Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

#LetsTalk: Understanding Social Media Usage of Self-Harm Users in India

Garima Chhikara[1,2], **Abhijnan Chakraborty**[1]

[1] Indian Institute of Technology Delhi, India
[2] Delhi Technological University, India

Abstract

Mental health concerns, such as depression, pose significant challenges for support systems in effectively identifying affected individuals. On the other hand, people suffering from depression often find it easier to discuss on social media rather than in face-to-face interactions. Additionally, the development of distressing conditions typically arises from a multitude of factors accumulated over time rather than a singular event. To gain a fine-grained understanding of these facets, in this work, we perform a longitudinal analysis of the tweeting behaviour of Indian users who post content related to self-harm. We categorise users based on their posting frequency and examine various aspects including their social network, bio descriptions, tweeting preferences, temporal variations and cognitive indicators. By elucidating these nuances, we aim to contribute insights that could aid in the early detection of mental health issues and prompt timely intervention from support networks.

Introduction

Today, a large number of people around the world suffer from mental health issues like depression and anxiety. Mental health support systems are in place to help people who are struggling with mental health problems. However, these systems often face the challenge of identifying people who need immediate help. This is because people who are depressed may not always seek help, or they may not be able to identify that they need help. Worryingly, unless attended to in early stages, many individuals resort to deliberate self-injury to cope with their acute feelings. In fact, self-harm is reported to be one of the main causes of death at a young age (Rane and Nadkarni 2014). Since 2005, the percentage of teenagers and young adults who experience major psychiatric distress, including suicidal ideation and suicide attempts, has climbed by 71% (Twenge et al. 2019). In India, around 85% of the total lives lost due to self-harm between 2015 and 2019 belonged to the age group 15 to 49 (Rane and Nadkarni 2014).

Besides identification, mental health support systems face another challenge – people suffering from depression find it hard to express their feelings because they are ashamed and have a constant fear of being judged (Wasserman et al. 2012). Social media platforms like Facebook, Instagram, X (Twitter) provide space for users to connect with others, build communities, share personal life events, spread awareness, and overcome fear and stigma (Ernala et al. 2018). Almost 60% of the Indian population are active on social media (Basuroy 2022a). Apart from entertainment, social media has evolved as a significant resource of information on mental health. Researchers have leveraged social media platforms for analysis and detection of self-harm content (Chancellor, Baumer, and De Choudhury 2019; Aldhyani et al. 2022; Ophir et al. 2020). Social media like X allows users to be pseudo-anonymous (disguised under imaginary profile names), allowing people to reach out to others for help without revealing their actual identity (Coppersmith et al. 2018; Robinson et al. 2015; De Choudhury et al. 2021). This also helps to mitigate the bias present in surveys and self-reported data (Yuan et al. 2023).

While there have been multiple studies on mental health issues observed from the prism of social media (Aldhyani et al. 2022; Ophir et al. 2020), we believe that such studies need to be undertaken in a particular geopolitical and cultural context and in this work, we focus our attention on India. Every nation possesses its own set of challenges, and India is no exception. With a growing economy and the dubious distinction of being the most populous nation with 1.43 billion people, India has been witnessing numerous highs and lows over the last few years (Mozumder 2023; Chatterjee 2023), ranging from the crises of unavailability of hospital beds and medical oxygen during the height of COVID-19 pandemic in May 2021 (Ghoshal 2021), the plight of the migrant workers during nationwide lockdown (Varma 2019) highlighting societal inequalities, to farmers' protest against three agricultural laws.[1] Moreover, often *Dalits* and religious minorities[2] have faced the problem of identity, security, social discrimination, communal tensions and riots (Insights 2022; LotusArise 2022; Nations 2021). Such socio-political activities have great implications on person's men-

[1]The farmers' protest took a toll on the agitating farmers' mental health leading to three suicides (BBC 2021).

[2]In December 2019, the Indian Parliament passed Citizenship Amendment Act (CAA) to provide a route to citizenship to members of six religious communities from Pakistan, Bangladesh and Afghanistan but not for Muslims, which led to nationwide protests spearheaded by Indian Muslims (BBC 2019).

Issues	Post
Covid Oxygen	My friends father was on the way to hospital and said pls see that I come back home, I dont want to die gasping on the want of oxygen. And on day 3 we get a call that he is no more. These stories will keep on going in our head with pain.
Labour Rights	Here in India - Migrant workers have died and are dying. Poor are dying. Children are suffering. Speak up India. Speak up. #SpeakUpIndia #DeathsofMigrantWorkers #MigrantLaboursDying #SocialRevolution #WorkersUnite #EndHunger #GiveJobs #EqualPay #ConstitutionalRights
Farmer Protest	How do we account for the harms caused to so many farmers and their families during the period preceding the withdrawal? How culpable is the union government and what should be the appropriate reparation? #FarmLawsRepealed #FarmersProtest"
Minority	#Minority communities could not avail govt benefits due to lack of documentation and fear of discrimination. #Muslims fear approaching state institutions for any kind of redressal for the fear of discrimination. No families received scheme benefits in 4%-34% of 53 Muslim areas.
Dalits	Dalit student other than me from our batch was attacked with so many questions in school. And we both had only each other to share these everyday 'casteism' we had to face from school.
CAA	SOS - I am currently being detained at Nagpada Police Station. The police is targeting me now for exercising my fundamental right to peacefully protest against CAA/ NRC. @MumbaiPolice I have not violated any laws. I will take this further to the courts if need be.

Table 1: Tweets highlighting major challenges faced by Indians over past few years (author info omitted for privacy).

tal health. Table 1 shows some example tweets about the major issues that have been surfacing in India over past years.

Earlier works on mental health of Indian social media users (Di Cara et al. 2023; Roy et al. 2021; Lathabhavan 2020; Thippaiah, Nanjappa, and Math 2019; Kumar and Nayar 2021; Barkur, Vibha, and Kamath 2020; Khasawneh et al. 2020) have looked at a binary categorisation: one either has mental health issues or not. However, in reality, people gradually transition from the initial stage of anxiety to depression, leading to extreme steps like suicide (Ageitos et al. 2021). Identification of this transition in the early phase can help in preventing cases of self-harm. Furthermore, changing social, political and personal elements can have an additive effect on a person's mental discomfort. Depression and anxiety level keeps on changing with external factors and thus there is a need to comprehend the change in users behaviour over time. Towards this, in this work, we undertake a longitudinal analysis of Indian social media users, categorising them into three categories based on the posting frequency of afflicting content – low, moderate and high frequency users.

Collecting extensive longitudinal data from X between 2017 and 2021 for all three user categories, we try to answer the following research questions in this paper: How engagement and activity levels vary across user categories? How users across different categories describe themselves? How tweeting preferences, temporal characteristics, and cognitive traits evolve over time for various user categories?

According to our analysis, users who post self-harm content more frequently follow many individuals but have fewer people following them. As we transition from high to low frequency users, the trend seems to reverse. The largest number of followers and lowest number of following are among low frequency users. Most high frequency users connect their identity with their hobbies and passions, while a larger proportion of low and moderate frequency users associate themselves with personal attributes and are vocal about their concerns. We also examine the reactions of other users to tweets posted by self-harm users. In contrast to tweets from low and moderate frequency users, tweets by high frequency users receive more comments. Additionally, self-harm content posted by high frequency users receive more attention as compared to non self-harm content posted by the same user set. High frequency users use mentions to spread the word about their issues, and they are often seen addressing similar topics in their original tweets, retweets, and comments.

Overall, our work helps in uncovering different categories of users undergoing mental health issues and how their social media usage evolves with time. We hope that this work will lead to follow-up efforts to identify early onset of mental health issues and enable support systems to act accordingly.

Related Work

Social media has become a crucial tool for analysing mental health as it is easily accessible (Kosinski et al. 2015), overcoming a number of difficulties in cutting-edge clinical mental health evaluation techniques that depend on subjectivity and retrospective recall bias (Lazer et al. 2009). Considerable work is done in identifying and predicting depression among social media users (Coppersmith et al. 2018; De Choudhury, Counts, and Horvitz 2013; De Choudhury et al. 2021); implication of anonymous nature of platform on self-disclosure (Andalibi et al. 2016; Ernala et al. 2017); comparing the online and offline mental health behaviours (Saha et al. 2017); comprehending social support measures to encourage positive mental health (Andalibi, Ozturk, and Forte 2017; De Choudhury and Kiciman 2017); identifying the norms and practices of community (Chancellor et al. 2016); and how social platforms can be leveraged to provide support (Inkster et al. 2016). Previous studies (Eichstaedt et al. 2018; Owen et al. 2023; Schemer et al. 2020; A. et al. 2022) have conducted longitudinal analysis to predict depression using social media data. However, our research

hang myself	hate myself	am lonely
self injury	take my life	want death
self harm	end my life	kill myself
want to die	been suicidal	cut myself
suicidal thoughts	commit suicide	have no one

Table 2: Keyphrases used for collecting self-harm posts.

specifically targets Indian users, offering a more comprehensive examination of posting behaviors and user traits. To the best of our knowledge, this is the first work that performs longitudinal analysis of self-harm users in India based on the users posting frequency.

Dataset Gathered

Our dataset curation went through the following stages: (i) collect tweets with several keyphrases, (ii) manually annotate them to identify true instance of self-harm and get rid of false positives, (iii) collect timeline of self-harm users, (iv) gather a collection of random tweets, and (v) build and apply a binary classifier to label timeline posts as self-harm or non self-harm. Figure 1 shows various stages of data collection.

Collecting Self-Harm Tweets

We collected all public tweets with geolocation India between 2017 to 2022 which included any of the keyphrases mentioned in Table 2, using the (now defunct) Twitter API with academic research access. These key phrases were curated using the phrases mentioned in prior literature (Coppersmith, Dredze, and Harman 2014; Coppersmith, Harman, and Dredze 2014). Through this data collection process, we obtained 3,759 posts from 3,026 unique users. However, the mere presence of these keyphrases does not necessarily indicate intention of self-harm. Table 3 shows example instances where the presence of a keyphrase can denote both presence and absence of self-harm intent. While there are tweets depicting real cases of anxiety, depression and self-harm; there are also tweets on disappointment with government rules and policies, political and economic situation, death of a famous celebrity, criminal offence or suicide in neighbourhood, suicide awareness and prevention, dialogue from a movie or drama, complaints about customer services etc.

To separate the true instances from false positives, we manually annotated 3,759 posts as valid (true case) or invalid (false case). We employed three annotators, and disagreements were resolved by a majority vote.[3] We find 1,147 posts from 939 users depicting actual instances of mental affliction. Since these 939 users have posted at least one mentally alarming post between 2017 and 2022, we refer to them as the *Focus Group*. We concentrate on this focus group for further analyses.

Compiling X Timelines of the Focus Group

Longitudinal studies have proven to be useful in explaining how people behave in various situations. We analyse the focus group over time from 2017 to 2022. Towards this, we attempted to collect the X timelines of all 939 users from the focus group. We found 49 accounts to be deleted or suspended, and gathered the timeline data for the remaining 890 users. The Twitter API generates an error message 'user not found' or 'user has been suspended' for unavailable accounts, we utilise this information to monitor deleted and suspended accounts. Overall, we obtained 1,385,535 tweets posted by the 890 users of the focus group.

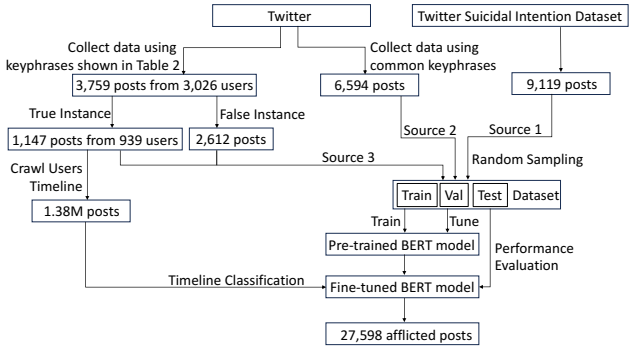


Figure 1: Steps followed for collecting data from users timeline. Transformer based BERT model is fine-tuned using data from three different sources. Model reported F1-score of 0.91 on test dataset. Fine-tuned model classified 27,598 posts as mentally alarming.

Classifier for Automated Labelling

For further analyses, we need to identify how many among these 1.38M tweets are related to self-harm. Since manual annotation of 1.38M tweets is challenging, we use the state-of-the-art transformer based BERT model (Devlin, Chang, and Lee 2019) for the classification task. Previous research has indicated BERT's efficacy in binary classification tasks in the realm of mental health (Owen et al. 2023). For our task, we fine-tune the BERT base model to classify a post as self-harm or non self-harm.[4]

Data for fine-tuning. The data for fine-tuning and testing is collated from three different sources: (i) *Twitter suicidal intention dataset* [5] with 9,119 tweets, amongst which 3,998 posts are labelled as suicidal intention and 5,121 are labelled as non suicidal intention; (ii) we additionally collect 6,594 tweets from X using keywords like 'shopping', 'travel' and 'happy'. Rationale behind this is that timeline posts can deal with routine activities and may not depict self-harm content always, hence such data is necessary for training the classifier;[6] and (iii) the initial manually annotated dataset with 1,147 self-harm tweets and 2,612 non self-harm tweets.

[3]The annotators are aware of the Indian context and have a good understanding of self-harm and non self-harm text.

[4]The classifier code is available at https://github.com/garimachhikara128/MentalHealth/

[5]The dataset is available at https://github.com/laxmimerit/twitter-suicidal-intention-dataset/

[6]We assume that since these tweets are obtained through keywords denoting happy state of mind, they are non self-harm posts.

Self-Harm Intent Present
I've decided to by myself take psychiatric help to give myself one last chance. If I see no improvement then I may just **end my life**.
My life at the current stage is making me to want to **end my life**. I'm slowly losing all purposes that have kept me alive to this day. I don't know how much longer before I can no longer take it and finally **end my life**.
Today's very very bad day, Anniversary of my parents, I'm the very unlucky daughter, I **hate myself** and my worst fate.
Yesterday and for many months before that, I've had intense stages of **self harm**. And these happen late, usually around and after 12 AM. And there's no suicide prevention helpline available. I called up one yesterday because I had self harmed and they asked me to try Savasana ???
Self-Harm Intent Absent
It would have **been suicidal** for congress to let priyanka fight from Varanasi. She is just entering political arena and would have surely lost big from Varanasi. A safe seat like Wayanad is required for her..!
After doing all household chores, I am loving to spend my time on Instagram, YouTube and WhatsApp more these days. I am not much on Facebook because I want to **cut myself** from all the food pics.
It's end of the crossroads now, wanna **end my life** long relationship with @Airtel_Presence, wanna port my SIM into @JioCare, Please let me know the process!
Today at 12:25AM, I have spent more than 45 mins in total waiting and searching for a ride at @Uber_India. I almost **hate myself** for being in situations like these where I need to rely on your service despite of the certainty of your services.

Table 3: Example instances showcasing presence and absence of self-harm intent.

Data	**Original**			**Sampling**			**Train + Validation**			**Test**		
	Class 0	Class 1	Total	Class 0	Class 1	Total	Class 0	Class 1	Total	Class 0	Class 1	Total
D1	5121	3998	9119	2000	3998	5998	1600	3198	4798	400	800	1200
D2	6594	0	6594	2000	0	2000	1600	0	1600	400	0	400
D3	2612	1147	3759	1140	1142	2282	912	913	1825	228	229	457
							4112	4111	**8223**	1028	1029	**2057**

Table 4: Table showcasing equal distribution of training, validation and test data from each data source. *Original* indicates the actual data collected through existing repository and through X API. Random *Sampling* is done to guarantee equal proportionality from both classes. 80% of the sampled data is used for training and rest 20% is used for testing. Amongst training data, 80% is used for fine-tuning the model and other 20% is used for validation. Test data is employed to evaluate the model's performance.

Then, we randomly sample the dataset to ensure an equal number of tweets from both classes, resulting in 8,223 tweets in the training+validation dataset and 2,057 tweets in the test dataset. Table 4 shows the detailed breakup of posts used for training, validation and testing. We observe that the final fine-tuned BERT model exhibits good performance on the test dataset with an accuracy of 91%. After applying this classifier to 1.38M posts collected from different users timelines, 27,598 tweets are classified as self-harm and the remaining tweets are classified as non self-harm.

Categorising Users in the Focus Group

Based on the frequency of self-harm posts over the years, we categorise the users from the focus group into three categories – **low frequency users**, **moderate frequency users**, and **high frequency users**. Figure 2 shows the frequency of self-harm posts between 2017 and 2022 posted by all users. We define low frequency as those who posted less than 50 self-harm posts, moderate frequency who posted more than 50 but less than 150 tweets, and high frequency who posted more than 150 tweets. Details about the user distribution across these categories are shown in Table 5. We observe that 1%, 3% and 8% of tweets posted by low, moderate and high frequency users respectively reveal the presence of mental discomfort. This suggests that when compared to low and moderate frequency users, high frequency users not only post more self-harm content but also a bigger percentage of their posts are distressing.

Characterising Different User Categories

In this section, we examine the long-term trends among different user categories based on social media neighbourhood, bio description, tweeting preferences and temporal variations. While there have been multiple works focusing on mental health concerns, to the best of our knowledge, this is the first work to consider the longitudinal differences amongst Indian users based on the posting frequency.

Social Media Neighbourhood

The number of followers and followees of a user reveal their attempt at socialising as well as the willingness to consume outside information. Earlier works have shown that power law best fits the in-degree distribution (followers), whereas log-normal best fits the out-degree distribution (followees) (Myers et al. 2014; Lerman, Yan, and Wu 2016). Interestingly, in our case (as shown in fig. 3), we observe that

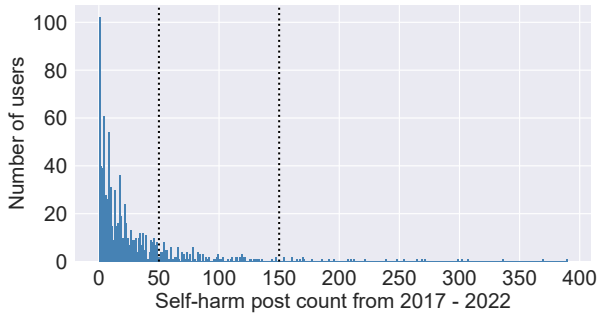


Figure 2: Total number of self-harm posts made by 890 users of the focus group from 2017 - 2022. We select 50 and 150 as threshold for low, moderate and high frquency users. Users with less than 50 self-harm posts are low frequency users, users with self-harm post count between 50 and 150 are moderate frequency users, and users with self-harm post count greater than 150 are high frequency users.

Category	No. of Users	No. of Posts	No. of SH Posts	Fraction of SH Posts
low	739	1023919	10892	0.011
moderate	122	286042	10086	0.035
high	29	75574	6620	0.088

Table 5: Distribution of self-harm (SH) users and posts across various categories.

both follower and followee counts obey log-normal distribution for all three user categories in the focus group. Table 6 shows the value of exponent α for power law and value of mean and standard deviation i.e., μ and σ for log-normal distribution. Moderate frequency users are reported to have the highest mean μ across both the measures of followers and following and it is interesting to observe that the standard deviation σ is the lowest across medium frequency users, indicating that most of the followers and following of moderate frequency users are clustered around the mean, whereas for low and high frequency users the spread is relatively higher as compared to moderate frequency users.

Engagement. To assess levels of engagement across three user categories, we establish three measures based on the volume of users followers, following and tweets (Wang et al. 2017). We characterise the level of engagement of category c in terms of measure m as follows:

$$Engagement(c, m) = \frac{\sum_{u=1}^{n} \#m_u}{n} \quad (1)$$

where $c \in \{Low, Moderate, High\}$ and $m \in \{Followers, Following, Tweets\}$, n are the number of users in category c and $\#m_u$ denotes the count of measure m for user $u \in c$.

Table 7 lists the statistics for social engagement. Low frequency users show the maximum engagement for #followers and high frequency users exhibit the maximum engagement

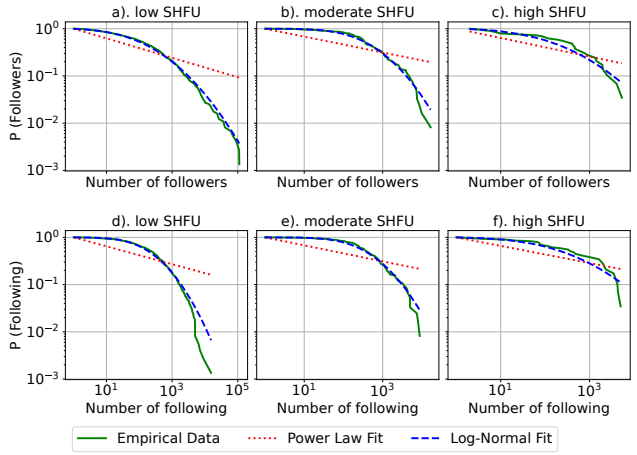


Figure 3: Complementary cumulative distribution function (CCDF) for count of followers and following. Red and blue curves denote fitted power law and lognormal distributions across low, moderate and high self harm frequency users.

Measure	Category	Power Law α	Log-normal μ	Log-normal σ
Followers	low	1.205	4.661	2.589
	moderate	1.168	**5.955**	**1.820**
	high	1.197	4.948	2.456
Following	low	1.189	5.272	1.766
	moderate	1.169	**5.903**	**1.673**
	high	1.181	5.415	2.521

Table 6: Statistics for distribution of followers and following. The parameter α assumes $P(x) \sim x^{-\alpha}$ for degree x (power law). The μ and σ parameters assume $P(x) \sim \frac{1}{x}\exp[\frac{(\ln x-\mu)^2}{2\sigma^2}]$ (log-normal).

for #following and #tweets. One might argue that it was implied for high frequency users to receive the maximum engagement on #tweets since this itself was the categorisation approach, but note that our method of categorisation is exclusively based on the number of **self-harm posts** regardless of the total number of posts. In addition to posting more instances of self-harm, high frequency users have the highest #tweets across all user types. Earlier works have reported the shrinkage of networks as a useful indicator of impending behavioural changes (De Choudhury, Counts, and Horvitz 2013). In contrast to low frequency users, who have a bigger network of #followers, high frequency users have less individuals following them. This suggests that users who publish more distressing content are less likely to be followed by other users. High frequency users have highest #following, which indicate they are more intrigued about others as compared to low and moderate frequency users. Kolmogorov-Smirnov (KS) test (Lilliefors 1967) revealed the results to be statistically significant with $p < 0.05$.

Activity. We investigate which user category is most active on Twitter. We quantify activity as the average number

Measure	Low	Moderate	High
Engagement			
#Followers	**1749.51**	1372.76	817.07
#Following	643.84	1035.57	**1203.97**
#Tweets	1385.55	2344.61	**2606.00**
Activity			
#Followers/day	0.65	**0.72**	0.37
#Following/day	0.61	0.65	**0.94**
#Tweets/day	1.05	3.06	**11.1**

Table 7: Social Media Neighbourhood: Statistics for Engagement and Activity across various user categories.

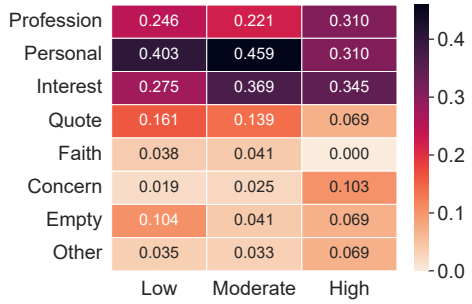


Figure 4: Normalised score for labels across three different user categories.

of followers, following and tweets per day.

$$Activity(c, m) = \frac{\sum_{u=1}^{n} \left(\frac{\#m_u}{d_u} \right)}{n} \quad (2)$$

where d_u is the number of days between – user u joining X and his last post. Higher activity levels indicate higher expressiveness (Ernala et al. 2017; Saha, Weber, and De Choudhury 2018; De Choudhury et al. 2021; Yuan et al. 2023). Moderate frequency users have active #Followers/-day whereas #Followers are not the highest for them, this indicate users with moderate frequency are engaged for a shorter amount of time on Twitter.

Bio Description

Users embark on significant communicative and identity disclose properties when they create their unique accounts, e.g., by filling up the bio sections, as opposed to being confined solely to the content that gets posted thereafter (Greene and Brownstone 2023; Chakraborty et al. 2017). We analyse bio description of the 890 users from the focus group and label them as 'profession', 'personal', 'interest', 'quote', 'faith', 'concern', 'empty', and 'other'.

Profession: where a user describes something about their profession, e.g., *"Associate Director"*, *"Practicing Gynaecologist"*, *"Assistant Professor(Pharmacology)"*, *"Fitness Coach"*.

Personal: information regarding age, gender, birthday, education, location, religion, qualities and relations, e.g., *"she/her 24"*, *"Astrophysics Scholar"*, *"Believe in Secularism — Peace — Love"*, *"A proud Resident of Bihar"*, *"Proud Father"*.

Interest: where users highlight interests or hobbies, e.g., *"BTS : Justin"*, *"Interest in Photography & Documentary Making"*, *"a singer and baker by hobby"*.

Quote: users use well known quotes or dialogue from movies or pen down their thoughts in bio description, e.g., *"Emptiness is better than something temporary"*, *"When life gives you lemons, make lemonade of it"*.

Faith: user mention about religion and God, e.g.,*"God Protect Me"*, *"Jai Shree Ram"*, *"Inservice to Lord Krishna"*, *"Burn the evil spirits and demons now and forever. Amen omshanthi Allah"*.

Concern: user raises concern regarding personal or societal issues, e.g., *"#cancel 12th board exam"*, *"Alone boy. Need adoption or family"*, *"Dream IIT but no money and guidance. Please help me for higher education."*, *"My life Dedicated towards Saving & Serving Innocent Speechless Souls (Animals)..In real,trying to be humane"*.

Empty: user prefers to leave their bio description unfilled.

Other: bio description is not significant enough to convey meaning, e.g., *'#'*, *'eh.'*, *'...'*, *'anyway,,,'*.

People use bio to express their *self-identification*, and individuals can opt to do so from a variety of perspectives, hence a bio description can be tied to several labels. Figure 4 shows the normalised count of labels across different user categories. 40.3% of low frequency and 45.9% of high frequency users mention about their *personal* lives in their bio descriptions, followed by *profession* and *interest*. Majority of the low and moderate frequency users discuss their personal traits on social media, on the contrary only 31% of high frequency users mention personal elements, which suggests fewer high frequency users prefer to divulge their personal identities on social media. The biggest percentage of high frequency users (34.5%) mention their amateur interest, suggesting that they possess or at least identify themselves with some leisure pursuits. An equal percentage (31%) of high frequency users mention about *profession* and *personal* elements, among high frequency users *identification* is nearly equally distributed across *profession*, *personal* and *interest*. Faith is mentioned in 3.8% and 4.1% of user bios from low and moderate frequency users respectively, on the contrary none of high frequency users referred to or exhibited trust in God in their bio descriptions. Low and moderate frequency users have only 1.9% and 2.5% bio descriptions tied to *concern*, whereas, among high frequency users 10.3% of tweets address critical issues, suggesting high frequency users are relatively more vocal about venting concerns.

Tweeting Preference

We analyse how users from the focus group engage with one another – interaction with posts and publishing of posts (Wang et al. 2017).

Post Interaction. We consider three measures for interaction with post - replies, likes and mentions. Figure 5 shows average number of replies, likes and mentions received on ***all set*** and ***self-harm set*** for different user categories.

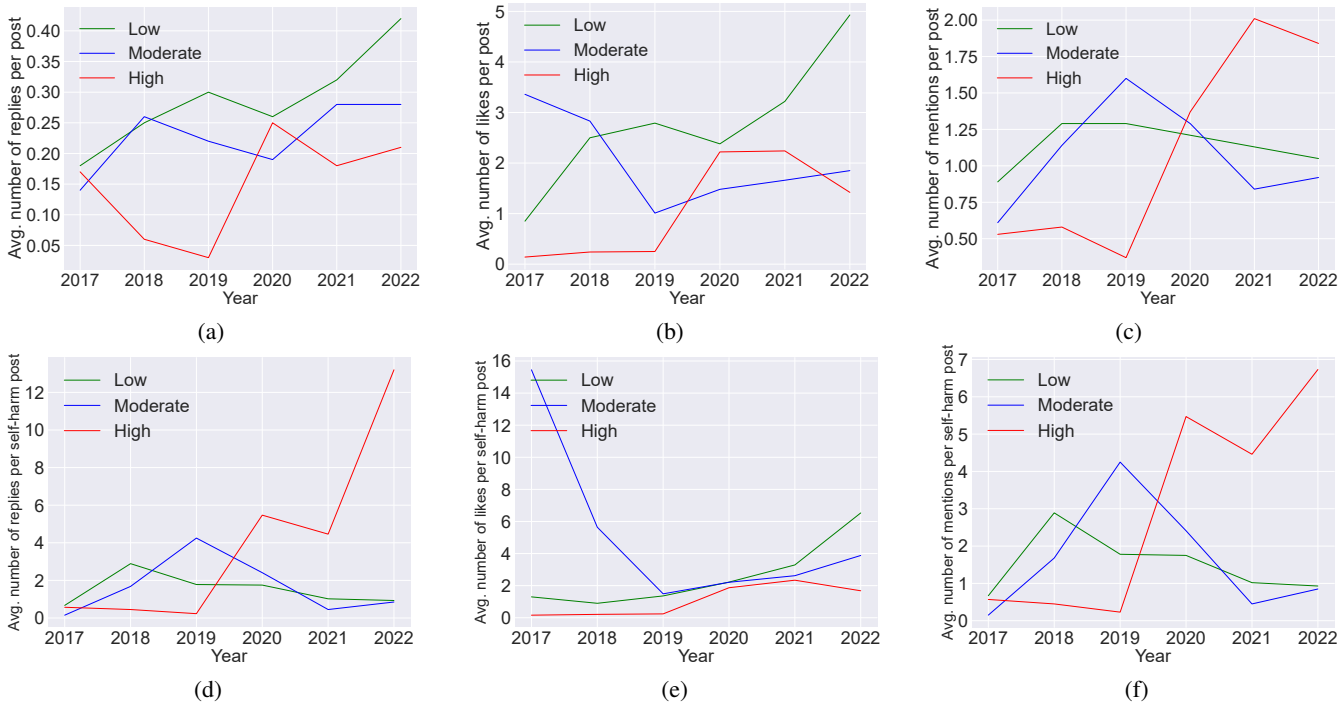
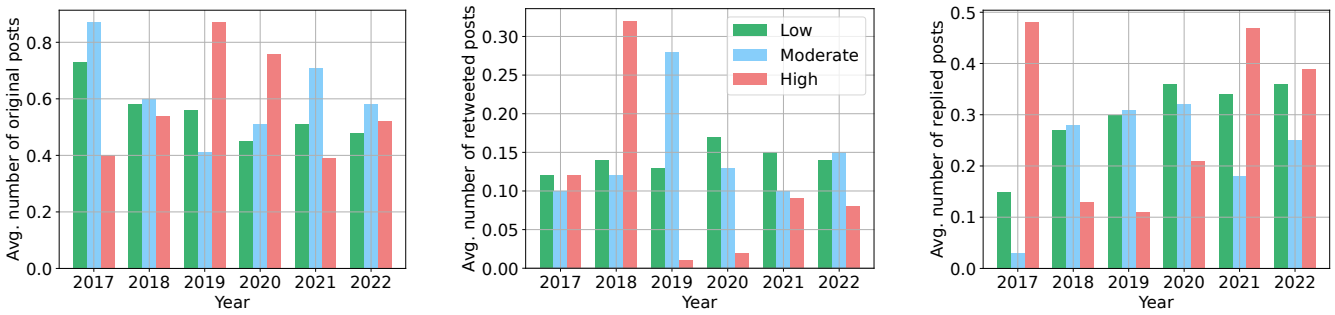


Figure 5: Post Interaction from 2017–2022 for various user categories. Average number of (a) replies (b) likes (c) mentions received on all set. Average number of (d) replies (e) likes (f) mentions on self-harm set.

We first compare the measures across all set and self-harm set. All set comprises of the entirety of timeline posts made by users of the focus group, self-harm set refers to mentally afflicting posts from the focus group's timeline. i) Replies for all set vary between [0,0.4] (fig. 5a), whereas for self-harm set replies range between [0,12] (fig. 5d). Self-harm set exert higher interaction – the average number of replies are greater in comparison to all set, people on social media exhibit higher interaction with self-harm posts indicating higher social assistance, some earlier works have shown replies can act as a proxy for social support (De Choudhury and Kiciman 2017). ii) Number of likes for all set varies between [0,5] (fig. 5b), whereas the spread for self-harm set is [0,16] (fig. 5e), however if we consider year 2018 and later, the number of likes drop to range of [0,7], which is sufficiently close to the likes received on all set. If we investigate recent years, we can deduce that all set and self-harm set receive a nearly identical pattern of interaction with regard to likes. iii) Figure 5c and 5f illustrates the mention count on all set to range between [0,2] and self-harm set to range in [0,7] respectively. Self-harm set have a higher count of mentions as compared to other posts, which indicates the intent to disseminate the self-harm post to a broader audience.

Secondly, we compare user categories across various measures. i) Replies. On all set, low frequency users receive highest replies and high frequency users seek minimum replies (fig. 5a), whereas the trend seems to be reversed when we consider the self-harm set – high frequency users gather maximum and low frequency users receive minimum replies (fig. 5d). This reveals that people on social media converse more with low frequency users when they post non self-harm content and people engage more with high frequency users when they post self-harm content. ii) Likes. Over recent years, from 2019 – 2022, the pattern of likes across user categories has been similar for all set and self-harm set (fig. 5b and 5e), low frequency users receive maximum number of likes whereas high frequency users obtain minimum number of likes. This indicates social media users do not prefer interacting with high frequency users through likes. iii) Mentions. Usage of mention implies the want to let the other party know about the problem or anticipate a solution. Mention for high frequency users increased drastically since 2019, high frequency users avail the maximum usage of mentions across all set and self-harm set (fig. 5c and 5f) indicating they want their posts to reach to the maximum audience, and in haste, the users tag all potential helpers. While we observe an increase in mentions for high frequency users since 2019, there is a dip in mentions for low and moderate frequency users across all set and self-harm set. KS test indicates statistical significance for results shown in Figure 5 with $p < 0.001$.

The takeaway is high frequency users garner maximum replies for self-harm posts, whereas for general posts low frequency users attract the most replies. For year 2018 – 2022, statistics for likes and mentions follow similar patterns for various user categories across all set and self-harm set. Likes received by a user on a normal post and on a self-harm post are alike, and likes are not the preferred way of interacting with high frequency users. Self-harm posts comprise higher mentions as compared to general posts.



(a) (b) (c)

Figure 6: Fraction of (a) original (b) retweeted and (c) replied posts in all set from 2017 - 2022 for various user categories.

Post Type	User Category	Distinct Frequent Keywords
original	low	cope, feelings, depressed, stressful, crying, unhappiness, emotionless, mistreatment, crying, attachment, loneliness, distress, longing, grief, stress
	moderate	existence, imprisons, calm, silenced, deprivation, isolate, relativesam, memories, suicide, psychopath, awareness, hide, disappear, paranoia, invisibility
	high	admission, prevention, exams, intervene, discrimination, depression, anxiety, stressed, sufferings, hospitalisation, cbse, exams, pandemic, students, covid
retweet	low	dharmapuri, prayer, spiritual, coping, depressed, treatment, depression, patanjali, psychotherapy, pranayama, ayurveda, meditating, buddha, yoga, depressant
	moderate	meditation, addictive, smoking, overcome, remorse, suicides, mindfulness, therapy, psychiatric, depressed, addict, debilitating, suicidal, craving, medications
	high	study, stress, grades, delaying, appeal, student, universities, silence, education, tuition, protest, campus, prevention, admission, exam
reply	low	abused, oppression, slander, threatening, crimes, murders, vaccination, grievance, assaulted, violence, hypocrisy, pervert, victim, rapists, molestation
	moderate	husbands, disorders, loneliness, marriage, dysfunction, introvert, societal, depression, stigma, divorce, therapy, coping, patriarchal, matrimonial, feminists
	high	castes, student, cbse, delhi, panchayat, conferred, bharat, exam, declared, india, xii, assessment, mandir, examination

Table 8: Distinct high frequency keywords in original posts, retweets and replies.

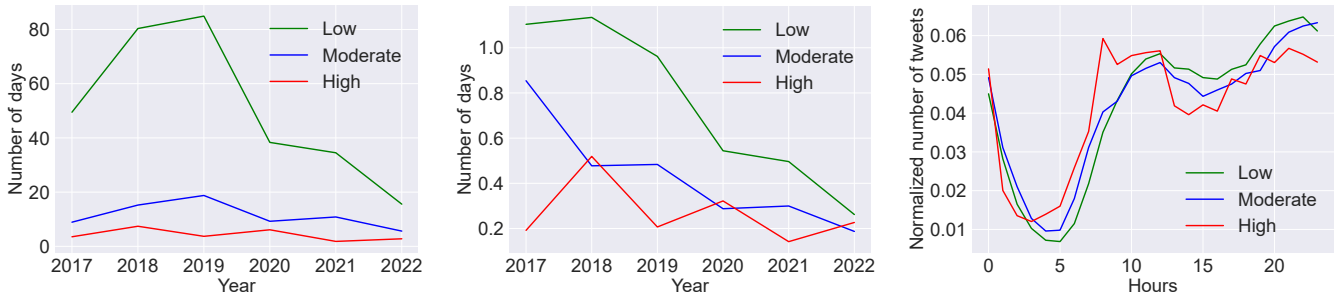
Publishing Posts. Users can publish posts in three ways: original post, retweet, or reply. We analyse the timeline posts of three user categories. Original post initiates a new discussion or thread, retweeted post is the retweet of another post, and replied post is the comment provided on another post.

In 2021 and 2022, all three user categories posted more original content followed by replied and retweeted posts (fig. 6). From 2017 - 2022, retweeted posts are observed to be minimal across all user categories, which indicate retweet is not extensively used by users.

We find the most frequently used words in original posts, retweets and replies across different user categories (table 8). We leverage the use of KeyBERT for keyword extraction, KeyBERT utilises BERT embeddings to generate keyphrases and keywords that are the closest to the given document (Grootendorst 2020). New conversation initiated by a low frequency user mostly talks about depression, stress, anxiety, and grief whereas a moderate frequency user posts about invisibility, isolation, and question their existence. In retweets, users suggest methods for coping with various mental issues through meditation, pranayama which is a breath regulation exercise, yoga, prayer, and medication. In replies, users are more expressive about their personal lives and share their perspectives about society, family, and personal matters like violence, divorce, and stigma. High frequency users, across all the posts, often discuss about exams, studies, students, results, and grades; this section of self-harm users seems to be mostly students and are trying to raise their concerns via new threads, retweets and replies. There is variation in the posts made by low and moderate frequency users across new posts, retweets and replies. Whereas high frequency users seem to be constantly discussing about the same issue across all the posts.

Temporal Variation

We study the relation between user categories and variation with time, we utilise the entire timeline data of 1.38M posts for temporal analysis. Low frequency users report highest average time gap between any two successive self-harm post, followed by moderate and high frequency users (fig. 7a). One might argue, this observation was implied as low frequency users post the least amount of self-harm posts, to



(a) (b) (c)

Figure 7: (a) Average time gap between two successive self-harm post (b) average time gap between any two successive post (c) time of the day when a post is created.

answer this, let us consider an example, a user posts three self-harm posts on day 1, day 4 and day 11 respectively, average time gap between two successive posts is 5 days. On the contrary, if the user posts on day 1, day 7 and day 15, the average time gap will be 7 days. In both cases, user posted three tweets but the average time gap differs, our method of categorisation is solely dependent on the number of self-harm posts. High frequency users tend to post self-harm content more frequently.

Similar statistics are observed for the average time gap between any two successive posts, low frequency users are reported to have the highest time gap (fig. 7b), which however has tremendously decreased over the past years. Low frequency users are less involved towards self-harm and general posts. On the other hand, high frequency users exhibit elevated levels of activity for both self-harm and general posts. If we look at the broader pattern, since 2017, the average time gap on self-harm posts has decreased for all three user categories, indicating people are getting more engaged and responsive to posts. KS test revealed results of fig. 7a and 7b to be statistically significant with $p < 0.001$.

We analyse the time of the day when users post. People with mental health issues may experience insomnia, they might act differently from regular users in terms of their online activities (Lustberg and Reynolds 2000). High frequency users are more engaged as compared to low and moderate frequency users from 3 am to 9 am and are least active from 11 am to 7 pm (fig. 7). Sleep cycle is particularly worse for high frequency users, who appear to be most active in early mornings and less active during the day. Prior work has suggested that disrupted sleep is one of the factors that is closely associated with depressive illnesses (Abdel-Khalek 2004). High frequency users are most prolific during odd hours which raises concerns about their mental health.

Cognitive Attributes

We study cognitive measures such as readability, complexity and repeatability for various user categories in all set.

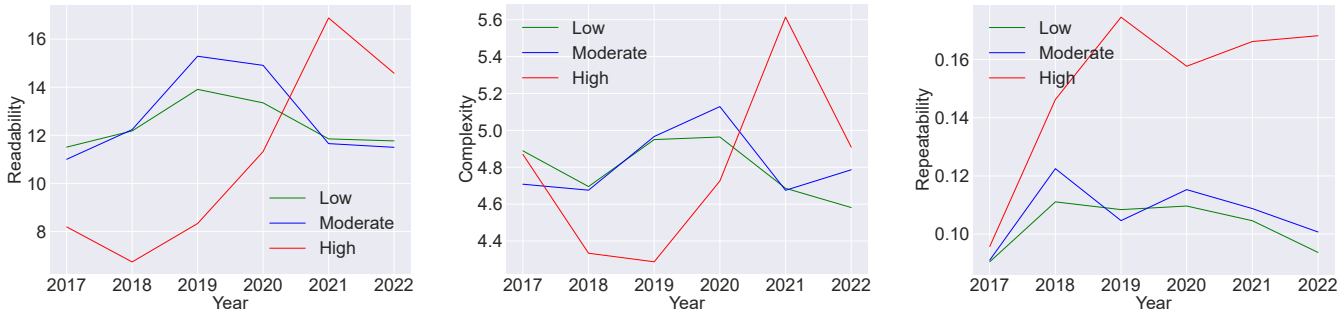
Readability. Readability is a measure to gauge the ease with which readers may understand a certain text (McCallum and Peterson 1982). Readability is an important indicator of people's cognitive behaviour, and earlier research has utilised this metric to comprehend the patterns of conversation in social networks (Ernala et al. 2017; Saha, Weber, and De Choudhury 2018; Saha and Sharma 2020). We employ Coleman-Liau Index (CLI) to calculate readability. CLI is calculated as, $CLI = 0.0588 * L - 0.296 * S - 15.8$, where L is the average number of letters per 100 words and S is the average number of sentences per 100 words (Pitler and Nenkova 2008).

We analyse the readability index in all set for different user categories (fig. 8a). CLI for low and medium frequency users are observed to be greater than 11 throughout the timeline of six years, depicting that the data is fairly difficult to read. For high frequency users CLI was lower than 8 until year 2019, indicating the text is ideal for average readers, but since then CLI has increased and in recent years is found to be higher than low and moderate frequency users. This indicates, as opposed to earlier years, high frequency users are now expressing their thoughts in fairly complex language.

Complexity. Complexity is the average length of words per sentence (Ernala et al. 2017; Saha and Sharma 2020). Since 2019, there has been an increase in the complexity for high frequency users with a decline in complexity for low and medium frequency users (fig. 8b).

Our analysis reveals that, since 2019, for low and medium frequency users there is a decrease observed in readability and complexity, and there is an increase for high frequency users. Earlier works (Ernala et al. 2017) have shown psychosocial health to have a positive correlation with readability and complexity. Note that, the data set used by (Ernala et al. 2017) is from year 2012 to 2016, which is disjoint with our data set which is from 2017 to 2022. If we base our analysis on the arguments made by (Ernala et al. 2017), from year 2017 to 2019, high frequency users show low scores for readability and complexity, indicating lower well-being. With many new Indian people joining X over the past few years (Basuroy 2022b), posting behaviour has changed substantially, high frequency users form more complex sentences as compared to low and moderate frequency users.

Repeatability. Repeatability is the normalised count of non-unique words, and correlates negatively with psychosocial health (Ernala et al. 2017; Saha, Weber, and De Choud-



(a) (b) (c)

Figure 8: (a) Readability (b) Complexity (c) Repeatability scores for all set across different user categories.

hury 2018). For high frequency users, repeatability is observed to be higher as compared to low and moderate frequency users (fig. 8), and higher repeatability is indicative of low well-being.

Takeaway. Since 2020, high frequency users from India have shown an increase in readability, complexity and repeatability, indicating they form more complex sentences and repeat words as compared to low and moderate users.

Concluding Discussion

We gathered timeline information from 1.38 million posts for 890 self-harm users in India. We opted for an automated transformer-based BERT model to classify 1.38 million tweets as self-harm and non-self-harm. On the basis of the number of self-harm tweets, we suggest a system to categorise self-harm users into three categories: low, moderate, and high. We conducted a number of analysis to learn more about the mental health of users in different categories. According to our findings, high frequency users are the most active; they follow many individuals and, in contrast, obtain fewer followers. High frequency users have highest comment counts, and likes are not the preferred method of communication with them. High frequency users had the worst sleep patterns of all and tend to write long, complicated phrases with repeated words. High frequency users show evidence of poor mental health, this category of users require quick assistance. The results of this study can provide assistance in early healthcare intervention in cases of self-harm.

Implications

Rich literature on mental health has considered two categories of users - one who face anxiety and depression and the other who does not. It is important not to place the users in either of the two axes – positive or negative. People who post mentally afflicting content can have different levels of affliction, motivated via this, we proposed a method to classify self-harm users in different categories based on their posting frequencies. Analysis reveals the psychosocial well-being of users from different categories, and the longitudinal analysis informs about the changing patterns across various measures.

Our work has practical implications for preventing self-harm. We currently assign each user a fixed category, but this work can be modified to allow for dynamic categorisation, where users who appear to change their categories from low to moderate or from moderate to high require immediate attention and users who change their categories from moderate to low or high to moderate can be seen benefiting from social media.

Limitations and Future Work

We acknowledge the limitations of our work, some of which suggest interesting areas for additional investigation. Our analysis is probably affected by selection bias, we collect publicly available data from X, given the stigma associated with mental illness (Corrigan 2004), only a small fraction of people prefer to post their opinions openly on social media. Our dataset encompasses the timeframe from 2017 to 2022, which coincided with the COVID-19 pandemic. This global event profoundly affected individuals' lives, with many experiencing personal losses. As a result, the user behavior depicted in our findings may be influenced by this unique period. It is imperative to recognise that conducting a similar study during a different timeframe could yield varying behavioral outcomes.

While India boasts a rich linguistic diversity, our analysis is confined to English tweets. It's noteworthy that many Indians utilise *code-mixing* (such as, Hinglish) for their social media interactions, suggesting a promising avenue for exploring tweet analysis across diverse Indian languages. Furthermore, our analysis overlooks user demographics. Individuals of varying genders, ages, and locations encounter distinct challenges. Delving into user characteristics at a more granular level, rather than merely at the national level, presents an intriguing opportunity for future research.

Ethics Statement

We utilise publicly available data from Twitter and we commit to safeguard the privacy of individuals. In this paper, we prioritise anonymity by removing all personal identity information and rephrasing the Twitter posts included in this document to prevent identification. Nevertheless, we recognise the potential negative impact of this work. Reporting of user characteristics of different categories may lead to

high anxiety levels among other users who may experience similar pattern. Moreover, there exists a potential for misuse of the findings presented in this paper, particularly in the development of commercial tools aimed at detecting levels of depression. Such tools could be exploited in various domains, including targeted hiring and insurance premium calculation, thereby amplifying existing societal inequities.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback which helped in improving the quality of the paper. We also extend our appreciation to our colleagues Kausik Hira and Vijay Kumar Meena for assistance during the early stages of this work.

References

A., A.; Chaudhuri, R.; Hussain, Z.; and Chatterjee, S. 2022. Social media usage and its impact on users' mental health: a longitudinal study and inputs to policymakers. *International Journal of Law and Management*.

Abdel-Khalek, A. 2004. Can somatic symptoms predict depression? *Social Behavior and Personality: An international journal*.

Ageitos, E. C.; Fabregat, H.; Araujo, L.; and Martínez-Romo, J. 2021. NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features. In *CLEF*.

Aldhyani, T. H. H.; Alsubari, S. N.; Alshebami, A. S.; Alkahtani, H.; and Ahmed, Z. A. T. 2022. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*.

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *CHI*.

Andalibi, N.; Ozturk, P.; and Forte, A. 2017. Sensitive Self-Disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *CSCW*.

Barkur, G.; Vibha; and Kamath, G. B. 2020. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian journal of psychiatry*.

Basuroy, T. 2022a. India: Social Network penetration 2025.

Basuroy, T. 2022b. Number of Twitter users in India from 2013 to 2022.

BBC. 2019. Citizenship Amendment Bill: India's new 'anti-Muslim' law explained.

BBC. 2021. Farm laws: India farmers end protest after government accepts demands.

Chakraborty, A.; Sarkar, R.; Mrigen, A.; and Ganguly, N. 2017. Tabloids in the era of social media? understanding the production and consumption of clickbaits in twitter. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–21.

Chancellor, S.; Baumer, E. P. S.; and De Choudhury, M. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *CSCW*.

Chancellor, S.; Lin, Z.; Goodman, E. L.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *CSCW*.

Chatterjee, R. 2023. India's population passes 1.4 billion — and that's not a bad thing.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring Post Traumatic Stress Disorder in Twitter. *ICWSM*.

Coppersmith, G.; Leary, R.; Crutchley, P.; and Fine, A. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical informatics insights*.

Corrigan, P. 2004. How stigma interferes with mental health care. *The American psychologist*.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *CHI*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2021. Predicting Depression via Social Media. *ICWSM*.

De Choudhury, M.; and Kiciman, E. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk. *ICWSM*.

Devlin, J.; Chang, M.-W.; and Lee, K. 2019. Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Di Cara, N. H.; Maggio, V.; Davis, O. S. P.; and Haworth, C. M. A. 2023. Methodologies for Monitoring Mental Health on Twitter: Systematic Review. *Journal of medical Internet research*.

Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D. A.; and Schwartz, H. A. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*.

Ernala, S. K.; Labetoulle, T.; Bane, F.; Birnbaum, M. L.; Rizvi, A. F.; Kane, J. M.; and De Choudhury, M. 2018. Characterizing Audience Engagement and Assessing Its Impact on Social Media Disclosures of Mental Illnesses. *ICWSM*.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *CSCW*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*.

Ghoshal, D. 2021. Why India is facing an oxygen crisis as COVID cases mount.

Greene, A. K.; and Brownstone, L. M. 2023. "Just a place to keep track of myself": eating disorders, social media, and the quantified self. *Feminist Media Studies*.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Inkster, B.; Stillwell, D.; Kosinski, M.; and Jones, P. 2016. A decade into Facebook: where is psychiatry in the digital age? *Lancet Psychiatry*.

Insights. 2022. Issues related to minorities.

Khasawneh, A.; Chalil Madathil, K.; Dixon, E.; Wiśniewski, P.; Zinzow, H.; and Roth, R. 2020. Examining the Self-Harm and Suicide Contagion Effects of the Blue Whale Challenge on YouTube and Twitter: Qualitative Study. *JMIR Ment Health*.

Kosinski, M.; Matz, S. C.; Gosling, S. D.; Popov, V.; and Stillwell, D. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*.

Kumar, A.; and Nayar, K. R. 2021. COVID 19 and its mental health consequences. *Journal of Mental Health*.

Lathabhavan, R. 2020. People and social media platforms for positive mental health- A paradigm shift: A case on COVID-19 impact form India. *Asian journal of psychiatry*.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Van Alstyne, M. 2009. Life in the network: the coming age of computational social science. *Science*.

Lerman, K.; Yan, X.; and Wu, X.-Z. 2016. The "majority illusion" in social networks. *PLoS One*.

Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*.

LotusArise. 2022. Problems of Religious Minorities in India.

Lustberg, L.; and Reynolds, C. F. 2000. Depression and insomnia: questions of cause and effect. *Sleep medicine reviews*.

McCallum, D. R.; and Peterson, J. L. 1982. Computer-Based Readability Indexes. In *Proceedings of the ACM '82 Conference*. ACM.

Mozumder, S. 2023. The World Bank In India.

Myers, S. A.; Sharma, A.; Gupta, P.; and Lin, J. 2014. Information Network or Social Network? The Structure of the Twitter Follow Graph. In *WWW*.

Nations, U. 2021. The Dalit: Born into a life of discrimination and stigma.

Ophir, Y.; Tikochinski, R.; Asterhan, C. S. C.; Sisso, I.; and Reichart, R. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*.

Owen, D.; Antypas, D.; Hassoulas, A.; Pardiñas, A. F.; Espinosa-Anke, L.; and Collados, J. C. 2023. Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation. *JMIR AI*.

Pitler, E.; and Nenkova, A. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *EMNLP*.

Rane, A.; and Nadkarni, A. 2014. Suicide in India: a systematic review. *Shanghai archives of psychiatry*.

Robinson, J.; Cox, G.; Bailey, E.; Hetrick, S.; Rodrigues, M.; Fisher, S.; and Herrman, H. 2015. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*.

Roy, A.; Singh, A. K.; Mishra, S.; Chinnadurai, A.; Mitra, A.; and Bakshi, O. 2021. Mental health implications of COVID-19 pandemic and its response in India. *International Journal of Social Psychiatry*.

Saha, K.; Chan, L.; De Barbaro, K.; Abowd, G. D.; and De Choudhury, M. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*

Saha, K.; and Sharma, A. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *ICWSM*.

Saha, K.; Weber, I.; and De Choudhury, M. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations after Student Deaths on College Campuses. *ICWSM*.

Schemer, C.; Masur, P. K.; Geiß, S.; Müller, P.; and Schäfer, S. 2020. The Impact of Internet and Social Media Use on Well-Being: A Longitudinal Analysis of Adolescents Across Nine Years. *Journal of Computer-Mediated Communication*.

Thippaiah, S. M.; Nanjappa, M. S.; and Math, S. B. 2019. Suicide in India: A preventable epidemic. *The Indian journal of medical research*.

Twenge, J. M.; Cooper, A. B.; Joiner, T. E.; Duffy, M. E.; and Binau, S. G. 2019. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. *Journal of abnormal psychology*.

Varma, S. 2019. Govt. Report Reveals Shocking Condition of Workers in India.

Wang, T.; Brede, M.; Ianni, A.; and Mentzakis, E. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *WSDM*.

Wasserman, C.; Hoven, C. W.; Wasserman, D.; Carli, V.; Sarchiapone, M.; Al-Halabí, S.; Apter, A.; Balazs, J.; Bobes, J.; Cosman, D.; Farkas, L.; Feldman, D.; Fischer, G.; Graber, N.; Haring, C.; Herta, D. C.; Iosue, M.; Kahn, J.-P.; Keeley, H.; Klug, K.; McCarthy, J.; Tubiana-Potiez, A.; Varnik, A.; Varnik, P.; Ziberna, J.; and Poštuvan, V. 2012. Suicide prevention for youth–a mental health awareness program: lessons learned from the Saving and Empowering Young Lives in Europe (SEYLE) intervention study. *BMC Public Health*.

Yuan, Y.; Saha, K.; Keller, B.; Isometsä, E. T.; and Aledavood, T. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *TheWebConf*.

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? Yes
   (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? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (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? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
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)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (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)? NA
   (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? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (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 FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA, we collected publicly available data from X.
   (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? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA