reddit employee), the average number of self-comments, and whether or not the user has been blocked before. The content-level features identify the average characteristics of users' comments. They include the eight LIWC linguistic features in Figure 3 and readability.

Based on each feature set, we applied three machine

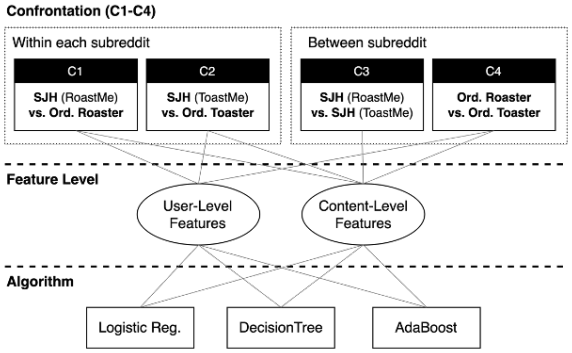


Figure 5: Framework of Predictive Models. We considered four distinct scenarios (confrontations) across two feature levels: user-level and content-level. The user-level features encompass aspects of a user's account information, including visit frequency, Karma scores, user titles, etc. Meanwhile, the content-level features pertain to the linguistic characteristics and readability of a user's postings. To ensure robustness, three algorithms were employed.

learning (ML) algorithms: Logistic Regression, Decision Tree, and AdaBoost. Because our goal in this section is not to achieve the best prediction performance but to compare the performances among different confrontation and feature level combinations, we used classical ML algorithms instead of more sophisticated ones such as deep learning algorithms. This $4\times2\times3$ prediction framework finally gives us 21 indicators of assessing SJHs' traits (21=24−3 indicators. Note that, in Figure 5, *C3* (RoastMe SJH vs. ToastMe SJH) does not connect with the user-level features since RoastMe SJHs and ToastMe SJHs share the same user-level features). Also, we have implemented several methodological preemptive measures. First, we up-sampled SJHs using the popular SMOTE algorithm (Chawla et al. 2002) on our training set because the proportions of SJHs are extremely small (0.51%). Also, we only focused on commenting since SJHs' language seems more distinct when they comment than when posting (see Figure 3). Lastly, we used 5-fold cross-validation and grid search techniques for tuning our hyper-parameters.

Figure 6 summarizes the results of our predictive models. Triangular points represent the average AUC of three ML algorithms for each combination of confrontations and feature levels. The interval shows the gap between the best- and the worst-performing model. As shown in *C1* and *C2*, SJHs' content-level traits are relatively less distinct than their user-level traits. As we have seen in SJHs' linguistic traits, they use language similar to that used by ordinary users of each subreddit. Therefore, the predictive models based on their linguistic traits show less accurate performance for both r/RoastMe (*C1*) and r/ToastMe(*C2*), where the AUC ranges from 0.627 to 0.688 and 0.552 to 0.694, respectively. Meanwhile, the user-level-based predictive models show better performance, as high as 0.797 and 0.845 for each subreddit. It becomes even more straightforward when it comes

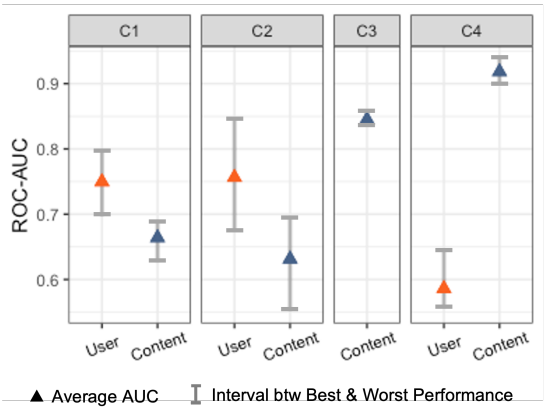


Figure 6: Results of Predictive Models. The prediction models' average performance is represented by red triangular points for user-level features and navy points for content-level features. In scenarios *C1* and *C2*, which involve distinguishing SJHs from ordinary users within each subreddit, the user-level features proved more distinguishable than those at the content-level. Conversely, in *C4*, the distinction based on content-level features was more pronounced between users, including SJHs, across the two subreddits.

to the direct comparison between r/RoastMe and r/ToastMe (*C3* and *C4*). In *C3*, where we distinguish SJHs in r/RoastMe from the same users in r/ToastMe based on their content-level features, we can clearly confirm that SJHs use completely different language in terms of not only the tone (positive or negative) but even general linguistic style. Moreover, by comparing ordinary Roasters and Toasters in *C4*, we find that their user-level traits are not quite different from each other (the AUC ranges 0.556 to 0.643), while their contents are more distinctive (0.898 to 0.938).

The results of *C1* to *C4* empirically support the SJHs' characteristics we investigated in the previous section. SJH's user-level characteristics–i.e., being more active and popular in both subreddits–are distinct so that we can detect SJHs well using the characteristics as predictors. Conversely, the linguistic traits captured in their content are not differentiated from ordinary users, and our models showed suboptimal performances.

## RQ 3: Why Do SJHs Do What They Do?

To better understand SJHs' behavior and their motivation, we conducted an online survey and one-on-one interviews with SJHs. In particular, we scrutinize the motivations behind JHs' participation in the two contrasting communities, employing a two-step approach. Initially, we conducted an online survey on 39 SJH users. This survey primarily consisted of multiple-choice questions supplemented by a few open-ended queries, aiming to obtain a comprehensive yet standardized understanding of JHs' participation and activities. Drawing insights from this survey, we then formulated an interview protocol to dive deeper into their motivations and the interplay between the two subreddits.

### Participants

We recruited the survey participants from the SJH user list. We sorted the list by the total number of posts and comments and contacted the most active SJHs (at the top of the list) first through Reddit's private messages. Of 1,000 invitees, 57 had started the survey, and 39 (68.4%) completed it. All participants who completed the survey were compensated with a $5 Amazon gift card ($8.77 per hour on average). Then, we invited all 39 survey participants to our interview session, and 8 (P1 to P8) of them accepted and had a 30-minute one-on-one interview. The median time for the interview was approximately 16 minutes. They were additionally compensated with a $20 Amazon gift card.

### Procedures

**Online Survey** The survey was conducted from December 2021 to July 2022. The online survey included five parts: (i) participation history and frequency; (ii) motivations for participating in each subreddit; (iii) interaction between two subreddits; (iv) adherence to community rules; and (v) demographics. For the participation history and frequency in each community, we questioned *how long they have been an active user and their visit frequency in a weekly manner*. For the motivation in participating in each subreddit, we gave them multiple-choice questions about *why they participate in r/RoastMe* ('To humiliate the roastee,' 'To join the wordplay with other roasters,' 'To gain more karma score' or 'Other') *and r/ToastMe* ('To encourage or comfort someone,' 'To make a return for what you were consoled by other users before,' 'To make myself feel comfortable for whatever reason,' 'To gain more karma score' or 'Other'). Also, we asked respondents who chose the 'Other' option to provide open-ended answers. To identify if an interaction exists between the two subreddits, first, we questioned *if participation in one subreddit affects the other subreddit's participation*. If they agreed with the influence, we asked them an open-ended question about *how one subreddit's participation affects the other*. For adherence to community rules, first, we asked them *whether they knew about the rules of each community*. Then, we questioned *how much they knew about the rules* using a 5-point scale. Toward the end of the survey, we asked about their demographics, such as age, gender, race, and educational level.

**Online Interviews** The one-on-one online interviews were implemented between May and July 2022 to delve deeper into JHs' motivations and potential interactions between their two contrasting activities. All sessions were audio-recorded and subsequently transcribed by two of the authors. We adopted an inductive approach to our analysis, where we initially focused on the participant's individual stories relating to each interview subsection. From these narratives, we conceptualized and delineated essential themes that aligned with our study's objectives through narrative analysis. Furthermore, the interviews provided participants with an opportunity to clarify and expand upon their survey responses to complement the survey data. Below, we present the survey results and identified themes for each subsection.

## Results

**Participation History and Frequency** Most survey respondents have participated in r/ToastMe and r/RoastMe for at least one year (61.5% and 69.2%, respectively), and a fourth and a half of the respondents have been members of each community for more than two years. Regarding the frequency of visiting both subreddits, 94.9% and 92.3% of respondents visit r/ToastMe and r/RoastMe less than or equal to twice per week, respectively. Only one respondent answered that he/she visits r/ToastMe every day. Also, for an additional question about their daily activity, all eight interviewees answered that they stay on either subreddit for less than one hour during their daily activity. Taken together, JHs are relatively long-standing members in both subreddits and consistently use the communities. Although they are usually more active than ordinary users, they do not use both subreddits obsessively; instead, they enjoy the communities as a lighter and milder daily activity.

**Motivation of Participation** Figure 7 summarizes the survey results about JHs' participation motivation. With multiple selections allowed in questions, 92.3% of respondents considered "Encourage others" as a main motivation for using r/ToastMe, and "Make a return for the encouragement they had received before" and "Feel comfortable" were also popular choices: 43.6% and 30.8%, respectively. Meanwhile, 87.2% of respondents picked "Join wordplay," and four (10.3%) selected "Humiliate others" for r/RoastMe. For both subreddits, "Gain Karma Score" is not a major purpose, which implies that gaining points or upgrading one's reputation within subreddits is not necessarily the primary motivation for their participation. Of the six respondents who chose "other," three answered they use r/RoastMe to find something funny and laugh. It shows how a malicious r/RoastMe comment can be accepted as socially valued humor (Allison, Bussey, and Sweller 2019). Once the threat is transformed into a socially valued one through either (or both) the social cognitive theory of morality (Bandura 2014) or (and) the benign violation theory of humor (McGraw and Warren 2010), it becomes a driving force fueling participation motivation.

In the one-on-one interviews, we separately asked about the interviewees' motivation for commenting on others' posts and posting their own content. We reconfirmed that they comment on r/RoastMe for entertainment and wordplay and on r/ToastMe for encouragement and empathy. An interviewee (P5) who considered entertainment as a primary purpose of using both communities pointed out the difference between the two types of entertainment. They described "*[ToastMe] is the kind of entertainment you get when you watch a cat be cute, and RoastMe is more like the entertainment of watching a train wreck*" (P5).

**r/RoastMe: Toxic Disinhibition.** Regarding their motivation for posting on r/RoastMe, they post their photos and ask random roasting to see what other people think of their appearance (P1), and just to be fun to have comments from others (P3, P4). They are not too worried about getting hurt by others' comments and ensure they are able to amuse themselves through wordplay. "*I was very confident that whatever anybody was going to throw at me wasn't going to hurt my feelings, and it didn't. I laughed at all the roast. And I ended up having a good time with it*" (P7). One important point is that Roasters' toxic disinhibition is not accepted as really toxic by a Roastee. Roastees post their photos expecting to get harsh language from strangers and being prepared to enjoy them. Indeed, it is not difficult to observe that Roastee responds to Roasters' wordplay by assessing the creativeness and severeness of their comments. It shows the r/RoastMe community can become a gateway for users to quench their thirst with the online disinhibition desire and relieve hatred and stress.

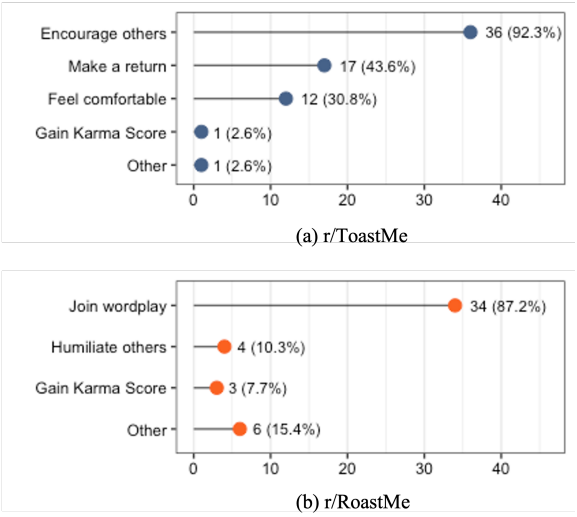


Figure 7: JHs' Participation Motivation. When multiple choices were allowed, "Encourage others" and "Join wordplay" emerged as the predominant reason for engagement in r/ToastMe and r/RoastMe, respectively.

**r/ToastMe: Benign Disinhibition.** In contrast, most respondents post on r/ToastMe when they feel down, and need encouragement or "*ego boost*" (P1, P2, P3, P4). P4 flashed back to their first time to use r/ToastMe, "*I post in there (r/ToastMe), I think I just did really poorly on an exam and I was feeling really down about myself. So I was like I'll just go there and talk to other people*" (P4). This is a typical example of a benign disinhibition effect in cyberspace (Suler 2004; Lapidot-Lefler and Barak 2015). Individuals disclose their personal information, not only their frontal photo but their health status, personal issues, and distress, and then random strangers commiserate with the Toastees.

**Interaction between r/RoastMe and r/ToastMe** If we look at these interviewees' behaviors and motivations in both subreddits separately, nothing might be peculiar. However, when we recall that the interviewees are SJHs and join both subreddits with the identified motivations simultaneously, it is reasonable to question if the two starkly opposing behaviors can affect each other. We look into whether one's activities in r/RoastMe affect those in r/ToastMe and vice versa. The survey results show that 61.5% of the respondents answered that there is no interaction between the two subreddits, while 38.5% said there is (28.2% said

only r/RoastMe affects r/ToastMe and 10.3% vice versa). Through the following one-on-one interviews, we further explored how activities in two subreddits act on each other. Three interviewees (37.5%) mentioned that there is no interaction, four (50%) mentioned that r/RoastMe activities affect participation in r/ToastMe, and one (12.5%) indicated the opposite direction. We identified two possible interactions between the activities of two subreddits.

**Influence of r/RoastMe on r/ToastMe.** First, we found that r/RoastMe activities can directly affect activities in r/ToastMe. Four interviewees mentioned that making hurtful comments on Roastees' posts makes them feel bad, and commenting on r/ToastMe posts helps them alleviate the guilt (P1, P3, P4, P6). "*When I feel like I've been mean enough in r/RoastMe, I will instead go to r/ToastMe to make myself feel a little bit better about what I did and commented on the other one on r/RoastMe*" (P6). Also, P1 answered similarly, "*if I feel bad about comments in r/RoastMe, then comments in r/ToastMe...(it) would make me feel better.*" In other words, activities in r/RoastMe give them the motive to participate in r/ToastMe. In the survey, about 30.8% of the respondents also considered "feel comfortable" as their primary purpose for using r/ToastMe, and it could include a similar context. Thus, r/RoastMe activities can promote r/ToastMe activities.

**Influence of r/ToastMe on r/RoastMe.** The other interaction occurs in the reverse direction. One interviewee alluded to their milder language when commenting on r/RoastMe after using r/ToastMe (P5). Participating in r/ToastMe makes them more considerate of other users' feelings, possibly creating empathy in themselves. It, in turn, allegedly seems to mitigate their ability to roast strangers mercilessly.

To sum up, although a majority (61.5%) of the survey respondents did not report any interaction between the two opposing activities, still, 38.5% had experienced that their activities in one subreddit affected the other. Moreover, we found that both directions affecting one another are plausible. This implies that SJHs wisely relieve their online disinhibition desire, both benign and toxic, by fully utilizing two opposing communities for their purpose. These findings conflict with the results of previous studies, where divergent online self-presentation was attributed to unstable real-world self-concept (Fullwood, James, and Chen-Wilson 2016; Strimbu and O'Connell 2019). Instead, SJHs' behavior shows one possible interpretation of divergent online personalities. It can result from one's concomitant participation in various online communities, who knows each community culture well and seeks to satisfy their insatiable appetites for online disinhibition in a sound and safe manner. Furthermore, this interaction offers insights into the literature on multi-community engagement. Prior research (TeBlunthuis et al. 2022) has introduced users' motivations for engaging in multiple communities as stemming from three primary desires: to access specific information, to connect with like-minded individuals, and to share content with a wider audience. Building upon the observed interplay between activities in the two contrasting communities, we propose that participation motivation in one community can be shaped by activities in another, especially to redeem prior actions in the former community.

**Adherence to Community Norms** According to the benign violation theory of humor, community norms are essential in the process where inappropriate and caustic humor is reframed as socially acceptable one (Allison, Bussey, and Sweller 2019; Kasunic and Kaufman 2018; McGraw and Warren 2010). Moreover, because community norms play a sandbox role, within which users can give off the disinhibition effect to their heart's content, it is essential to investigate SJH's awareness of community norms. Both r/RoastMe and r/ToastMe have explicit community rules and display them on their first page, and we asked SJHs how much they are aware of and abide by the community rules.

**Perception of Community Rules.** The survey results show that most respondents recognize that each subreddit has its own rules (94.9% for r/ToastMe and 97.4% for r/RoastMe). However, 92.3% (r/ToastMe) and 76.9% (r/RoastMe) do not know each of the rules clearly or have never read them. It implies that they hardly read the community rules when joining the community. However, they casually pick up some rules through their Reddit activities or interactions with other users, moderators, or auto-moderators. Even though they were not fully acquainted with the rules, every interviewee thought its presence is significant. P1 pointed out that the rules are essential since "*they avoid too extreme of hate, hate speech and racism.*" Also, according to the rules, all participants should post their frontal photo with a piece of paper in hand displaying their username, and it "*helps in preventing people from posting their friends and other people to make fun of them*" (P1). Another interviewee (P3) emphasized, "*the RoastMe group is for entertainment. It's not to be bullying, and there can be a fine line between banter and bullying.*" So, they think "*having guidelines...is essential to keep it*" (P3).

**Safety Boundaries.** Another critical point is that the presence of community rules makes users feel comfortable when commenting on other users' posts. P6 said, "*the rules are very good... by using the rules as a way to vent frustrations without having any guilt about stepping over the line with somebody.*" This well explains how the community norms administer to users' sandbox. Unless the online community members overstep the sandbox made by the community norms, their activities, which would otherwise be unacceptable offline, can be reframed as acceptable and even valuable (Allison, Bussey, and Sweller 2019). In other words, the online disinhibition effect, which loosens an individual's public morality sense in cyberspace (Joinson 2007; Suler 2004), can be protected to be safely realized within the boundary drawn by the norms.

## Discussion and Limitations

So far, we have explored JHs' characteristics and participation motivations. If we pass all the findings in review, we conclude that SJHs are more digital culture-savvy than ordinary users. They wisely utilize online community benefits and enjoy each community's culture within the boundaries set by the communities' norms.

These results from exploring SJHs' behavior speak volumes about online self-presentation and how it has changed among the digital generation. SJHs actively participate in r/RoastMe and r/ToastMe, abiding by each community's rules. Also, they have the apparent motivation and take advantage of both communities. It suggests that digital culture-savvy users, like SJHs, present themselves in various ways suitable for each online community and can be very different from their real-life personalities. Most of all, when users follow the online community's rules and assimilate into the community's unique culture and tone, any self-presentation instigated by the disinhibition effect can be accepted and even regarded as socially valued. One SJH interviewee described r/RoastMe as a bustling club where "*they (Roasters) are really stepping over people's toes...I bet if you met any of these people outside of the club, they'd be perfectly fine, perfectly normal. However, in the context of this club or r/RoastMe, the people get very aggressive because that's the culture of the subreddit*" (P6).

Our research implies avenues for both understanding online community user behavior and crafting healthier online environments. To start, online communities can benefit from an optimized design emphasizing clear and accessible norms. Our findings stress the importance of explicit community norms, implying that web designers should prioritize their visibility. For example, in r/RoastMe and r/ToastMe, while rules are on the front page, users must scroll down and click on the rule to grasp the specifics. Indeed, many survey participants were aware of the presence of these rules, but few comprehended them fully. A solution could involve placing these rules at the uppermost visible spot on the homepage with easily digestible text and arresting visuals. However, since some users may still overlook or forget these rules, moderators can play a pivotal role. All of our interviewees agreed with the necessity of moderators. When removing comments, moderators should provide clear reasons for their actions. For instance, instead of the vague "Comment removed by moderator" on r/RoastMe, giving a specific reason, like violation of the "Don't Be Evil" rule, can foster community learning. Hyperlinking to the related rule can enhance this effect.

Moreover, our findings on the positive interplay between activities in contrasting communities offer insights for community managers. They might consider establishing or linking to complementary communities where the tone and norms are diametrically opposite to their own, as seen in r/RoastMe and r/ToastMe. Such paired communities can synergize, allowing users to both express their disinhibition desire and find relief from any guilt stemming from prior activities, all within a healthy environment.

Lastly, our research indicates that online users may have multifaceted personas portrayed in different online spaces. The JH concept, capturing this duality, could provide a framework to understand the dynamic nature of online behaviors, particularly their fluidity and adaptability. This concept holds potential for diverse research scenarios that encounter entities with seemingly contradicting behaviors, where static categorizations might not suffice. For instance, it could be useful in studies examining a user's simultaneous benign and toxic postings across social platforms or their active engagement in two contrasting social movements.

Although our study's mixed research design allowed us to investigate our target JH users and quantitatively validate the insights from our findings, there are certain limitations to consider. First, we did not account for our target users' activities in other subreddits. There is a chance that the users are active members of other communities, meaning other subreddit activities might have affected their r/RoastMe and r/ToastMe activities. Secondly, the generalizability of our interview findings may be limited, as they stem from inductive reasoning based on a relatively small set of interviewees. Thus, extrapolating these findings to broader populations or other online communities should be approached cautiously. Finally, for future studies, we suggest measuring user popularity through the count of upvotes and downvotes rather than the Karma Score. The latter may not accurately represent a user's popularity within a single community due to privacy constraints. Also, juxtaposing users' offline personalities with their online personas, an aspect we were unable to explore in this study could be an intriguing avenue for future research.

## Conclusion

Unlike Dr. Jekyll in Stevenson's novella, JHs can easily switch from Jekyll to Hyde or vice versa. They present their personality at their pleasure in accordance with the communities' tone and alternative norms and enjoy the communities' own culture. In this regard, the community can play a positive role as an "outlet for sentiments" (Allison, Bussey, and Sweller 2019), which maximizes cyberspace's application. By the same token, r/ToastMe may also serve as a buffer zone that preemptively lessens community users' aggression on the one hand and mitigates their guilt subsequently on the other hand. However, establishing explicit norms and providing understandable self-purifying mechanisms such as moderators must take precedence to perform the proper function and contribute to creating a healthier and sound net culture.

## Ethics Statement

While our study used public data, we understood users' concerns about data use for research. We safeguarded identities by analyzing data at group levels. Our surveys and interviews were structured to respect participants' rights and dignity, receiving prior approval from our institution's Institutional Review Board (IRB). We obtained informed consent from participants and acquainted them with their rights. We also carefully avoided intrusive questions. Most of all, from a research-oriented perspective, our research underscores the multifaceted motivations and behaviors of online community users. By highlighting the synergetic relationship between activities in contrasting communities, we demonstrate the potential pitfalls in drawing conclusions from user engagement in a single community. Ethical research should avoid stigmatizing users based on their specific activities, considering the broader context of their online interactions.

## Acknowledgements

We express our gratitude to the anonymous reviewers for their insightful feedback. This research was supported in part by The Center for Social Data Analytics (C-SoDA) Accelerator Award Program at Penn State in 2021.

## References

Allison, K. R.; Bussey, K.; and Sweller, N. 2019. 'I'm going to hell for laughing at this' Norms, Humour, and the Neutralisation of Aggression in Online Communities. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–25.

Ancis, J. R. 2020. The age of cyberpsychology: An overview. *Technology, Mind, and Behavior*.

Attrill-Smith, A.; Fullwood, C.; Keep, M.; and Kuss, D. J. 2019. *The Oxford handbook of cyberpsychology*. Oxford University Press.

Bandura, A. 2014. Social cognitive theory of moral thought and action. In *Handbook of moral behavior and development*, 69–128. Psychology press.

Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The Development and Psychometric Properties of LIWC-22.

Butler, B. S.; and Wang, X. 2012. The cross-purposes of cross-posting: Boundary reshaping behavior in online discussion communities. *Information Systems Research*, 23(3-part-2): 993–1010.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357.

Christopherson, K. M. 2007. The positive and negative implications of anonymity in Internet social interactions:"On the Internet, Nobody Knows You're a Dog". *Computers in Human Behavior*, 23(6): 3038–3056.

Dumont, G.; and Candler, G. 2005. Virtual jungles: survival, accountability, and governance in online communities. *The American Review of Public Administration*, 35(3): 287–299.

Dynel, M.; and Poppi, F. I. 2020. Quid rides?: Targets and referents of RoastMe insults. *Humor*, 33(4): 535–562.

Fullwood, C.; James, B. M.; and Chen-Wilson, C.-H. 2016. Self-concept clarity and online self-presentation in adolescents. *Cyberpsychology, Behavior, and Social Networking*, 19(12): 716–720.

Halligan, S.; Altman, D. G.; and Mallett, S. 2015. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *European radiology*, 25(4): 932–939.

Hollenbaugh, E. E.; and Everett, M. K. 2013. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18(3): 283–302.

Hwang, S.; and Foote, J. D. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Joinson, A. N. 2007. Disinhibition and the Internet. In *Psychology and the Internet*, 75–92. Elsevier.

Jung, S.-G.; An, J.; Kwak, H.; Salminen, J.; and Jansen, B. J. 2018. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Twelfth international AAAI conference on web and social media*.

Kabay, M. E. 1998. Anonymity and pseudonymity in cyberspace: deindividuation, incivility and lawlessness versus freedom and privacy. In *Annual Conference of the European Institute for Computer Anti-virus Research (EICAR), Munich, Germany*, 16–8. Citeseer.

Kasunic, A.; and Kaufman, G. 2018. " At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Lapidot-Lefler, N.; and Barak, A. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2): 434–443.

Lapidot-Lefler, N.; and Barak, A. 2015. The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 9(2).

Lin, Y. 2021. 10 reddit statistics you should know in 2021 [infographic].

Marwick, A. E. 2013. Online identity. *A companion to new media dynamics*, 355–364.

McGraw, A. P.; and Warren, C. 2010. Benign violations: Making immoral behavior funny. *Psychological science*, 21(8): 1141–1149.

Pennebaker, J. W.; Booth, R. J.; Boyd, R.; and Francis, M. E. 2015. Linguistic Inquiry and Word Count: LIWC 2015 (Pennebaker Conglomerates, Austin, TX).

Shen, J.; Brdiczka, O.; and Liu, J. 2015. A study of Facebook behavior: What does it tell about your Neuroticism and Extraversion? *Computers in Human Behavior*, 45: 32–38.

Strimbu, N.; and O'Connell, M. 2019. The relationship between self-concept and online self-presentation in adults. *Cyberpsychology, Behavior, and Social Networking*, 22(12): 804–807.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior*, 7(3): 321–326.

Tan, C.; and Lee, L. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, 1056–1066.

TeBlunthuis, N.; Kiene, C.; Brown, I.; Levi, L.; McGinnis, N.; and Hill, B. M. 2022. No community can do everything: why people participate in similar online communities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–25.

Zhu, H.; Kraut, R. E.; and Kittur, A. 2014. The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 281–290.

## Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? No, because we could not identify any significant potential misuses of our work that warranted discussion in the manuscript.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, because we viewed our machine learning analysis section as supplementary, meant primarily to support other key sections, and therefore did not consider it essential.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? No. For the same reasons mentioned previously.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, because our machine learning experiment is relatively simple, requiring minimal computing resources, and we considered it unnecessary.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? No, because we did not consider it essential. However, if needed, we are prepared to provide the full text as a supplemental material.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes
   (d) Did you discuss how data is stored, shared, and deidentified? Yes