Unpacking the Layers: Exploring Self-Disclosure Norms, Engagement Dynamics, and Privacy Implications

Ehsan-Ul Haq
euhaq@hkust-gz.edu.cn
Hong Kong University of Science and Technology (GZ)

Shalini Jangra
s.jangra@surrey.ac.uk
University of Surrey

Suparna De
s.de@surrey.ac.uk
University of Surrey

Nishanth Sastry
n.sastry@surrey.ac.uk
University of Surrey

Gareth Tyson
gtyson@ust.hk
Hong Kong University of Science and Technology (GZ)

ABSTRACT

This paper characterizes the self-disclosure behavior of Reddit users across 11 different types of self-disclosure. We find that at least half of the users share some type of disclosure in at least 10% of their posts, with half of these posts having more than one type of disclosure. We show that different types of self-disclosure are likely to receive varying levels of engagement. For instance, a *Sexual Orientation* disclosure garners more comments than other self-disclosures. We also explore confounding factors that affect future self-disclosure. We show that users who receive interactions from (self-disclosure) specific subreddit members are more likely to disclose in the future. We also show that privacy risks due to self-disclosure extend beyond Reddit users themselves to include their close contacts, such as family and friends, as their information is also revealed. We develop a browser plugin for end-users to flag self-disclosure in their content.

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **Information systems** → *Internet communications tools.*

KEYWORDS

Self-disclosure, Privacy, Reddit, Engagement

ACM Reference Format:
Ehsan-Ul Haq, Shalini Jangra, Suparna De, Nishanth Sastry, and Gareth Tyson. 2025. Unpacking the Layers: Exploring Self-Disclosure Norms, Engagement Dynamics, and Privacy Implications. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3701716.3717740

WWW Companion '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3717740

1 INTRODUCTION

Online self-disclosure is sharing information about oneself with other users in online communities [27, 46]. This may include information on health, gender, sexual orientation, or general opinions, *e.g.* liking or disliking a person [30]. Reasons for disclosure are diverse, including the perceived sense of anonymity [34], adherence to community norms [8], and to improve engagement with community members [7]. There are many benefits of self-disclosure (such as support seeking [9]). However, it can also raise privacy concerns. For example, self-disclosure can lead to serious risks related to mental health [39, 44, 48] and intimacy [29]. Indeed, there have been numerous cases where accidental self-disclosure has led to physical harms, *e.g.* loss of employment [40], harassment [15], and geo-tagging [24]. Consequently, we argue that it is vital to better understand how such privacy-invasive disclosures occur, and develop tools to mitigate such risks.

Although there have been prior works that study self-disclosure online, these are either specific to a particular type of self-disclosure (*e.g.* gender [38]) or specific to a particular type of community, *e.g.* support seeking subreddits [8] and mental health forums [9]. Hence, we know little about the true scale of self-disclosure across diverse disclosure categories and communities. Key questions include: (*i*) How often do users self-disclose, and what types of self-disclosure are made together? (*ii*) How common are high-risk self-disclosures? (*iii*) What is the effect of self-disclosure specific communities on other users' self-disclosure propensity? Answering such questions is critical for formulating better user support against associated privacy risks. Yet, the subtle complexities of online self-disclosure raise a number of challenges that must be overcome to study this. Specifically, to date, we lack a methodical way of identifying and classifying forms of self-disclosure in online posts. This is further complicated by the fact that self-disclosures are not always atomic. For instance, a user may disclose more than one type of identifiable information within one post or across multiple (seemingly unrelated) posts. This is particularly common on platforms like Reddit, where users may participate in multiple subreddits [11]. Take the following posts (paraphrased to retain anonymity) as an example: *"I am a 20 years-old male who does not have a good relationship with my parents (mother is 59 and father is 60) and I am suffering from mental health issues. I have an appointment at ER tomorrow, but I have not told my parents."* The same user in an earlier post wrote – *"I am 19 year old male, a 19 year old female friend of mine is*

interested in dating me because I am less masculine as compared to her previous boyfriends." Here, by combining these posts, a third party can study the user's age, gender, and health, alongside garnering information about the user's parents and potential partner, along with the reason why she wanted to date the user.

With the above in-mind, this paper characterizes the multifaceted nature of self-disclosure on Reddit, across 11 distinct categories related to identity and sensitive information [14]. These categories encompass age, gender, religion, ethnicity, sexual orientation, and more. To study how these patterns vary across communities, we study a diverse pool of Reddit users from the top-10 (by number of users) subreddits. We then use the outcomes to design a tool that can alert users to (potentially unknown) disclosures in their social media posts. Our contributions are:

- We design a novel classifier to detect the presence of 11 different types of self-disclosures in a piece of text. We make our classifier open-source to assist other researchers. (§4)
- We show that at least 50% of users self-disclose in at least 10% of their total posts. Moreover, 50% of disclosing posts have more than one type of self-disclosure, revealing prior works miss significant volumes of information [35, 50]. We build on this to identify the social norms of self-disclosure by highlighting pairs more likely to co-occur in a post, such as *Age* and *Gender*, or *Gender* and *Relationship*. (§5)
- We show that users' engagement varies significantly with the type of self-disclosure in the posts. For instance, *Sexual Orientation* gets 2.6x more comments than posts without any disclosure. Moreover, users who have received interactions from self-disclosure-specific community members (such as the LGBT subreddit) are more likely to disclose in the future than those who have not. (§6)
- We embed our work in a browser tool that can automatically alert users to inadvertent online self-disclosure. (§7)

2 BACKGROUND

Self-Disclosure Types and Detection. Most of the prior literature on self-disclosure focuses on one type of self-disclosure or focused user groups and communities. For example, disclosure about mental health [3, 9] and subreddits related to health support [3], empathy and intimacy [41]. Some studies have looked at more than one type of self-disclosure. However, such studies are based on the discourse related to specific events such as the COVID-19 pandemic [6]. Qualitative methods such as manual coding [8], surveys, and interviews [50] have been used to detect self-disclosure. Other researchers have used quantitative methods, consisting of supervised learning [3, 46], unsupervised learning [6], and large language models (LLMs) [14] for self-disclosure detection. A key contrast between our work and the above is that we focus on identifying multiple types of self-disclosure.

Self-Disclosure Characterization Online self-disclosure characterization is a multidimensional area [5, 52]. Several studies focus on self-disclosure as means of users' support [9, 16, 31], other have explored linguistic characteristics within self-disclosure [16], and its impact on privacy [4, 13, 49]. One particular direction focuses on investigating social and communication norms among users, which can lead to nuances in self-disclosure behavior [12, 17]. For example, establishment of self-disclosure norms in mental health discussions as part of communication reinforces self-disclosure [12]. However, differences can be observed within demographics; for example, younger people are more open about their sexuality [17]. Users' gender may play an important role in their disclosure and reaction to the disclosure of others within the blogging community [26].

Our work is distinct in that it focuses on characterizing *multiple* types of self-disclosure of a *general* set of users, spanning various communities. Thus, our insights generalize to a wide population of Reddit users, and captures a general scale of disclosure.

Self-Disclosure and Privacy. Another key line of work focuses on building tools for end-users to help control their disclosure [14, 18]. A recent study proposed a task to help users rewrite the disclosure in their social media posts [14]. Similarly, Guarino et al. developed a web browser extension to help users control their disclosure based on keywords. However, it does not cover six of our identified disclosure types and relies on multiple classifiers, increasing the computational cost [18]. Our work provides data-driven insights and a tool to improve user privacy. Our tool covers 11 types of self-disclosure and does not require maintaining users' profiles. Additionally, our co-occurrence analysis offers insights into self-disclosures likely to occur together, which is useful in downstream research to increase the efficiency of privacy-preserving methods.

3 DATA COLLECTION METHODOLOGY

Our data collection aims to solicit *all* posts from a Reddit user in a given time window. This ensures a holistic view of self-disclosure by a given user. Thus, we use the pushshift data dump of all Reddit from October 2020 to June 2021 [23], making it possible to track a user's all interactions within this time period. We first gather a seed list of users to start our data collection. For this, we extract all users who have written a post in any of the top 10 largest subreddits[1] (by community size) during two months (Jan. and Feb. 2021). We call these *general users*. Note that posts collected from these general subreddits are more likely to contain general discourse than topic-specific discourse like health and intimacy [3, 9].

We then gather all their posts for these general users, from October 2020 to June 2021. The raw dataset contains 16,706,119 posts from 365,385 users. We remove accounts containing any single or combination of the words *'bot', 'moderat', or 'auto'* in their usernames to filter bot and moderator accounts [51]. This removed 1,103 accounts and 428,275 posts.

4 SELF-DISCLOSURE CLASSIFICATION

To analyze self-disclosure at scale, it is first necessary to devise a methodology that can automatically identify the presence of disclosure in user posts. Thus, we first design a classifier.

4.1 Defining Self-Disclosure Types

We start by defining a taxonomy of self-disclosure types to form our supervised label set. Given our focus on privacy, we focus on self-disclosures that may reveal one's identity and information that may be used in a potential prejudice, such as health, age, gender, and sexual orientation [31, 46]. A recent study proposed a list of

[1]The subreddits are *funny, AskReddit, gaming, aww, worldnews, todayilearned, Music, movies, science, pics*

19 self-disclosure types related to demographics and personal experiences [14]. We take inspiration from these 19 categories and a review of different self-disclosure types studied in prior literature [50]. Based on our analysis, in Table 2 (Appendix B), we present the high-level categories of self-disclosure. These cover various aspects of one's life, including identity, relationship, work, health, group affiliations, and opinions. For simplicity, we do not consider the group affiliations and opinion categories; these categories include far more subjective information. It is also challenging to judge if this subjective information carries sensitive insight.

Note that we consider relationship, profession/economics, and health as broader categories in this study. For instance, we consider health to be one type of self-disclosure, as our research questions do not strive to differentiate between mental and physical health at a granular scale. Similarly, finance, profession, and job-related self-disclosure are grouped as "Job." These considerations help us improve our classifier's performance and answer our research questions more effectively. Our final list of self-disclosure types is: **Age, Education, Ethnicity, Gender, Health, Job, Location, Physical Appearance, Relationship, Religion, and Sexual Orientation.** We acknowledge that our selected self-disclosure types do not form an exhaustive list. However, this list matches commonly studied self-disclosure types [50] and covers the GDPR definition of personal data, except affiliations, opinions, and genetic data [1].

4.2 Self-Disclosure Training Data Annotation

It is next necessary to label a training dataset, for which we take a two-step approach. To optimize the training sample, we first use ChatGPT to extract more relevant posts for manual inspection. Then, two researchers label the posts to create the training set.

Identification of Self-Disclosure. There are two commonly used approaches for self-disclosure identification: (*i*) A binary coding (*yes, no*) representing the presence or absence of self-disclosure [28]; and (*ii*) Identifying self-disclosure as an ordinal variable that quantifies the sensitivity of the information provided by a user [46]. Usually, this is in the form of *low, medium, and high*. There is, however, no fixed criteria for using either of the approaches. We follow the first approach and use a binary variable to indicate the presence or absence of a particular type of self-disclosure. This is because we are interested in the overall presence of self-disclosure instead of a fine-grained analysis within each self-disclosure type. We also note that an ordinal approach will require a more extensive data annotation exercise and will likely result in reduced performance.

Data Annotation. We use a two-step data annotation exercise, employing ChatGPT (3.5) followed by manual annotation. The use of ChatGPT is to improve the sample selection, as only the samples with positive outcomes from ChatGPT are subsequently manually analyzed. We randomly select 2000 samples from the dataset for ChatGPT annotation. We then use the standard ChatGPT guidelines on the chain of thought [47] along with a role assignment [25] to design our prompt. The GPT annotations are performed in a single session, and 1,764 posts are selected as containing at least one type of disclosure. Two researchers then perform manual validation of these annotations. After a briefing session, the two researchers are assigned a pool of 110 posts (taken from the positive ChatGPT annotations), consisting of 10 posts randomly selected for each of the 11 self-disclosure types. The results from the two researchers are compared and disagreements are discussed. Inter-rater reliability (IRR) is calculated, and the Cohen's Kappa score is in the range of 0.51 to 0.80, (median 0.70 and mean 0.70) for all labels. In addition, 19 posts are identified to be completely incorrectly classified as self-disclosing by ChatGPT, with another 30% posts missing at least one label identified in manual inspection (where both raters have agreements). After the disagreements are discussed between the researchers, another round of manual annotations is followed on a sample of 250 posts, followed by the IRR measurements for round 2. Here, the Cohen's Kappa score for each self-disclosure type is in the range of 0.79 to 0.94 (median 0.85, and mean 0.82); showing a high agreement between the two annotators [20]. After the second debriefing round, following prior work [36], one researcher continues with the rest of the manual annotations. In total, 1,329 posts are labeled as self-disclosing. One post can have more than one type of self-disclosure. We use this final manually annotated dataset for training the classifier.

4.3 Classifier Training

Fine-tuning of pre-trained large models has shown positive results in downstream research tasks like ours. Thus, we use the labeled data to fine-tune pre-trained BERT and Roberta models [33, 45]. We define this as a multi-label classification task — the input is a post, and the output is the probability of each type of self-disclosure it contains (if any). We stratify the labeled data into 70-15-15% for training, testing, and validation. Roberta performs better than BERT, with an average weighted F1 of 0.83. Table 4 (Appendix D) reports the F1 scores within each self-disclosure type, along with precision and recall scores. We achieve the best F1 scores for Ethnicity (0.95), Religion (0.94), and Age (0.89). All self-disclosure except *Gender* and *Job* have >0.8 F1 score. Note that we consider a positive class if the sigmoid function has a value of 0.5 or higher. Any change in this limit can affect the data size; however, 0.5 gives a moderate approach and is commonly used in classification tasks. We use this model to predict the self-disclosure labels of every post. We predict labels for the text and title of posts separately due to difference in the length and writing styles. For each post, we then take the union of the self-disclosure labels identified in both the text and title. Our classifier is open source for use by the community.[2]

5 QUANTIFYING SELF-DISCLOSURE

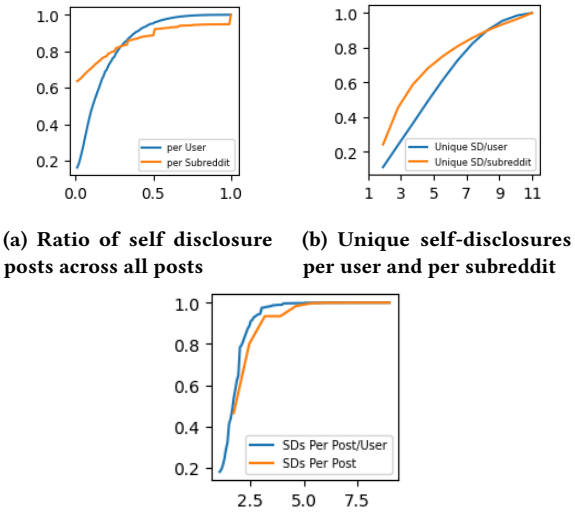
We now measure the scale of self-disclosures to understand what sort of (privacy-sensitive) information is disclosed.

5.1 Quantifying the Scale of Self-Disclosure

First, we evaluate the count self-disclosures across all users and posts, to quantify potential privacy-compromising exposure.

Number of Self-Disclosures Per User. For each user, we inspect the ratio of posts that contain at least one type of self-disclosure vs. the total number of posts from that user. Figure 1a shows the distribution of this ratio for all users in our dataset. The mean of the ratio is 0.15 (median 0.10, 75th percentile 0.23). Half of the users in our dataset self-disclose in at least 10% of their posts, and 25% of

[2]https://huggingface.co/euhaq/self_disclosure



(a) Ratio of self disclosure posts across all posts

(b) Unique self-disclosures per user and per subreddit

(c) Unique self-disclosure types per post and per user

Figure 1: Cumulative distributions of self-disclosure.

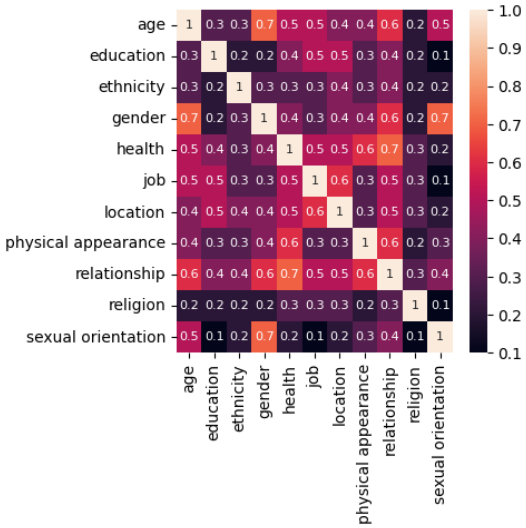


Figure 2: Correlation between different self-disclosure types that occur commonly per user. Most users who share Age also share Gender and Relationship-related self-disclosures.

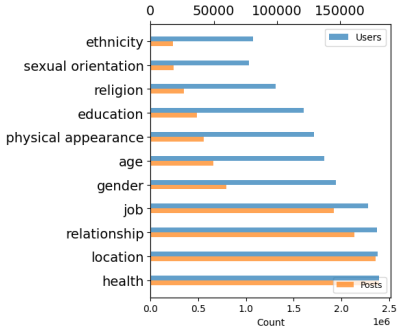


Figure 3: Number of users (top x-axis) and posts (bottom x-axis) for each self-disclosure type.

users have self-disclosure in at least 23% of their posts. This confirms that a significant number of users are involved in self-disclosure. We emphasize that our dataset is based on users from general purpose subreddits, and we do not focus on specific subreddits that are more likely to have self-disclosure. Thus, this is surprisingly high.

Number of Self-Disclosure Types Per User. Whereas the above may raise privacy concerns, it is less problematic if a user discloses the same information repeatedly (rather than many instances of new information). Thus, we inspect the number of *unique* self-disclosure types per user. For example, if a user shares *Age* in four posts and *Gender* in two posts, this user still only has two unique self-disclosure types (*Age* and *Gender*).

Figure 1b plots the number of unique self-disclosure types per user. We see that users share an average of four ($\sigma = 2$) different types of self-disclosure across their timelines. In fact, 50% of users have at least five self-disclosure types in their timelines. Thus, most users exhibit a diverse range of disclosures. In many cases, we find that these multiple disclosures are embedded within individual posts. To quantify this, Figure 1c (orange line) plots the distribution of the number of unique disclosures per post, and the blue line shows the distribution of the per-user average of this measure. We find that, on average, posts contain close to 2 ($\mu = 1.8$, $\sigma = 2$) self-disclosure types. This confirms that users expose substantial and diverse information. We posit that it is possible for third parties to easily combine such information. This naturally increases users' susceptibility to malicious actors in online spaces.

5.2 Quantifying Types of Self-Disclosure

The above has identified that a large number of self-disclosures are exposed by users. We proceed to study the specific categories of information disclosed and which are co-located within posts.

Frequency of Self-Disclosure Types. Figure 3 presents a histogram of the number of posts and users who disclose each type of information. Health, location, and relationship are the most commonly shared. Location is particularly concerning, as this can expose users to physical risks. Equally, health information tends to be particularly sensitive. That said, support-seeking for users with medical issues is commonplace, and some may argue the benefits outweigh the risks in certain scenarios [43, 53]. Information about ethnicity, sexual orientation, and religion is the least shared. these disclosures can also be used to harm users, *e.g.* sharing such information can be used for catfishing and cyberbullying [15].

Co-occurrence of Self-Disclosures. Next, we inspect which combinations of disclosure types are exposed by each user. We wish to understand if certain combinations of information disclosure are common (*e.g.* revealing *both* age and gender). For this, we first compute the sum of all types of self-disclosure shared by each user. This generates a matrix of $nx11$, where n is the total number of users (recall, we cover 11 types of self-disclosure). We then calculate the Pearson correlation for each pair of self-disclosure types. Figure 2

Table 1: The likelihood of self-disclosures to co-occur in the same post. Each column and row represent model and confounding factors, respectively. $***$ indicates statistical significance. The top value displays the β estimate and the bottom shows the standard error. Blue and Red colors indicate the most positive and negative estimates, respectively.

	Age	Education	Ethnicity	Gender	Health	Job	Location	Phy. Appearance	Relationship	Religion	Sexual Orientation
age		0.013***	-0.005***	0.271***	0.012***	0.006***	0.002*	-0.020***	0.119***	-0.008***	0.002***
		0.001	0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0
education	0.017***		-0.014***	-0.053***	-0.168***	-0.136***	-0.112***	-0.057***	-0.103***	-0.045***	-0.023***
	0.001		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
ethnicity	-0.019***	-0.038***		0.069***	-0.166***	-0.198***	0.009***	-0.005***	-0.087***	-0.029***	-0.003**
	0.002	0.002		0.001	0.002	0.002	0.002	0.001	0.002	0.001	0.001
gender	0.364***	-0.056***	0.025***		-0.134***	-0.141***	-0.069***	0.060***	0.141***	-0.023***	0.067***
	0.001	0.001	0.001		0.001	0.001	0.001	0.001	0.001	0.001	0.001
health	0.007***	-0.080***	-0.028***	-0.061***		-0.307***	-0.208***	-0.002***	-0.147***	-0.063***	-0.036***
	0.001	0.001	0	0.001		0.001	0.001	0.001	0.001	0	0
job	0.003***	-0.063***	-0.032***	-0.062***	-0.294***		-0.188***	-0.075***	-0.198***	-0.070***	-0.037***
	0.001	0.001	0	0.001	0.001		0.001	0.001	0.001	0	0
location	0.001*	-0.058***	0.002***	-0.034***	-0.223***	-0.211***		-0.044***	-0.129***	-0.043***	-0.018***
	0.001	0.001	0	0.001	0.001	0.001		0.001	0.001	0	0
physical_appearance	-0.034***	-0.074***	-0.002***	0.074***	-0.005***	-0.211***	-0.110***		-0.105***	-0.042***	-0.008***
	0.001	0.001	0.001	0.001	0.001	0.001	0.001		0.001	0.001	0.001
relationship	0.069***	-0.047***	-0.014***	0.062***	-0.140***	-0.197***	-0.115***	-0.037***		-0.027***	0.002***
	0.001	0.001	0	0.001	0.001	0.001	0.001	0.001		0	0
religion	-0.024***	-0.107***	-0.024***	-0.051***	-0.307***	-0.357***	-0.198***	-0.076***	-0.138***		-0.032***
	0.002	0.001	0.001	0.001	0.002	0.002	0.002	0.001	0.002		0.001
sexual_orientation	0.008***	-0.065***	-0.003**	0.179***	-0.210***	-0.227***	-0.095***	-0.017***	0.012***	-0.038***	
	0.002	0.001	0.001	0.001	0.002	0.002	0.002	0.001	0.002	0.001	

shows the correlations on a per-user level. Each cell in the figure shows the correlation between a pair of self-disclosure types — a higher value indicates that the self-disclosure types from the pair have been shared together by more users. We do see certain pairs commonly occurring, suggesting that certain self-disclosure norms have emerged within Reddit. Importantly, we emphasize that this disclosure behavior is not limited to support-seeking subreddits, but is instead a *general* self-disclosure norm.

Modeling Self-Disclosure Relationships. To systematize the above analysis, we model the precise relationships using 11 different fixed effect regression models, as follows:

$$s_j = \sum_{i}^{11} \beta_i s_i + U + T \quad i, j \in [1, 11], i \neq j \tag{1}$$

where, s_j is the dependent variable for the j^{th} self-disclosure, and B_i is the estimate (impact) of the remaining self-disclosure types. U and T are the fixed effects for users and time. We develop 11 regression models, one for each self-disclosure type, with the goal of identifying potential patterns of risky co-disclosure.

Table 1 summarizes the results, where each column reports a regression model for a self-disclosure type mentioned in the column header. The rows of the table report the regression estimate of the remaining self-disclosure types. Additionally, each estimate is accompanied by the standard error along with p-values indicated by $*$. For each model, we highlight the most positive estimate with blue color and the most negative one with red color.

Confirming our previous results in Figure 2, we observe both positive and negative correlations in self-disclosures, reflecting interesting trends. For instance, if a post adds a self-disclosure about *Gender* the likelihood of *Age* disclosure increases by 0.364x. However, disclosure about *Ethnicity* will increase the likelihood of disclosing *Age* by just 0.02x. Similarly, mentioning *Gender* increases the likelihood of mentioning a *Relationship* by 0.14x, while *Job* reduces the likelihood of mentioning *Relationship* by -.198x. Overall, we find that *Age* and *Gender* have the largest effect sizes for most of the regression models. This implies that users most often disclose their age and gender together on Reddit. Such combinations may have benefits for support but also pose risks associated with mental health to certain demographics*e.g.* youth during gender transitions [21]. This combination is also visible in Figure 2, which shows a 0.74 correlation between the presence of *Age* and *Gender* disclosure. Although this is arguably privacy-sensitive, it does reveal a common norm, whereby users are expected to express such information (*e.g.* "[30 F]") as part of daily discussion [8].

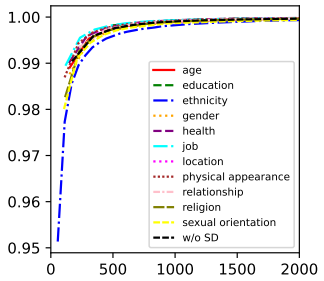
At the opposite end of the spectrum, *Religion* and *Job* generally have the least effect size for most regression models. This suggests these are least likely to co-occur with other self-disclosure types in the same post. For instance, the likelihood of *Health* disclosure is reduced by 0.357x with *Religion*. *Religion* and *Job* generally have a negative impact on most of self-disclosure types to co-occur with them. For instance, the likelihood of *Health* disclosure is reduced by 0.357x with *Religion*. The positive correlation of *Gender* with *Physical Appearance* and *Sexual Orientation* shows that the latter two are mostly accompanied by the former. We argue that topics related to physical appearance, such as body shaming together with gender, may lead towards negative outcomes such as body shaming and psychological abuse [10, 37], or hate speech concerning the sexuality of the users [11, 32].

6 SELF-DISCLOSURE ENGAGEMENT

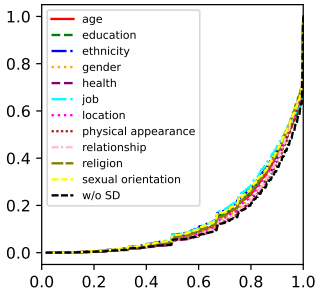
Next, we explore user engagement with self-disclosing posts, focusing on how interactions in self-disclosure communities (*e.g.*, /r/AMA) may encourage others to disclose, potentially creating privacy-compromising attack vectors.

6.1 Quantifying Engagement with Self-Disclosure

From a privacy perspective, higher engagement reflects potentially increased exposure to users' information and may also affect users' future posts [22]. We therefore start by measuring the difference in engagement levels across self-disclosing posts.

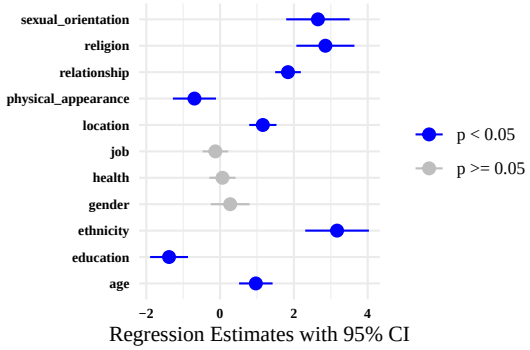


(a) Number of Comments

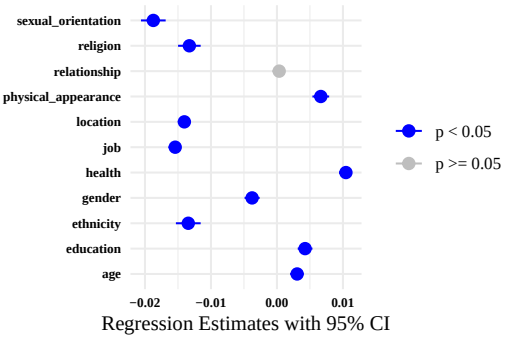


(b) Upvote Ratio

Figure 4: Cumulative distributions of per-user mean engagement values, per disclosure type.



(a) Effect on Number of Comments



(b) Effect on Upvote Ratio

Figure 5: Regression Models for Engagement.

Figure 4 presents the distribution of two engagement metrics: (*i*) number of comments and (*ii*) the ratio of upvotes to downvotes, across all posts containing self-disclosure. A non-parametric Kruskal-Wallis test confirms significant differences across the distribution of each disclosure type. For instance, posts with *Sexual Orientation* have more comments ($\mu = 16.4$) than posts with *Job* disclosure ($\mu = 11.6$).

To systematically analyze these differences, we turn to regression analysis with users and time-fixed effects. For each engagement metric (number of comments and upvote ratio) and self-disclosure type, we design a separate engagement prediction task based on whether the posts contain a particular self-disclosure or not (not including other types of self-disclosure). The posts without self-disclosure from all users remain identical across the models, hence providing a common baseline across models to compare the results with non-disclosing posts and within different types.

Figure 5a and 5b show the regression results for number of comments and upvote ratio, respectively. The Y-axis shows the corresponding self-disclosure used as a confounding factor. The X-axis shows the regression estimate with a 95% confidence interval. The values in the blue color are statistically significant ($p < 0.05$), and the grey color shows non-significance.

We see that the presence of self-disclosure does have a significant effect on both the number of comments and the upvote ratio. Interestingly, the effect is not similar across all self-disclosure types, and there is also a disparity in each engagement metric for a given self-disclosure. For instance, the presence of *Sexual Orientation* disclosure increases the number of comments (2.65x); however, the same has a negative effect on the upvote ratio($\approx -.02x$).

To further see the difference between heterosexual and potentially more vulnerable non-hetero, we do a keyword filtering of the posts (with sexual orientation disclosure) containing the words (*gay, lesbian, bisexual, and straight*), and use a non-parametric Kruskal-Wallis test ($\chi^2 = 251.2, df = 3, p < 0.001$), followed by Dunn's test, to see the engagement metric distributions difference across each keyword's post. We observe that posts containing words 'gay' receive more comments (almost double) ($\mu = 102$) and lower upvote ($\mu = .80$) than other non-hetero (*e.g.* upvote ratio for 'lesbian' $\mu = 0.83$). However, it receives fewer comments than the posts containing 'straight' keywords ($\mu = 119$). This contrast shows that, although *Sexual Orientation* increases engagement in terms of comment count, the engagement quality is less positive, on average. We posit that such engagement metrics may influence users' future self-disclosure likelihood, as prior Reddit studies find engagement do affect topic choice [22]. Moreover, a lower upvote ratio with higher comments may signify negative discussions and discontent with the original poster [42], potentially leading to negative experiences, especially in disclosures like *Sexual Orientation* and *Ethnicity*.

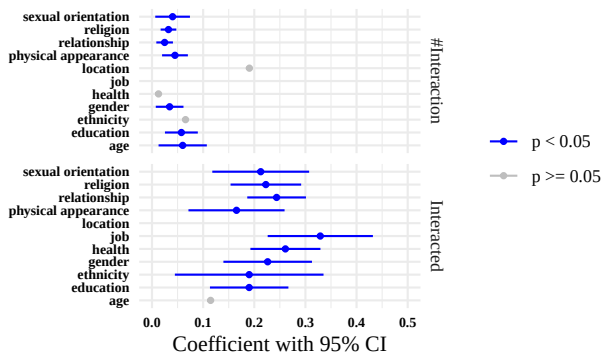


Figure 6: Self-disclosure types on the y-axis represent one regression model with two independent variables (#interactions and Interacted). The x-axis shows the β estimates.

6.2 Quantifying Impact of Self-Disclosure Communities

There are various Reddit communities directly related to specific forms of self-disclosure, *e.g.* `r/aznidentity` is related to discussions on ethnicity, `AskDocs` has health-related discussions, and `r/AskMen` has gender-related discussions. These pertain to *Ethnicity*, *Health*, and *Gender*, respectively. We hypothesize that engaging with members of such communities may increase one's own likelihood of disclosing, even to other communities. This may create attack surface where malicious actors purposefully post (fake) self-disclosure to encourage others to share. We next explore the potential presence of such behaviors, quantifying the impact of receiving comments that contain self-disclosure.

Self-Disclosure related Subreddits. For the above analysis, we first obtain a set of subreddits dedicated to self-disclosure. The names of the subreddits indicate their association with a focused topic [2]. We extract the top 50 largest subreddits, in terms of their number of posts classified as containing each type of self disclosures. As some subreddits are associated with multiple self-disclosure types, we end up with 250 unique subreddits out of 550 extracted subreddits. We then manually review the name of each subreddit and annotate whether it is associated with one of our self-disclosure types. Some examples are shown in Table 3 in Appendix C.

Regression Task. We next ask (*i*) what is the effect of a user receiving an interaction from a users in a self-disclosure related subreddit compared to not having received an interaction? and (*ii*) If a user receives an interaction, what is the impact on the number of such interactions? We model this as a regression task to predict whether the user will have a self-disclosure in future:

$$Y_{st} = \alpha + \beta N I_{t-1} + U + T$$

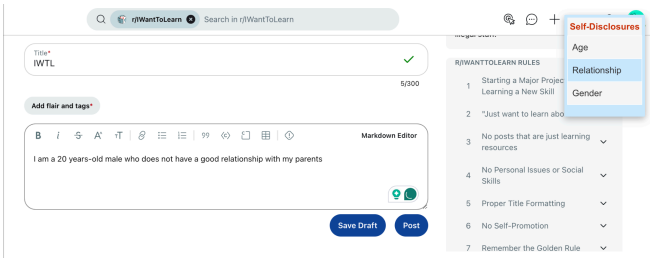
Y_{st} is the number of self-disclosures by a user in a period t (1 week). β shows the coefficient for number of interactions(N_{t-1}) at time $t-1$ with the users from selected subreddit communities (SD-specific or general subreddits). $I = [0, 1]$ whether a user received an interaction $I = 1$ or not $I = 0$. Note, we consider interaction to be any action initiated by users from a selected community. Thus, interaction occurs when a user from a selected subreddit writes a comment (at any level of the post) on a post by our selected users. Finally, U represent all users, and T represents time (in weekly brackets) to control any user and time-dependent fixed-effect. In total, we run 11 regression models. Each model is specific to the self-disclosure-specific subreddits labeled with that type of disclosure. Note, we consider any self-disclosure as a positive instance of self-disclosure and do not differentiate within different types.

Results. Following these steps, we run our fixed effect regression model while controlling heteroskedasticity [19]. Figure 6 plots the regression model results. The figure consists of two panels, one for each of the independent variables. The lower panel shows the binary impact of the user receiving an interaction or not, whereas the upper panel shows the impact of the number of interactions. The x-axis shows the β estimates for variables, and the y-axis refers to the regression model depending on the self-disclosure for which the specific subreddits are used. The blue color shows the results that are statistically significant (with $p - values$ being lower than 0.05). The gray color shows the corresponding estimates are statistically not significant. The missing values are also statistically not significant and the values are less than zero, so they are not shown due to the figure's x-axis scale.

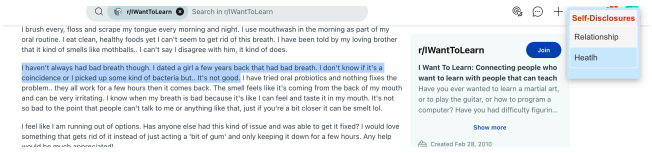
Confirming our hypothesis, we observe a positive effect (β estimates) for future self-disclosure, *i.e.* users are more likely to share a self-disclosing posts if they receive an interaction from a user who has previously posted in self-disclosure-specific subreddits. Distinct effect sizes for the *Interacted* variable are observed within each model. The most significant effect is observed for *Job*, with the odds of future self-disclosure being 0.34x, followed by *Health* at 0.26x, and *Relationship* at 0.24x. Similarly, the *#interactions* shows the positive effect for each such interaction. Given these confounding factors, we hypothesize two scenarios where Reddit users could be exploited: (*i*) A user may be influenced to self-disclose through repeated interactions with other users who engage in self-disclosure, potentially including bots; or (*ii*) The creation of subreddits designed to *maliciously* foster a sense of community to increase the likelihood of users engaging in self-disclosure.

7 INSIGHTWATCHER - A BROWSER PLUGIN

To help users control their self-disclosure, we have developed a browser plugin, *InsightWatcher*. The tool automatically scans for self-disclosure, in real time, within any text box that the user loads in their browser. It works across any webpage that contains a text box (including Twitter/X, Facebook, and WhatsApp Web). Whenever a self-disclosure is identified, a small non-invasive popup is raised, notifying the user of the information they are exposing if they proceed. The plugin achieves this using our classifier. It does not require any user login and does not record any inputs. Note, the tool also allows users to select any text on a web page, and request a list of self-disclosures within the text. Figure 7a shows a screenshot of the active reminder, featuring a pop-up on the right side of the screen in real-time. In Figure 7b, a user selects text from a webpage,



(a) Active Reminder



(b) Passive Reminder

Figure 7: Working Example of Browser Plugin

and a pop-up displays the self-disclosures from it. The plugin will be open source and available for users install.[3]

8 DISCUSSION

Our work presents the first large-scale and multi self-disclosure characterization on Reddit based on general discourse.

Privacy Control Measures. Our work highlights that a large number of posts have self-disclosure; 50% of users in our dataset have self-disclosure in at least 10% of their posts. In addition, half of the users disclose more than one type of self-disclosure. In addition to the empirical insights, there are other latent risks that lie in multiple disclosures from a user. For example, when users self-disclose, they often include information from their past and other people involved around that time. Thus, this can increase users' vulnerability to malactors who can combine such information to find more about a user. This behavior increases the need for tools, like our browser plugin InsightWatcher, that help users control their disclosure. However, Self-disclosure moderation should be context-specific, guided by ethical principles. In support-based communities, self-disclosure can enhance support, so strict moderation might lessen their positive impact. In contrast, disclosure in generic communities can be more harmful to users, thus requiring a nudge on self-disclosure.

Co-located Self-Disclosure Types. Our work highlights the importance of studying multiple self-disclosures together. Users do this to add more information and context while describing their life events or sharing their thoughts. Our work highlights that pairs of self-disclosure types, such as *Gender* and *Relationship*, are likely to appear together. This has implications when studying such disclosures alone. The occurrence of one particular disclosure may be inspired by another, hence deviating from commonly associated characteristics with certain disclosures. The study of co-located self-disclosure can offer insights into several use-cases, such as uncovering the mechanism of identity establishment, particularly for vulnerable populations. For instance, studying *Sexual Orientation* together with *Age* and *Gender* and *Relationship* can uncover how different groups open up about their sexuality.

Disclosure about Close Contacts. We observe that users often disclose about people around themselves, including immediate family members, friends, and co-workers. We find that over 88,709 posts mention at least one word from— *father, mother, sister, brother, boyfriend, girlfriend, bf, gf* in their main body. Thus, privacy risks can extend from one user to their extended social network. Reddit users often put such details and other characters to add more contextual information to refer to some event while sharing their experiences or seeking information. Thus, any tools designed must be flexible enough to accommodate these norms while minimizing privacy risk. We conjecture that dynamically rewriting such information, while not compromising the context could be valuable (*e.g.* changing the specific age).

9 CONCLUSION

We have presented the first multiple-type self-disclosure characterization on Reddit. We identified 11 types of self-disclosure associated with a user's identity and demographics, such as age, gender, relationship, and job. We first developed our open-source multi-label self-disclosure classifier for the 11 different self-disclosure types. Our characterization shows that at least half of users self-disclose in more than 10% of their posts. We highlight that user posts are not limited to one particular type of self-disclosure. Instead, users share more than one type of information to add more context to posts. Through thematic analysis, we show that user self-disclosure reveals information about themselves and extends to their social connections, such as parents, siblings, and partners. Building on this, we have developed and deployed a browser plugin tool that can automatically notify users when they are self-disclosing.

We note our study has several limitations, which form the basis of our future work. Most notably, our list of self-disclosure types is not exhaustive, albeit covering key identifiable information. Future work can extend disclosure types, such as disclosure through opinions (the types we have not included in our analysis) or increasing granularity as high, medium, and low. Moreover, our work is limited to Reddit. However, a similar approach can be extended to other platforms; for example, self-disclosure-related Facebook groups or hashtag-specific discussions can be taken as a proxy of Reddit communities for Facebook and X, respectively. Furthermore, our work is focused only on Reddit; similar studies are required on other platforms to analyze the generalization of our findings. Finally, within our work, we consider the presence of self-disclosure but do not quantify whether the self-disclosure is high or low risk. Our future work will focus on better understanding the exact nature of the risks involved and how they can be mitigated.

10 ACKNOWLEDGMENTS

This work was supported in part by the Guangzhou Science and Technology Bureau (2024A03J0684); the Guangzhou Municipal Science and Technology Project (2023A03J0011); the Guangzhou Municipal Key Laboratory on Future Networked Systems (024A03J0623), the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007); and by AP4L (EP/W032473/1).

[3]https://github.com/ehsanulhaq1/InsightWatcher

REFERENCES

[1] 2023. What is personal data? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/personal-information-what-is-it/what-is-personal-data/what-is-personal-data/ Publisher: ICO.

[2] David Ifeoluwa Adelani, Ryota Kobayashi, Ingmar Weber, and Przemyslaw A. Grabowicz. 2020. Estimating community feedback effect on topic choice in social media with predictive modeling. *EPJ Data Science* 9, 1 (Dec. 2020), 1–23. https://doi.org/10.1140/epjds/s13688-020-00243-w

[3] Sairam Balani and Munmun De Choudhury. 2015. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Seoul Republic of Korea. https://doi.org/10.1145/2702613.2732733

[4] Natalya N. Bazarova and Yoon Hyung Choi. 2014. Self-Disclosure in Social Media: Extending the Functional Approach to Disclosure Motivations and Characteristics on Social Network Sites1. *Journal of Communication* 64, 4 (Aug. 2014), 635–657.

[5] Thales Bertaglia, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi. 2024. Influencer Self-Disclosure Practices on Instagram: A Multi-Country Longitudinal Study. http://arxiv.org/abs/2407.09202 arXiv:2407.09202.

[6] Taylor Blose, Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2020. Privacy in Crisis: A study of self-disclosure during the Coronavirus pandemic. https://doi.org/10.48550/arXiv.2004.09717 arXiv:2004.09717 [cs].

[7] Hsuan-Ting Chen. 2018. Revisiting the Privacy Paradox on Social Media With an Extended Privacy Calculus Model: The Effect of Privacy Concerns, Privacy Self-Efficacy, and Social Capital on Privacy Management. *American Behavioral Scientist* 62, 10 (Sept. 2018), 1392–1412. https://doi.org/10.1177/0002764218792691

[8] Yixin Chen, Scott Hale, and Bernie Hogan. 2024. "I Am 30F and Need Advice!": A Mixed-Method Analysis of the Effects of Advice-Seekers' Self-Disclosure on Received Replies. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 276–288. https://doi.org/10.1609/icwsm.v18i1.31313

[9] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 71–80. https://doi.org/10.1609/icwsm.v8i1.14526 Number: 1.

[10] Enrico Corradini. 2023. The Dark Threads That Weave the Web of Shame: A Network Science-Inspired Analysis of Body Shaming on Reddit. *Information* 14, 8 (Aug. 2023), 436.

[11] John Patrick Crowley. 2014. Expressive Writing to Cope with Hate Speech: Assessing Psychobiological Stress Recovery and Forgiveness Promotion for Lesbian, Gay, Bisexual, or Queer Victims of Hate Speech. *Human Communication Research* 40, 2 (April 2014), 238–261. https://doi.org/10.1111/hcre.12020

[12] Beth Dietz-Uhler, Cathy Bishop-Clark, and Elizabeth Howard. 2005. Formation of and Adherence to a Self-Disclosure Norm in an Online Chat. *CyberPsychology & Behavior* 8, 2 (April 2005), 114–120. https://doi.org/10.1089/cpb.2005.8.114 Publisher: Mary Ann Liebert, Inc., publishers.

[13] Tamara Dinev and Paul Hart. 2006. Privacy Concerns and Levels of Information Exchange: An Empirical Investigation of Intended e-Services Use. *e-Service Journal* 4, 3 (2006), 25–60. https://doi.org/10.2979/esj.2006.4.3.25

[14] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. https://doi.org/10.48550/arXiv.2311.09538 arXiv:2311.09538 [cs].

[15] Lauckner et. al. 2019. "Catfishing," cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males. *Journal of Gay & Lesbian Mental Health* 23, 3 (2019), 289–306. https://doi.org/10.1080/19359705.2019.1587729

[16] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "Let Me Tell You About Your Mental Health!": Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. 753–762.

[17] Connor Gilroy and Ridhi Kashyap. 2021. Digital Traces of Sexualities: Understanding the Salience of Sexual Identity through Disclosure on Social Media. *Socius* 7 (Jan. 2021), 23780231211029499. https://doi.org/10.1177/23780231211029499

[18] Alfonso Guarino, Delfina Malandrino, and Rocco Zaccagnino. 2022. An automatic mechanism to provide privacy awareness and control over unwittingly dissemination of online private information. *Computer Networks* 202 (Jan. 2022), 108614. https://doi.org/10.1016/j.comnet.2021.108614

[19] Damodar N. Gujarati and Dawn C. Porter. 2009. *Basic econometrics* (5. ed ed.). McGraw-Hill Irwin, Boston, Mass.

[20] Reza Hadi Mogavi, Yankun Zhao, Ehsan Ul Haq, Pan Hui, and Xiaojuan Ma. 2021. Student Barriers to Active Learning in Synchronous Online Classes: Characterization, Reflections, and Suggestions. In *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21)*. 101–115. https://doi.org/10.1145/3430895.3460126

[21] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2015. Disclosure, Stress, and Support During Gender Transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 1176–1190. https://doi.org/10.1145/2675133.2675152

[22] Ehsan-Ul Haq, Tristan Braud, Lik-Hang Lee, Anish K. Vallapuram, Yue Yu, Gareth Tyson, and Pan Hui. 2022. Short, Colorful, and Irreverent! A Comparative Analysis of New Users on WallstreetBets During the Gamestop Short-squeeze. In *Companion Proceedings of the Web Conference 2022 (WWW '22)*. Association for Computing Machinery, 52–61. https://doi.org/10.1145/3487553.3524202

[23] Ehsan-Ul Haq, Yiming Zhu, Zijun Lin, Haodi Weng, Gareth Tyson, Lik-Hang Lee, Reza Hadi Mogavi, Tristan Braud, and Pan Hui. 2023. Understanding Characteristics of Catalyst Users in the WallStreetBets Community. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. ACM, Kusadasi Turkiye, 320–324. https://doi.org/10.1145/3625007.3627595

[24] Keith Harrigian. 2018. Geocoding Without Geotags: A Text-based Approach for reddit. http://arxiv.org/abs/1810.03067

[25] Muhammad Imran and Norah Almusharraf. 2023. Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology* 15, 4 (Oct. 2023), ep464. https://doi.org/10.30935/cedtech/13605 Publisher: Bastas.

[26] Chyng-Yang Jang and Michael A. Stefanone. 2011. Non-Directed Self-Disclosure in the Blogosphere: Exploring the persistence of interpersonal communication norms. *Information, Communication & Society* 14, 7 (Oct. 2011), 1039–1059.

[27] Adam Joinson. 2007. *Oxford handbook of internet psychology*.

[28] Yubo Kou and Colin M. Gray. 2018. "What do you recommend a complete beginner like me to practice?": Professional Self-Disclosure in an Online Community. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–24. https://doi.org/10.1145/3274363

[29] Juwon Lee, Omri Gillath, and Andrew Miller. 2019. Effects of self- and partner's online disclosure on relationship intimacy and satisfaction. *PLOS ONE* 14, 3 (March 2019), e0212186. https://doi.org/10.1371/journal.pone.0212186

[30] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. 2023. Online Self-Disclosure, Social Support, and User Engagement During the COVID-19 Pandemic. *Trans. Soc. Comput.* 6, 3-4 (Dec. 2023), 7:1–7:31.

[31] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 31:1–31:27. https://doi.org/10.1145/3392836

[32] Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. *Behaviour & Information Technology* 39 (2020).

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs].

[34] Xiao Ma, Jeff Hancock, and Mor Naaman. 2016. Anonymity, Intimacy and Self-Disclosure in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, 3857–3869. https://doi.org/10.1145/2858036.2858414

[35] Philipp K. Masur, Natalya N. Bazarova, and Dominic DiFranzo. 2023. The Impact of What Others Do, Approve Of, and Expect You to Do. *SM + Society* 9 (2023).

[36] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–23. https://doi.org/10.1145/3359174

[37] Jenny McMahon, Kerry R. McGannon, and Catherine Palmer. 2022. Body shaming and associated practices as abuse: athlete entourage as perpetrators of abuse. *Sport, Education and Society* 27, 5 (June 2022), 578–591.

[38] Yelena Mejova and Anya Hommadova Lu. 2023. Gender in the disclosure of loneliness on Twitter during COVID-19 lockdowns. *Frontiers in Digital Health* 5 (Dec. 2023). https://doi.org/10.3389/fdgth.2023.1297983 Publisher: Frontiers.

[39] Anne Nobels, Charlotte Meersman, Gilbert Lemmens, and Ines Keygnaert. 2023. "Just something that happened?": Mental health impact of disclosure and framing of sexual violence in older victims. *International Journal of Geriatric Psychiatry* 38, 12 (2023), e6036. https://doi.org/10.1002/gps.6036

[40] Larissa Ott. 2013. Reputation in danger: Selected case studies of reputational crises created by social networking sites. (2013).

[41] Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the Language of Self-Disclosure across Corpora. Association for Computational Linguistics, Dublin, Ireland.

[42] Julian Risch and Ralf Krestel. 2020. Top Comment or Flop Comment? Predicting and Explaining User Engagement in Online News Discussions. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May 2020), 579–589.

[43] Bárbara Silveira Fraga, Ana Paula Couto da Silva, and Fabricio Murai. 2018. Online Social Networks in Health Care: A Study of Mental Disorders on Reddit. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. 568–573.

[44] Stacie Tay, Kat Alcock, and Katrina Scior. 2018. Mental health problems among clinical psychologists: Stigma and its impact on disclosure and help-seeking. *Journal of Clinical Psychology* 74, 9 (2018). https://doi.org/10.1002/jclp.22614

[45] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. https://doi.org/10.48550/arXiv.1908.08962

[46] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling Self-Disclosure in Social Networking Sites *(CSCW '16)*. https://doi.org/10.1145/2818048.2820010
[47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. (2022).
[48] Benjamin T. Wood, Olivia Bolner, and Phillip Gauthier. 2014. Student Mental Health Self-Disclosures in Classrooms: Perceptions and Implications. *Psychology Learning & Teaching* 13, 2 (June 2014). https://doi.org/10.2304/plat.2014.13.2.83
[49] Haejung Yun, Gwanhoo Lee, and Dan J Kim. 2019. A chronological review of empirical research on personal information privacy concerns: An analysis of contexts and research constructs. *Information & Management* 56, 4 (June 2019), 570–601. https://doi.org/10.1016/j.im.2018.10.001
[50] Azma Alina Ali Zani, Azah Anir Norman, and Norjihan Abdul Ghani. 2022. Motivating Factors to Self-Disclosure on Social Media: A Systematic Mapping. *IEEE Transactions on Professional Communication* 65, 3 (Sept. 2022), 370–391.
[51] Yiming Zhu, Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Yuyang Wang, and Pan Hui. 2024. A Study of Partisan News Sharing in the Russian Invasion of Ukraine. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 1847–1858. https://doi.org/10.1609/icwsm.v18i1.31430
[52] Arne Freya Zillich and Kathrin Friederike Müller. 2019. Norms as Regulating Factors for Self-Disclosure in a Collapsed Context: Norm Orientation Among Referent Others on Facebook. *International Journal of Communication* 13, 0 (June 2019), 20. https://ijoc.org/index.php/ijoc/article/view/10943 Number: 0.
[53] Wenxue Zou, Lu Tang, Mi Zhou, and Xinyu Zhang. 2024. Self-disclosure and received social support among women experiencing infertility on reddit: A natural language processing approach. *Computers in Human Behavior* 154 (May 2024).

A ETHICS STATEMENT

This research complies with the SAGE (Self-Assessment Governance and Ethics Form for Humans and Data Research) self-check process provided by the University of Surrey, UK for ethics approval. No governance risks or ethical concerns falling under the higher, medium, or lower risk criteria were identified so Ethics and Governance Application (EGA) was not required for this study. No unauthorized access or collection of private data has occurred during this research. All datasets created and collected are sourced from publicly available materials. Reddit, a platform used in this project, openly shares its content as free and open data. Since this project does not involve interactions with human subjects, there was no requirement for informed consent. We confirm compliance with the University's Code on Good Research Practice, Ethics Policy, and all relevant professional and regulatory guidelines. We highlight that we do not use the data to identify any user. The analysis is performed and reported at aggregate levels, thus minimizing the privacy risks. The examples shown in the paper are chosen at random, and the text is paraphrased (e.g., line 75, page 1) so that readers cannot directly search for the same text on Reddit.

B SELF-DISCLOSURE TYPES

Table 2 shows the disclosure types identified through the literature review in this paper. The highlighted boxes show the types that are used in this paper.

Table 2: Self-Dislcosure Types and Categories: Highlighted color shows the types that are used in the paper.

Category	Type
Identity	Birthday/Age, Ethnicity, Sexual Orientation, Location, Physical Appearance, Gender
Relationship	Family, Relations, Friendship
Profession/ Economic	Job/Finance, Education
Heath	General, Physical, and Mental Health
Group Affiliation	Religion, Politics, Community
Opinions/ Interests/ Feelings	Sports, Politics, Current Affairs, Religion, Administration

C SELF-DISCLOSURE SPECIFIC COMMUNITIES

Table 3 lists subreddits specific to *Sexual Orientation* self-disclosure. These subreddits are more likely to have disclosures related to the given type of self-disclosure. In contrast the general category of subreddit are not restricted to any self-disclosure.

Table 3: Examplars of subreddits' association with self-disclosure.

Self-disclosure Type	Subreddits
Sexual Orientation	'lgbt', 'bisexual', 'askgaybros', 'BisexualTeens', 'LGBTeens', 'actuallesbians', 'gay', 'asexuality', 'AreTheStraightsOK', 'me_irlgbt', 'SuddenlyGay', 'comingout', 'pansexual', 'TwoXChromosomes', 'AskGayMen', 'gaybros'
General Subreddits	'AskReddit', 'memes', 'cats', 'Showerthoughts'

D CLASSIFIER PERFORMANCE

Table 4 shows the performance metrics of the classifier.

Table 4: $F1$ performance for Roberta Fine-tuning.

Self-Disclosure	Precision	Recall	F1
Age	0.84	0.93	0.89
Ethnicity	1.00	0.90	0.95
Gender	0.86	0.60	0.71
Education	0.87	0.81	0.84
Health	0.88	0.84	0.86
Job	0.89	0.70	0.78
Location	0.79	0.88	0.83
Physical Appearance	0.81	0.87	0.84
Relationship	0.84	0.86	0.85
Religion	0.88	1.00	0.94
Sexual orientation	0.88	0.79	0.83