# Quantifying and Understanding Gender and Cross-Cultural Differences in Mental Health Expression via Social Media

Munmun De Choudhury[†], Sanket Sharma[†],
Tomaz Logar[§], Wouter Eekhout[‡], René Nielsen[§]
[†] Georgia Tech, [§] United Nations Global Pulse, [‡] Leiden University

August 27, 2016

## 1 Introduction

For a long time, mental illness was treated as a gender and culture-neutral problem. However recent research shows that several social and psychological characteristics are unique to women [6], and hence there is a need to investigate in depth the causes and antecedents of emotional distress in women [49]. In developing countries, socio-political and economic issues such as poverty, inequalities in the workplace, societal and cultural expectations of women are often known to heighten women's risk to distress and mental illness [54].

However, the gender and cultural dimensions of mental health, especially in countries of the majority world, are less understood[1]. The United Nations' Millennium Development Goals[2], established in 2000, notes that the majority of data we have on global burden of mental illnesses comes from massive data aggregation exercises conducted only once every few years [38]. This data often does not include gender information; and often different countries have different policies and practices in place in terms of how mental illnesses are assessed [48].

Overlooking gender or culture based differences can have drastic consequences. This includes, misdiagnosis, misappropriation of interventions, constrained help-seeking along stereotypical lines, and a "one-size-fits-all" approach to extend help to those who may have unique needs [49, 48]. With the lack of an adequate global surveillance system, and with the accelerating pace of economic and cultural change, more frequently updated information on mental disorders for different gender and cultural subgroups is needed [39, 41], especially from subgroups who may not necessarily appear in formal statistical surveys of health and well-being.

In recent years, a new research direction has established social media data as a way to understand mental health challenges in people [16, 14, 26]. It has been found that these approaches can greatly supplement and complement traditional mental health assessments, such as the Behavioral Risk Factor Surveillance System (BFRSS) survey conducted by the Centers for Disease Control and Prevention (CDC) in the US [13, 12, 10, 18]. However, cultural and gendered expression of different subgroups may differ markedly from the "typical", largely western populations, on which these current social media investigations of mental health are based [16, 14]. Per the perspective laid out by *postcolonial computing*, rarely are analytical insights and intervention mechanisms able to be transplanted without modification from one subgroup to another, and still provide the same value [29]. Therefore, identifying gender based and cross-cultural context is critical in the use of such passively sensed big data for making sense of mental health and well-being.

[1] http://data2x.org/wp-content/uploads/2014/11/Data2X_MappingGenderDataGaps_FullReport.pdf
[2] http://www.un.org/millenniumgoals/

In this paper, we present a gender based and cross-cultural quantitative examination of mental health content shared on social media. Specifically, we study mental illness related expression made on Twitter, and analyze and characterize them along gender based and cross-cultural dimensions. Our contributions include:

- A machine learning approach to identify genuine self-disclosures of mental illness from noisy social media posts. We use vector-representations of content shared on external mental health support communities as weak labels to infer genuine social media disclosures in a semi-supervised manner. Based on psychology expert consultation, this method is found to yield 96% accuracy.
- Statistical comparisons between the expressions of female and male users, and by users from four prominent English speaking countries: western countries U.S., U.K., and majority world countries India, and South Africa. Specifically, we examine different linguistic style, affective, behavioral and cognitive attributes, and content characteristics based on a topic modeling approach.

For our work, we use a dataset of half a million Twitter users and nearly 1.5 million posts. Our findings reveal significant differences in how different gender and cultural subgroups, who disclose their mental health concerns, express themselves on Twitter, compared to equivalent control subgroups. We discuss how naturalistically generated and unobtrusively gathered social media data can help understand gender and culture based differences among individuals who disclose their mental health challenge. Our work also bears implications for new gender and culturally aware policy decisions, so as to bring tailored, timely help to individuals in need.

## 2 Related Work

### 2.1 Gaps in Gender and Cultural Dimensions of Mental Health

**Gender.** Prior work has identified significant differences in the health challenges experienced by gender subgroup populations [40]. Prior research has specifically identified the need for data collection efforts that can address largely unreported causes of women's excess disease burdens, and parse out the contributions of sex and gender, and their interaction, in the etiology, onset, progression and prevention of different health conditions [37]. Sexual violence and mental health issues among women are especially known to suffer from paucity of adequate data [47]. For a long time, clinicians have reported that women receive more services for mental illness in clinical settings than men [1]. There is therefore a need to understand the expression of mental illness in gender subgroups to identify those at greater risk.

Other works exploring gender-based dimensions of mental illness include the work of Kawachi and Berkman [31]. They reported that gender differences in support derived from social network participation partly account for the higher prevalence of psychological distress among women compared to men. In general, Astbury found that gender differences in mental disorders extend beyond differences in the rates of various disorders or their differential time of onset or course, and include factors that can affect susceptibility, disclosure, diagnosis, and adjustment to mental disorder [1].

**Culture.** Cultures vary in the extent to which expression of distress is socially sanctioned and reported. However, cross-cultural and cross-national studies of mental illnesses are limited [50]. Guillemin et al. [22] noted that cultural groups vary in disease expression and in their use of various health care systems. They go on to argue for the need to develop mental health and quality of life assessment measures specifically geared toward populations in non English-speaking countries. Similarly, Yeomans and Forman noted that many diagnostic methods and models have been derived from studies of samples from industrialized countries; however their application to diverse cultural populations needs attention to reliability and validity [56].

In majority world countries, it is noted that socio-political and economic issues, inequalities in the workplace, societal and cultural expectations are often known to impact individuals' risk to distress and mental illness. The 14-countrywide WHO Multinational study conducted in 1995 around the prevalence, nature and determinants of mental illnesses in general medical care settings provides some broad insights into some of these cross-national dimensions of mental health [52]. Patel et al. [41] reported that the startling finding of this study was that, despite the use of standardized methods in all country-specific centers, there were enormous variations in most variables known to be linked to mental health. Indeed, the only similarities across centers were the general observations of the ubiquity of mental illnesses, and the association of mental illness and disability after adjustment for physical disease severity. On the other hand, specific variables showed substantial variations; the prevalence rates of mental illness ranged from 7-52%, clinician recognition of mental illness varied from 5-60%, and the association of key variables such as gender, physical ill-health and education with mental illness were in opposite directions in different centers. These findings demonstrate the need for comparative cross-cultural and cross-national studies that can identify disclosure practices of mental illnesses, the local needs and thereby inform local policy and interventions [42].

We envision that leveraging social media data to understand gender and culture based differences in mental health disclosures can contribute toward closing these data gaps. More data driven capabilities of mental health inferencing and monitoring can also increase the visibility of these issues and provide an impetus for vulnerable subgroups to seek help and for clinicians to offer more treatment or intervention options.

### 2.2 Social Media and Health

Research in recent years has revealed that social media data, especially language and conversational patterns can be a powerful source of information toward understanding and detecting health challenges in individuals and populations. These efforts include utilizing social media to understand conditions and symptoms related to diseases [43], substance abuse [36, 34], postpartum depression [14], eating disorders [7], and other mental health disorders [16, 10, 51].

We note that gender and cross-cultural examinations of health states and health behaviors using social media are limited. UNICEF recently undertook a study of social media content in Eastern Europe to look at attitudes towards vaccination [53]. De Choudhury et al. [16] identified differences between experiences of depression between women and men as measured from Twitter. Andalibi et al. [2] found that men are significantly more likely to adopt anonymous social media identities when engaging in sexual abuse related self-disclosure. Tsugawa et al. [51] examined how individuals with non-English speaking backgrounds expressed depressive thoughts and emotions on Twitter, and thereafter built classification models to identify markers of depression among Japanese speaking users. In other work, Ramirez-Esparza et al. [45] analyzed linguistic attributes of English and Spanish posts shared in depression forums. They found that depressed people who wrote in Spanish were more likely to mention relational concerns than depressed people who wrote in English, who were more likely to mention medicinal concerns.

Our work presents an in-depth large-scale data-driven study of individuals who choose to express their mental illness on Twitter. We examine gender based and cross-cultural differences in attributes of these expressions in four predominantly English speaking countries.

## 3 Data

### 3.1 Twitter Data

Our study uses publicly shared data collected from the social media Twitter. We started by obtaining a large sample of English language candidate self-disclosure posts from Twitter's Firehose stream around a variety of mental health concerns. Specifically, we filtered the Twitter posts shared in March 2015 containing any of the keyphrases included in Table 1. These keyphrases were collated by a combination of reference to prior work [10, 11], and consultation with a trained psychiatrist practitioner. Through these keyphrases, we sought to identify individuals who publicly state that they have been diagnosed with, or suffering from some form of mental illness. As noted by Coppersmith et al., users may make such a statement to seek support from others in their Twitter social network, to fight the stigma of mental illness, or perhaps as an explanation of some of their behavior [10]. We obtained 1,319,064 posts from 534,829 unique users at the end of this initial data collection phase.

| | | |
|---|---|---|
| i want to die | i want to end my life | i want to suicide |
| i tried to suicide | i [*] thinking of suicide | i thought of suicide |
| i am depressed | i [*] diagnosed [*] depression | i [have/had] mental illness |
| i [*] diagnosed [*] mental illness | i attempted suicide | i [have/had] depression |
| killing myself | ending my life | |

Table 1: Keyphrases used for obtaining candidate mental health disclosure posts from Twitter.

Parallelly, we obtained a candidate control data sample from Twitter's Firehose stream, so as to allow robust statistical comparisons between Twitter users who choose to express their mental illness, and those who do not. This dataset included a random sample of 1,513,279 posts from 673,898 unique users made on Twitter during March 2015, ensuring that none of these posts matched any of the keyphrases given in Table 1.

Thereafter, for both of the candidate mental health disclosure sample and the candidate control sample of posts, we utilized Twitter's official API[3] to obtain the last 3,200 posts for each of the unique users in both datasets. For the control dataset, if any of the users had any post in their crawled posts matching one of the keyphrases above, we disregard them from our analysis. Further, employing the Google Compact Language Detector[4], we disregarded any users, if at least 75% of their posts were not written in English. Our final candidate disclosure sample contained 51,038,914 posts from 470,337 users ($\mu = 108.5$), while the control sample contained 66,214,850 posts from 480,685 users ($\mu = 137.7$).

### 3.2 Gender and Country Inference

In order to allow gender based and cross-cultural comparisons of the above data, we now present an automated gender and country name inference method — Twitter does not allow individuals an ability to self-report their gender, whereas location information is known to be highly noisy [24]. We also note the need for such a method given the size of our two datasets, that makes human coding of gender and country names challenging from a practical perspective.

**Gender Inference.** For inferring gender of a user in our two datasets, we utilized the self-reported name string shared by the users in our datasets in their Twitter profile's name field. Thereafter, we performed a 1-gram lookup of these strings in a large names database complied from Latin script sources (English, French, Spanish, Italian) as well as that released by the US Census Bureau. For analytical simplicity, we only consider binary gender (female/male) in this paper.

[3]https://dev.twitter.com/overview/documentation

[4]https://code.google.com/p/cld2/

**Country Name Inference.** For inferring the country name corresponding to a user, we adopted a stepwise approach as follows: First, we cleaned the location strings reported in the location field of Twitter user profiles — including normalization of character case and removal of non-English word roots. Then we performed a location matching exercise, wherein we split the cleaned location strings into single words, iteratively created all possible 5-to-1-gram substrings, and then matched each substring to a location database curated from Wikipedia and travel resources. Third, we performed disambiguation by computing geographic distances between matched text-adjacent places and assigned high likelihood to those matches that are close to each other geographically. For instance, "Viana, MA" matches Brazil, not US, because distance between city Viana and state of Maranhao is less than between city Viana and state Massachusetts. We then sorted equally likely location alternatives by population size and choose the top one. We note that, compared to geo-located Twitter posts, this location field string lookup method has been known to yield better coverage in social media data [24]. We also note that while more sophisticated machine learning approaches have been used in the past to infer a user's "home" location [19] or hyper-local location information, since our interest is in country name inference, we deemed our approach to be adequate in reliably capturing this course location information.

**Validation.** Finally, we validated both our gender and country name inference methods based on annotations obtained from two independent raters on a sample of 100 users. We found agreement between the raters' annotations and the one given by our method for 79% of the cases for gender, and 86% of the cases for country names ($\kappa = .77$). In our candidate disclosure and candidate control datasets, we were able to infer binary gender for 325,873 candidate mental health (of which 59% were female) and 439,224 candidate control users (of which 46% were female) respectively, and country information for 131,890 and 328,468 users for the same two groups respectively.

Within the scope of this paper, we restrict our attention to four most populated countries where English is a predominant form of expression — western countries US, UK (GB), and majority world countries India (IN), and South Africa (ZA). All of these countries fare in the 25 countries with most population[5]. Focusing on these countries provides us with a lens to examine cross-cultural differences in the disclosures of mental illnesses on social media.

**Handling Population and Internet Penetration Bias.** We note that the populations of the four countries are widely different, along with their overall reported internet penetration rates[6]. Hence we devise a subsampling strategy to filter users belonging to one of the four countries, from the set of users in the candidate disclosure and the control datasets with inferrable country information. First, for population based subsampling, we use the inverse of the population ranks of the countries[4] as the respective rates of sampling. Then we use the internet penetration percentages of the countries[5] to randomly sample that fraction of users from the population subsampled sets. In this manner, across both the datasets, we obtained 211,132 Twitter users from the US, 61,816 users from the UK, 10,808 from IN, and 5,769 from ZA.

## 4 Methods

### 4.1 Obtaining Genuine Mental Health Expressions

We note that the candidate mental health disclosure data sample is prone to significant noise. It is possible that although a user uses one of the mental health keyphrases in their Twitter post, it may not indicate a genuine disclosure (e.g., "when I have to wake up at 6am, I feel like killing myself" does not indicate a person's real intention of taking their life); similarly, other keyphrases

[5] https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)

[6] http://www.internetlivestats.com/internet-users-by-country/

may be used in a sarcastic, humorous context or as a flippant reference).

To eliminate such users, we adopt a semi-supervised machine learning method [57] in which we compare the language of each user in our Twitter dataset, with the language of a self-identified set of social media users suffering a from mental illness. For the purpose, we obtain a large sample of 79,833 posts from 44,262 users made on the Reddit sub-communities r/depression, r/mentalhealth, and r/SuicideWatch between February and November 2014. We use this dataset a weak signal of the language used by individuals identifying with a form of mental illness. Prior work has also indicated that the disclosures made on these forums are genuine disclosures of mental illnesses, and have also been validated through consultation with a psychiatrist [17].

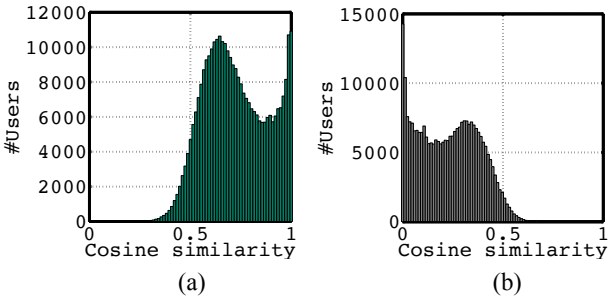


Figure 1: Distribution of the number of Twitter users over their cosine similarties with Reddit mental health content. The distributions are shown for both the candidate mental health disclosure as well as the candidate control sample.

Our approach proceeds as follows:

- *Step 1.* We create Twitter user-centric vector-representations by collating all of their posts – this would give us as many vectors as the number of users in the candidate disclosure sample. We also similarly create a single vector representation by collating all posts in our Reddit dataset.
- *Step 2.* Next we establish comparative validity across the Twitter vectors and the Reddit vector. This is an important step because the language of Twitter and Reddit cannot be directly compared due to the unique affordances of each site, and the seeming differences in the demographics of the two[7]. For each of these vectors, we perform linguistic normalization of tokenized items in them[8]. Then we compute the average automated readability index (ARI [46]) on all of the normalized Twitter vectors, and the Reddit vector. We observe the ARI differential between the two to be $\sim$6.7%[9], indicating that the languages are close to standardized internet speak [27], and hence comparable following normalization.
- *Step 3.* Next we build $n$-gram language models ($n = 3$) for both the normalized Twitter vectors, and the single Reddit vector. Language models are commonly used to estimate how likely a given sequence of words is. We then determine cosine similarity scores between the language models of each normalized Twitter vector and the Reddit vector.
- *Step 4.* Finally, we obtain the distribution of cosine similarities over all normalized Twitter vectors (see Figure 1(a)). We construct the "genuine disclosure dataset" of users to be those vectors (of users) for whom the cosine similarity of their language models with that of Reddit's is *greater than or equal to* the median similarity across all vectors (median distance=.71 $\sigma = .157$). Our final dataset of genuine mental illness disclosures consisted of 231,611 users; we will refer to this set as the `MID` users.

[7]Twitter posts are only 140 characters long, whereas Reddit posts tend to be more detailed and longer; Twitter posts often contain highly irregular syntax and nonstandard use of English.

[8]For normalization, we perform syntactic disambiguation of the Twitter vectors based on PoS (part-of-speech) tagging [30]. This method converts non-standard forms into dictionary forms, expands abbreviations, handles misspellings, punctuation/omission errors, phonetic spelling and intentional misspelling for verbal effect: "rite" to "right", "soooorryyy" to "sorry".

[9]ARI for the normalized Twitter posts was 15.1, whereas that for Reddit was 16.2.

| **Genuine Disclosures** |
|---|
| I feel my mental state slowly declining. *I have been depressed.* |
| I'm in a dark place today. It's strange, after admitting that *I am depressed* I can recognize the slippery slope that I'm on earlier. |
| All I want to do is die, *I tried suicide* six times and couldn't even do it right. |
| I am open about most things, the latest is that *I have Mental Illness* and this has made my life extremely difficult to live and be happy |
| **Noisy Disclosures** |
| My headphones broke so now *I'm ending my life.* Bye |
| The kindergarten class I'm in charge of today doesn't know what Little Sally Walker is..... *I am depressed* |
| lost my headphones brb *killing myself* |
| ok im literally *killing myself* bc my global warming class fuck this final paper |

Table 2: Example (paraphrased) genuine and noisy mental health disclosure users from Twitter, obtained using our semi-supervised learning approach.

**Expert Verification of Mental Health Disclosures.** Next we qualitatively verified whether posts from these users indeed were self-disclosures of mental health challenges. For the purpose, we consulted a licensed psychologist, and also included two researchers, who were familiar with mental health content shared on social media. Over a random sample of 100 posts complied from the timeline of 100 randomly selected `MID` users, we obtained independent binary annotations on whether a post was likely to be related to mental health. The Fleiss' $\kappa$ for interrater agreement was found to be high, .87, along with an accuracy of 96% in distinguishing users who engage in genuine disclosures from those who do not. This establishes adequacy of our approach. Table 2 gives some paraphrased mental health disclosure posts of Twitter users who were identified to be genuine mental health disclosers by our approach.

In Figure 2 we show the pipeline of steps involved in our approach.

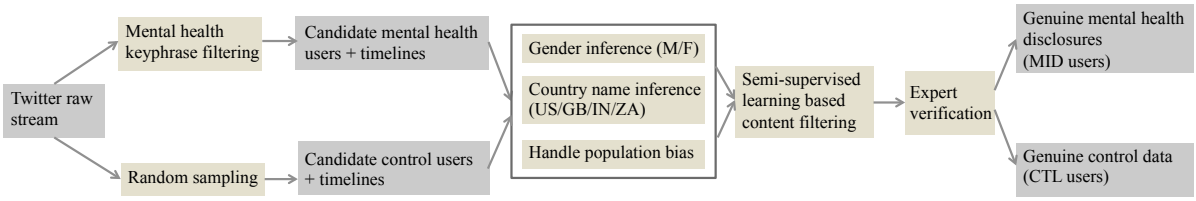


Figure 2: Flow chart of the steps involved in obtaining the `MID` (mental health disclosure) and `CTL` (control) user cohorts starting from the raw stream of Twitter data.

## 4.2 Developing an Accurate Control Dataset

Next we also note that it is possible that our candidate control dataset includes users who engage in mental illness in their posts, however did not use any of the keyphrases from Table 1. To eliminate such users, like above, we compare the language used in the Twitter posts of these users with that of the Reddit mental health posts. However in this case, we are interested in the users whose language is most distinct (or least similar) from that used in the Reddit content. Hence our final control dataset is obtained by filtering for those users whose language model based similarity is *less than* the median similarity across all control user vectors (238,860 users; median distance=.38; $\sigma = .125$). See Figure 1(b) for a distribution of the cosine cosine similarities. We will refer to this dataset as `CTL` users (ref. Figure 2 for the approach).

## 4.3 Quantifying Differences in Disclosures

In this subsection, we present methods to quantify the differences between the disclosure characteristics of females and males, and individuals reported to be from one of the four countries of

interest: US, GB, IN, and ZA in the MID dataset.

#### 4.3.1 Linguistic Measures

Language is a powerful source of expression [28]. A rich body of work, such as Boroditsky et al. [5] showed how the perception of objects in different languages can relate to as well as impact one's social and pscyhological status. It is recognized that language, specifically one's native language, shapes and drives one's thoughts, actions, and social relationships [9]. Further, it is established that cross-cultural and sex differences exist in one's underlying thought processes [23, 55]: according to Kövecses [32], cultural models are known to define one's emotional concepts.

To quantify gender and cross-cultural dimensions in the language of individuals who engage in mental health disclosure on social media, we propose three categories of measures: (1) **affective attributes**, (2) **cognitive attributes**, and (3) **linguistic style attributes**. Measures belonging to all of these attribute categories are largely based on the psycholinguistic lexicon LIWC [44], and were motivated from prior literature that examines associations between the behavioral expression of individuals and their psychological distress, including vulnerability to mental illness [8, 15].
(1) We consider two measures of affect derived from LIWC: positive affect (PA), and negative affect (NA), and four other measures of emotional expression: *anger*, *anxiety*, *sadness*, and *swear*.
(2) We use LIWC to define the cognitive measures as well: (a) cognition, comprising *cognitive mech*, *discrepancies*, *inhibition*, *negation*, *death*, *causation*, *certainty*, and *tentativeness*; and (b) perception, comprising set of words in LIWC around *see*, *hear*, *feel*, *percept*, *insight*, and *relative*. Quantifying one's cognition and perception, as manifested linguistically, can provide insights into emotional stability and cognitive complexity—these attributes are important with regard to understanding one's mental well-being [21].
(3) Next, we consider four measures of linguistic style: (a) **Lexical Density**: consisting of words that are *verbs*, *auxiliary verbs*, *nouns*, *adjectives* (identified using NLTK's [3] POS tagger), and *adverbs*. (b) **Temporal References**: consisting of *past*, *present*, and *future* tenses. (c) **Social/Personal Concerns**: words belonging to *family*, *friends*, *social*, *work*, *health*, *humans*, *religion*, *bio*, *body*, *money*, *achievement*, *home*, and *sexual*. (d) **Interpersonal Awareness and Focus**: words that are *1st person singular*, *1st person plural*, *2nd person*, and *3rd person* pronouns. Together, linguistic styles are known to indicate one's underlying psychological processes (lexical density), personality (temporal references), social support and connectivity (social/personal concerns), and awareness of one's surroundings and environment (interpersonal focus). Prior work identifies all of these cues to be valuable in understanding mental health, in both offline and online contexts, including social media [45].

#### 4.3.2 Topic Modeling

Our second method for comparing mental illness disclosures uses a topic model, which have been commonly employed to analyze health data [43]. We obtain topics by running Latent Dirichlet Allocation (LDA) [4] over all posts. We pre-processed the data by removing a standard list of Twitter-specific stop words, words with very high frequency ($> 0.25\times$ datasize), and words that occur fewer than five times. Thereafter, we used Gensim's implementation of online LDA from [25]. We used the default hyper-parameter settings and 100 topics, which we determined based on the value of average corpus likelihood over ten runs.

To measure topic differences in one cohort (e.g., IN MID users) over the other (e.g., UK MID users), we first compute the posterior probability of each topic separately for all posts in both cohorts. We then compute three comparison metrics: (1) the rate of change for each topic, given as the difference between the posterior topic probabilities of the cohorts, divided by the probability of the first cohort; (2) the pointwise mutual information between the posterior topic probabilities of

the same cohorts; and (3) the Spearman's rank correlation between the topic distributions for the two cohorts. Additionally, we compare all gender and culture cohorts based on significance tests (e.g., Mann Whitney U test for gender, and the Kruskal Wallis test for cultural differences).

We also present a method to qualitatively examine the differences between the topics used by different MID user cohorts. For the purpose, two researchers familiar with mental health content on social media independently inspected the words associated with each of the topics given by the above topic model. They used a semi-open coding approach to develop a codebook and extracted descriptive topical themes for the topics (Cohen's $\kappa$=.74). During the codebook development, the two annotators referred to prior literature on gender and cultural differences in mental health [47, 22, 52]. In the results section, we will present an examination of these qualitative differences.

# 5 Results

## 5.1 Gender Differences

### 5.1.1 Linguistic Differences

Based on Table 3, we observe considerable differences, in terms of the linguistic measures, in the Twitter posts of female and male MID users.

**Affective Attributes (Gender).** Starting with the measures of affect, female MID users show higher sadness ($z = -39.0$; 15.4% higher) and anxiety ($z = -26.4$; 10.7% higher). Prior literature indicates expression of these emotions to be associated with depression symptoms such as mental instability and helplessness, loneliness, and restlessness [14].

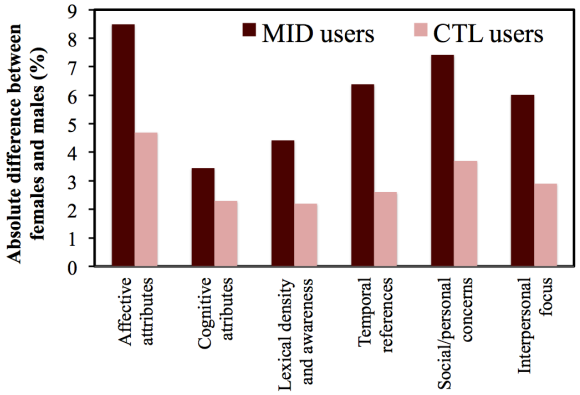


Figure 3: Mean absolute differences between female and male MID and CTL users per the various categories of linguistic measures. Difference for a specific measure is calculated as the ratio of the difference between the values of the measure for females and males, to the value of the measure among males.

> "My health has been defining me lately. #depression has invaded my peace and #anxiety has exhausted my thoughts. Pain isn't always physical." (↑ female)

> "Why am I even here ... No one needs or wants me ... I'm useless" (↑ female)

Whereas we find that male MID users express more NA ($z = 9.8$; 2.6% higher), anger ($z = 15.8$; 5.3% higher), and use more swear language ($z = 22.5$; 9.5% higher) in their posts.

> "The past week has been horrible. Depression is robbing me of the peace I have felt. I'm isolated in a house full of people." (↑ female)

> "Honestly fuck everyone, y'all gonna miss me when I'm gone" (↑ male)

Interestingly, female users also tend to use more PA in their shared content ($z = -33.9$; 7.1% higher), perhaps as a way to demonstrate a positive outlook publicly, despite the mental health challenge that they might be experiencing.

> "There's something about those eyes helping to wake me up everyday that makes the days brighter." (↑ female)

We also report the extent to which the female and male CTL users differ, along all of these affective attributes. Per Figure 3, this mean difference is found to be only 4.7%, which is considerably low compared to 8.5% in the case of the MID cohort.

**Cognitive Attributes (Gender).** Next, female MID users show lowered cognition and perception, in other words, greater cognitive impairment in their Twitter posts compared to male MID users. For instance, inhibition ($z = 29.9$) is lower in females relative to males by 6.5%. Through lower usage of certainty words, female peers tend to demonstrate heightened emotional instability ($z = 4.7$).

> "So now here I am, confused and full of questions. Am I born to lose or is this just a lesson?" (↑ male)

The female and male users in the CTL cohort, however, differ by only 2.3% across the various cognition and perception attributes, as shown in Figure 3.

**Lexical Density and Awareness (Gender).** Next, lexical density of the social media content of female MID users is higher compared to their male peers, as observed through the usage of prepositions ($z = -18.4$; 6.4% higher), conjunctions ($z = -4.6$; 1.6% higher), adverbs ($z = -18.1$; 3.0% higher), inclusive ($z = -7.1$; 2.3% higher), and exclusive ($z = -16.9$; 4.9% higher). However, compared to the male MID users, they show lowered awareness of objects and their surroundings, as measured via the proportion of verbs ($z = 21.5$; 1.6% lower), auxiliary verbs ($z = 13.1$; 1.9% lower), articles ($z = 37.6$; 13.4% lower) in Twitter posts. Note that, in the case of the female and male users in the CTL cohort, per Figure 3, the mean difference across all of these measures is observed to be only 2.2%. This indicates that the female and male MID users show differences beyond that accounted for in the control sample.

| | $\mu$ (male) | $\mu$ (female) | $z$ | $p$ |
|---|---|---|---|---|
| **Affective attributes** | | | | |
| PA | 0.0647 | 0.0693 | -33.928 | *** |
| NA | 0.0148 | 0.0144 | 9.858 | ** |
| anger | 0.0215 | 0.0203 | 15.879 | *** |
| anxiety | 0.0041 | 0.0046 | -26.451 | *** |
| sadness | 0.0077 | 0.0089 | -39.002 | *** |
| swear | 0.0144 | 0.0130 | 22.562 | *** |
| **Cognitive attributes** | | | | |
| Cognition | | | | |
| cognitive_mech | 0.1363 | 0.1338 | 19.832 | *** |
| inhibition | 0.0407 | 0.0380 | 29.935 | *** |
| causation | 0.0147 | 0.0144 | 9.553 | ** |
| tentativeness | 0.0221 | 0.0205 | 42.423 | *** |
| Perception | | | | |
| see | 0.0140 | 0.0134 | 17.609 | *** |
| hear | 0.0087 | 0.0085 | 8.042 | ** |
| feel | 0.0092 | 0.0087 | 17.496 | *** |
| percept | 0.0337 | 0.0326 | 14.963 | *** |
| relative | 0.0183 | 0.0180 | 8.797 | ** |
| **Lexical Density and Awareness** | | | | |
| verbs | 0.1553 | 0.1528 | 21.155 | *** |
| auxiliary verbs | 0.0710 | 0.0696 | 13.084 | ** |
| articles | 0.0320 | 0.0277 | 37.686 | *** |
| prepositions | 0.0642 | 0.0684 | -18.451 | *** |
| adverbs | 0.0483 | 0.0498 | -18.163 | *** |
| inclusive | 0.0218 | 0.0223 | -7.161 | ** |
| exclusive | 0.0198 | 0.0208 | -16.967 | *** |
| **Temporal references** | | | | |
| present tense | 0.1077 | 0.1114 | -33.391 | *** |
| future tense | 0.0071 | 0.0064 | 35.784 | *** |
| **Social/Personal Concerns** | | | | |
| family | 0.0048 | 0.0054 | -23.292 | *** |
| friends | 0.0032 | 0.0037 | -27.404 | *** |
| health | 0.0092 | 0.0097 | -18.017 | *** |
| religion | 0.0042 | 0.0039 | 10.615 | ** |
| bio | 0.0429 | 0.0452 | -21.303 | *** |
| body | 0.0150 | 0.0155 | -8.910 | ** |
| achievement | 0.0157 | 0.0170 | -32.293 | *** |
| home | 0.0047 | 0.0055 | -36.581 | *** |
| sexual | 0.0156 | 0.0148 | 14.940 | *** |
| **Interpersonal focus** | | | | |
| 1st p. singular | 0.0489 | 0.0443 | 26.127 | *** |
| 1st p. plural | 0.0035 | 0.0038 | -16.785 | *** |
| 2nd p. | 0.0185 | 0.0199 | -8.021 | ** |
| 3rd p. | 0.0069 | 0.0071 | -6.974 | ** |

Table 3: Differences between posts from female and male MID users based on linguistic measures. Statistical significance is reported following Bonferroni correction.

**Temporal References (Gender).** Female MID users tend to be more focused on the here and now, due to their greater use of present tense words ($z = -33.3$; 3.4% higher). On the other hand, male MID users show a greater future orientation compared to the females, via the usage of more future tense in their Twitter language ($z = 35.7$; 9.2% higher). Such differences in temporal references

in language are however not observable between the female and male CTL cohort. Compared to a mean difference of 6.4% in the case of the MID cohort, it is only 2.6% for the CTL users.

**Social/Personal Concerns (Gender).** There are a variety of differences in the social and personal concerns that manifest in the Twitter posts of female and male MID users. First, male MID users display lower sense of achievement ($z = -32.2$; 8.1% lower)—a known sign of lowered self-esteem [7]. On the other hand, female MID users express greater concern about health ($z = -18.0$; 6.0% higher) and body ($z = -8.9$; 2.7% higher) compared to their male peers. This might indicate their greater self-awareness of their wellness status or perceptions of their physical health.

> "Over the past 2 years I have been hit with physical and mental pain. The pain is real. It is still there." (↑ female)

> "My stomach sinks everytime." (↑ female)

An interesting finding here is the observation that male MID users exhibit lower use of social ($z = -4.7$), friends ($z = -27.4$; family ($z = -23.2$; and bio words ($z = -21.3$). This may imply that these users are less socially concerned or bothered. By the same token, the female peers might be using such language more extensively in their Twitter posts in order to explicitly seek help from their social networks or to feel supported.

> "Hard to really feel sick with this support group. #Family" (↑ female)

> "I miss having someone, a friend to talk to all night" (↑ female)

Finally, male MID users show a greater interest in religious discussions compared to females ($z = 10.6$; 5.1% higher).

> "God is working things out for you, even when you don't feel it. Have faith and be thankful."

While analyzing the social and personal concerns expressed by the female and male users in the CTL cohort, based on Figure 3 we do not observe such extensive differences: the mean difference between the cohorts per these measures is only 3.7%.

**Interpersonal Focus (Gender).** Increased use of first person singular pronouns ($z = 26.1$; 10.2% higher) in the posts of male MID users shows their self-focus and disclosure of personal stories.

> "I decided yesterday evening to go back to this thinking place. there was no one else there, just me. I had let my parents know where I was" (↑ male)

Additionally, lower use of second person pronouns ($z = -8.0$; 3.0% lower) and third person pronouns ($z = -6.9$; 3.4% lower) in the content of the male users tell us that they tend to be less interactive, and engage in lesser discourse about others. For the female and male users in the CTL cohort (Figure 3), the interpersonal focus measures account for only 2.9% of the difference.

### 5.1.2 Topical Differences

Digging deeper, between the female and male gender MID cohorts on Twitter, we find significant differences in topics of the posts. Specifically, we find that, per the three topic comparison metrics, the Spearman rank correlation for topical comparison between females and males is $\rho = .31$ ($p < .01$ based on a Mann Whitney U test), the pointwise mutual information (PMI) is .38 ($p < .01$ based on a Mann Whitney U test), while the mean percent difference in the likelihoods of the 100 topics is 59.3% ($p < .001$ based on a Mann Whitney U test). However, the respective metrics between

males and females in the CTL cohort are $\rho$ =.14, PMI=.61, and a 33.6% difference in the topic likelihoods. This reveals significant topical differences for the MID cohort, above and beyond the differences that exist due to underlying gender based latent factors.

We now present some insights we gained by examining the posts associated with the topics that distinguish the female and male cohorts most distinctively. We first identify the two topics that are more likely to be prevalent in posts from male MID users than female. The posts associated with topic #13, for instance, express contemplative negative thoughts and hopelessness (e.g., 'never', 'low', 'mess', 'lonely', 'drunk'):

> "Sometimes I wonder if **anyone still** looks out for me. I am a **mess** that **nobody** wants to clean up. I'm a **wreck**"

> "Some things are better left **unsaid**. **Lonely** nights make for **long** nights. Being **drunk** sometimes makes it **easier**."

The second topic more prevalent among male MID users is topic #57, and it shows detachment from the social realm and hesitation to seek help (e.g., 'invisible', 'help', 'ask', 'relationship'):

> "If I were going to **kill** myself, I wouldn't **tell** anyone. If I'm already **invisible**, why see me to favor your own self righteousness?"

> "I love my **relationship** but it's fun to wonder who was just too shy or too **afraid** to ask me out. I have **suspicions** but I'm self centered."

Contrastively, we examine the top two topics that manifest more extensively in posts from female MID users, compared to the males. Posts associated with topic #86 indicate the presence of a positive outlook, motivational spirit to cope with mental health challenges, and a desire for disclosure and help seeking (e.g., 'fear', 'rejection', 'bury', 'lose', 'stay', 'say', 'okay'):

> "you're afraid to tell **people** how you feel because you **fear rejection**, so you **bury** it deep inside yourself where it only **destroys** you more."

> "Sadly, you can't **lose** what you never had, you can't keep what's not yours, and you can't **hold** on to something that doesn't want to **stay**."

Finally, through topic #45, the topic with the second highest likelihood of prevalence in female MID users' posts over males, the female MID users share personal experiences around mental illness, including self-assessments and self-realization (e.g., 'pain', 'control', 'realization', 'reminder'):

> "I used to hurt **myself**, because it was the only **pain** I could **control**."

> "Daily **reminder** that life has **taught** me: don't waste your time **worrying** about the **people** who don't like you, you don't live to please anyone."

## 5.2 Cross-Cultural Differences

### 5.2.1 Linguistic Differences

The second part of our empirical investigation focused on characterizing the usage of different linguistic measures across the four cultures: western countries US, GB, and majority world countries IN and ZA. In Table 4, we report the mean values of each statistically significant linguistic measure in each of the four cultures, as well as the results of Kruskal-Wallis significance tests comparing them. For the purposes of the ensuing discussion, we performed post hoc multiple comparison Wilcoxon tests to report which specific pairs of cultures differed most in terms of the different linguistic measures. Further, we discuss our results by comparing the majority world cultures IN and ZA, with the western cultures US and GB. The linguistic differences between these two cohorts in the CTL group are shown in Figure 4.

**Affective Attributes (Culture).** First, MID users from IN and ZA express relatively higher PA ($H = 104.3$ and lower NA ($H = 17.5$), anger ($H = 127.9$), anxiety ($H = 87.9$) and sadness ($H = 79.6$) compared to those from US and GB. In other words, users from the latter two cultures, US and GB, tend to express emotional content that is more emotionally outspoken in comparison with the majority world cultures (11.6% higher). We conjecture those in the majority world cohort might be expressing more positivity as a coping mechanism or as a face-saving mechanism [15]. This difference is also considerably larger than the affective differences observed for the cultures in the CTL cohort (6.8%); see Figure 4.

> "I hate myself with a burning passion right now, I should be out on the sled" (↑ US, GB)
>
> "I've been crying all afternoon I feel like a bag of shit" (↑ US, GB)
>
> "Being unhappy is the only thing I do right." (↑ US, GB)

| | $\mu$ (US) | $\mu$ (GB) | $\mu$ (IN) | $\mu$ (ZA) | $H$ | $p$ |
|---|---|---|---|---|---|---|
| **Affective attributes** | | | | | | |
| PA | 0.062 | 0.063 | 0.070 | 0.066 | 104.3 | ** |
| anger | 0.020 | 0.019 | 0.018 | 0.016 | 127.9 | *** |
| anxiety | 0.004 | 0.005 | 0.004 | 0.004 | 87.9 | ** |
| sadness | 0.008 | 0.009 | 0.007 | 0.007 | 79.6 | ** |
| swear | 0.014 | 0.012 | 0.010 | 0.009 | 239.7 | *** |
| **Cognitive attributes** | | | | | | |
| **Cognition** | | | | | | |
| cognitive_mech | 0.135 | 0.133 | 0.142 | 0.139 | 150.8 | *** |
| negation | 0.023 | 0.023 | 0.025 | 0.026 | 87.2 | ** |
| tentativeness | 0.021 | 0.021 | 0.023 | 0.021 | 142.0 | *** |
| **Perception** | | | | | | |
| see | 0.013 | 0.012 | 0.014 | 0.014 | 69.9 | ** |
| hear | 0.009 | 0.008 | 0.008 | 0.008 | 314.3 | *** |
| feel | 0.008 | 0.009 | 0.009 | 0.009 | 52.7 | ** |
| percept | 0.030 | 0.032 | 0.032 | 0.034 | 120.1 | *** |
| **Lexical Density and Awareness** | | | | | | |
| auxiliary verbs | 0.069 | 0.070 | 0.072 | 0.074 | 215.2 | *** |
| articles | 0.030 | 0.030 | 0.033 | 0.034 | 186.7 | *** |
| prepositions | 0.067 | 0.066 | 0.072 | 0.072 | 210.6 | *** |
| adverbs | 0.048 | 0.051 | 0.046 | 0.049 | 208.8 | *** |
| **Temporal references** | | | | | | |
| past tense | 0.030 | 0.031 | 0.025 | 0.025 | 369.9 | *** |
| future tense | 0.007 | 0.007 | 0.008 | 0.008 | 103.8 | ** |
| **Social/Personal Concerns** | | | | | | |
| family | 0.005 | 0.005 | 0.004 | 0.004 | 154.1 | *** |
| social | 0.101 | 0.101 | 0.093 | 0.082 | 1613.5 | *** |
| health | 0.010 | 0.010 | 0.009 | 0.009 | 124.5 | *** |
| religion | 0.004 | 0.005 | 0.004 | 0.004 | 162.3 | *** |
| body | 0.015 | 0.015 | 0.015 | 0.013 | 56.6 | ** |
| work | 0.007 | 0.008 | 0.006 | 0.007 | 281.9 | *** |
| home | 0.006 | 0.006 | 0.004 | 0.005 | 158.2 | *** |
| sexual | 0.014 | 0.015 | 0.013 | 0.012 | 131.0 | *** |
| **Interpersonal focus** | | | | | | |
| 1st p. singular | 0.038 | 0.041 | 0.046 | 0.041 | 128.0 | *** |
| 2nd p. | 0.023 | 0.023 | 0.018 | 0.015 | 290.1 | *** |
| 3rd p. | 0.007 | 0.007 | 0.007 | 0.006 | 49.9 | ** |

Table 4: Differences between posts of MID users from the four cultures: US, GB, IN and ZA, per different linguistic measures. Statistical significance is reported following Bonferroni correction.

**Cognitive Attributes (Culture).** We also observe greater manifestation of cognitive impairment and emotional instability among MID users from US and GB compared to those from IN and ZA. Mean difference between the groups is 6.3%, while the same in the case of the CTL cohort is only 2.9%. This is observable especially from the lower use of measures of cognitive_mech ($H = 150.8$), certainty ($H = 41.2$), discrepancies ($H = 25.5$), and percept ($H = 120.1$).

> "you think youre doing okay, then suddenly its a nighttime and you feel alone and youre not sure how to distract yourself anymore" (↑ IN, ZA)

**Lexical Density and Awareness (Culture).** The lexical density measures show lower values in the posts of US and GB MID users, relative to those in the case of IN and ZA users. That is, the former group uses fewer prepositions ($H = 210.6$), adverbs ($H = 208.8$), and exclusive words ($H = 38.4$). Additionally, they also show reduced awareness of their social and environmental context, as measured via the usage of articles ($H = 186.7$). We note this to be more than twice relative to the differences in terms of lexical density and awareness observed in the two groups within the CTL cohort (ref. Figure 4).

**Temporal References (Culture).** Next, the MID users from IN and ZA show an increased future orientation in their Twitter posts, compared to those from US and GB, as indicated in the usage of future tense words ($H = 103.8$). On the other hand, the latter group discussed more of their past experiences and events, via the use of past tense words ($H = 369.9$). This difference is also much larger than what is observed in the case of the CTL cohort (7.2%, compared to 14.1% int he former).

> "I felt so lonely and I started to cry but nobody understood.. Nobody saw just how broken I really was." (↑ US, GB)

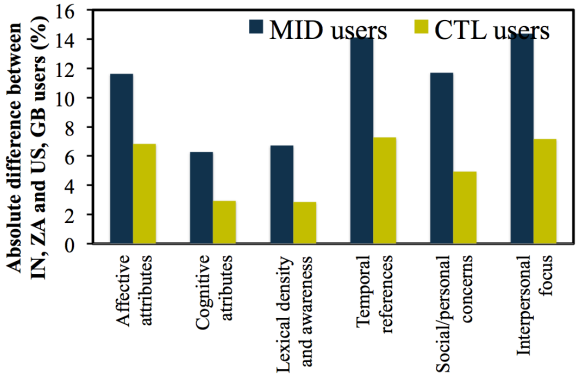


Figure 4: Mean absolute differences between IN, ZA and US, GB MID and CTL users per the various categories of linguistic measures. Difference for a specific measure is calculated as the ratio of the difference between the aggregated values of the measure for IN, ZA and US, GB users, to the value of the measure among US, GB users.

**Social/Personal Concerns (Culture).** Increased levels of social concerns, measured via family ($H = 154.1$), and social words ($H = 1613.5$), also tend to be observable in the posts shared by MID users from western cultures US and GB, compared to IN and ZA. These users also discuss more able health ($H = 124.5$), body ($H = 56.6$), and bio ($H = 27.8$) compared to the IN and ZA users. We conjecture the latter (majority world) group to be more self-conscious in utilizing a public platform like Twitter to discuss about the health aspects relating to a stigmatized illness.

> "I miss friendships with everyone that has ever held me close. Please be patient with me and I promise you that I'll be back!" (↑ IN, ZA)

> "I go through these spouts of depression and I could lay in bed for weeks. I just dont care about a single thing about this health condition" (↑ US, GB)

We further observe that the IN and ZA MID users converse less about sensitive or 'taboo' topics like religion ($H = 162.2$), death ($H = 38.4$), and sexual ($H = 131.0$) in their Twitter posts, compared to their peers from US and GB. This might indicate a latent social norm in the former cohort to avoid sensitive discourse on a publicly accessible platform like Twitter. Overall, the differences in these two sets of MID culture groups is much higher (11.7%) compared to the difference observable in the CTL cohort (4.9%).

**Interpersonal Focus (Culture).** Finally, we observe the MID users from the western cultures US and GB to demonstrate a greater tendency of social interaction (use of second person pronoun words ($H = 290.1$)), as well as attention to people around them (use of third person pronouns ($H = 49.9$). However, the users from IN and ZA express greater self pre-occupation and self-attentional focus, as measured in their use of more first person singular pronouns ($H = 128.0$). We also note that the aggregated interpersonal focus in these two sets of MID users (14.4%) is much higher in contrast to the CTL users (7.2%).

> "I suffered and was embarrassed to talk. When I spoke up my suffering lifted. There are people who will listen." (↑ IN, ZA)

### 5.2.2 Topical Differences

We now move over to examining the differences in the topics prevalent in the four cultures based on the LDA topic model.

First, we present topical differences per the three metrics: (1) Mean Percentage Difference in Topic Likelihood; (2) Spearman's Rank Correlation; and (3) Pointwise Mutual Information (PMI). The results are given in Table 5. Across all pairwise differences for the three metrics, we observe that US and GB `MID` users aggregatively differ less topically compared to US-IN, US-ZA, GB-IN or GB-ZA. For instance, the Spearman's rank correlation of topics used by US and IN `MID` users is only .17, however it is much higher for US-GB (.43). This shows that US-GB topics are more similar to each other in their usage, compared to the topics prevalent in US and IN posts. Another example is the measure of pointwise mutual information (PMI) for US and ZA. It is only .18, indicating their there are considerable differences in the likelihoods of various topics across these two cultures.

| | US | GB | IN | ZA |
|---|---|---|---|---|
| *Mean Percentage Difference in Topic Likelihood (%)* | | | | |
| US | 0 (0) | 22.1 (24.5) | 84.4 (69.6) *** | 71.8 (64.2) *** |
| GB | | 0 (0) | 58.6 (50.2) ** | 49.2 (37.5) ** |
| IN | | | 0 (0) | 44.6 (34.9) * |
| ZA | | | | 0 (0) |
| *Spearman's Rank Correlation* | | | | |
| US | 1 (1) | 0.436 (.519) ** | 0.174 (.295) *** | 0.209 (.352) *** |
| GB | | 1 (1) | 0.430 (.502) ** | 0.535 (.683) * |
| IN | | | 1 (1) | 0.321 (.504) ** |
| ZA | | | | 1 (1) |
| *Pointwise Mutual Information (PMI)* | | | | |
| US | 1 (1) | 0.491 (.638) * | 0.123 (.224) *** | 0.183 (.327) *** |
| GB | | 1 (1) | 0.285 (.391) ** | 0.391 (.482) * |
| IN | | | 1 (1) | 0.313 (.542) ** |
| ZA | | | | 1 (1) |

Table 5: Topical differences between the posts of `MID` users from the four cultures. Corresponding numbers in gray indicate the same for the `CTL` users.

Examining these numbers computed on the `CTL` users, we observe less significant differences in topic likelihoods. Together, these indicate topical differences in the posts of `MID` users from the four cultures to be higher than that observed in their `CTL` cohort peers.

We now discuss the context of usage of the top two topics that have a greater likelihood of occurrence in western cultures (US and GB posts) versus that in the majority world cultures (IN and SA posts shared by the `MID` users). Topic #29 centers around words like 'kill', 'stop', 'hate', 'pain', 'life', which revolve around self-deprecating, self-critical thoughts and self-destruction.

> "my god **life** makes me want to **kill** myself"

> "**everything** good is taken from me. **EVERYTHING**"

Next, topic #54 includes words like 'alone', 'lonely', 'people', indicative of loneliness and lack of social support. US and GB `MID` user posts associated with high likelihood of this topic tend to bear a tone of decreased self-esteem, and greater self-loathing.

> "**lonely** even when I'm not **alone**"

> "no one will ever be able to **understand**"

On the other hand, the two top topics that are more prevalent in `MID` users from majority world cultures, IN and ZA, compared to those from western cultures, US and GB, revolve around significantly different content. For instance, topic (#17) manifests confessions and regrets of individuals (e.g., 'faith', 'regret', 'strong').

> "I know I need to be **strong** but its just **hard** to now a days. **Regret**"

> "Will I ever **beat** this? I was doing so **good** but what happened to me"

Through topic #86, the MID users in IN and ZA express bereavement and marginalization due to the stigma associated with mental illness (e.g., words 'pretend', 'ashamed', 'embarrassed', 'struggle'). We observe that, these MID users may be trying to resolve their problems through posts associated this topic, believing that mental health can be maintained by avoiding bad thoughts and exercising will power. In fact, Wang et al. [54] have suggested that talking to another individual or group (e.g., a mental health worker) about psychological problems may be viewed by non-American sufferers as bringing disgrace on the family.

> "Beyond tired. I want this to end. I can't **pretend** anymore, need to be **strong**"

> "I'm still so **embarrassed**, but I'm **hurt** more by him not being there for me. I don't want to **struggle** alone."

## 6 Discussion

### 6.1 Theoretical and Practical Implications

Our work has provided some of the first detailed insights into how gender and cultural attributes relate to the content shared by individuals who disclose about their mental health challenges on these platforms.

**Reflecting on Gender Differences.** In the sample of mental health disclosing Twitter users we studied in this paper, there were considerable differences in the manner in which females and males appropriated the platform. Specifically, based on our different linguistic measures, females expressed higher sadness and anxiety, but lower anger and NA. These observations align with prior work in social psychology. Lieberman and Goldstein [33] also used the LIWC program to find that women in online support groups who experienced depression used high volume of anxiety words. Female mental health disclosers in our data also expressed greater social and familial concerns, compared to males. Literature has indicated that women tend to rely more on the social network of their family and the community, rather than the individual; whereas men exhibit a relative orientation toward stoicism [22]. These observations provide credence to our findings.

Per our topic model analysis, we further observed differences between female and male mental health disclosers. Male users appropriated Twitter to show greater detachment from the social realm and hesitation to seek help. Examining literature on mental health trauma, we find evidence supporting this observation. Norris et al. [37] found that the differences in PTSD severity were greater women compared to men. The authors speculated this to be attributable to adherence to more traditional gender roles, in which male "machismo" inhibits disclosure of distress.

**Reflecting on Cultural Differences.** Turning to cultural comparison of social media disclosers of mental health concerns, one of our findings revealed that individuals from the majority world countries, IN and ZA were less likely to express negative emotions in their Twitter posts compared to those from western countries, US and GB. The former group also expressed relatively less cognitive impairment and lowered tendency of social interaction as measured via pronoun use. As Marsella and Christopher [35] have argued, intrusive symptoms of many mental health concerns may be universal, however, avoidance or numbing may be more culturally based. Yeomans et al. [56] also discovered cultural factors in how people handle traumatic stress: for example, Asians are more reluctant to express distress in public, which aligns with our findings.

Moreover, based on our topic analysis, the IN and ZA cohorts were less likely to be disinhibiting or candid in their social media discourse, compared to their western peers. For instance, they used fewer words relating to self-critical or self-loathing thoughts, but shared evidence of bereavement and shame. This finding supports observations in prior cross cultural investigations of mental

health in eastern and western cultures; where individuals in the former culture were found to feel more socially stigmatized while sharing these experiences socially [20]. These practices have been attributed to social, political, economic, and communication barriers. Broadly, these observations indicate notable cultural differences in the Twitter activities of IN and ZA based mental health disclosers. They support the view that a conflict exists between traditional cultural values for these groups and the way in which psychological and social support are sought in western cultures.

Summarily, as pointed out earlier, gender and culture based dimensions of mental health are less understood today, primarily because of the challenges in obtaining reliable and normalized data on the same [49, 48]. In this paper, we showed the potential of social media as a way to quantify and assess these differences. To do so, we proposed, developed, and evaluated a rigorous and principled methodology that automatically and accurately identified disclosures of mental health challenges on social media, specifically Twitter. Obtaining adequate access to gold standard data (aka ground truth) is of prime importance in quantitative studies of mental health and social media. The most reliable data is gathered via self-reported means like surveys and user studies, however are difficult to scale. Gathering public data around mental health topics from social media incurs a much lower overhead, but the false positive rates in such data acquisition approaches can be alarmingly high. Our technique balances this through a semi-supervised approach, in which we leverage a much more reliable source of mental health data to glean genuine disclosures in another large but noisy source. Domain expert validation revealed the effectiveness of our approach. In short, beyond mental health, we believe our method provides a generalized template for gathering reliable data in problem contexts where either data acquisition is difficult, or noise, social and psychological considerations pose significant challenge in obtaining and curating quality data.

Besides, while our findings largely focus on validating known attributes of gender and culture in the context of mental health, our work brings to the fore novel mechanisms to do so. It proposes a rigorous quantitative framework through which the findings can be studied in large populations, tested for generalization, or adapted to multiple online, gender and cultural contexts. Finally, we were also able to discover nuances in mental health expression across the gender and culture subgroups, that may be challenging to quantify through traditional means. These primarily include recognizing the role of specific linguistic constructs and topics in mental health disclosures. Thus, a gender and cross-cultural approach to mental health, examined from naturalistic, unobtrusive data collected from social media, can provide guidance to the identification of appropriate responses from the mental healthcare system, as well as for healthcare policy globally.

### 6.2 Limitations

There are some limitations to our work that one should consider. First, we acknowledge that our findings are limited by our data acquisition capabilities. We relied on a set of hand curated keyphrases to seed our Twitter data collection. Although these were validated via consultation with a psychologist, they do not possibly include all of the ways in which Twitter users self disclose their mental health concerns. We were also limited by our ability to accurately infer gender and country information for the users in our dataset. While the methods we used have been employed in prior work and yield high precision content, we cannot rule out that our method suffers from a low recall problem.

Further, our method of obtaining mental health self-disclosure information only captures a subpopulation of Twitter users with a form of mental illness (i.e., those who are speaking publicly about what is usually a very private matter). This may not truly represent all aspects of the population as a whole. Moreover, this method in no way verifies whether this diagnosis is genuine (i.e., people are not always truthful in self-reports on Twitter), or whether they are clinically validated. However, given the stigma often associated with mental illness, it seems unlikely users

would self-disclose that they are diagnosed with a condition they do not have.

We also note caveats in our semi-supervised machine learning method of identifying genuine mental illness disclosures, and eliminating noisy control data samples. We used Reddit as a way to obtain weakly labeled mental health disclosure information. However, we do note that Reddit demographics and platform usage practices might be considerably different from that of Twitter. We believe our normalization step for comparative validity to be helpful in curbing the effects of these differences, but future work could explore other means.

Finally, our results, while suggestive, are correlational by nature. Therefore, it is impossible to say whether the expression of specific linguistic or topical constructs was the driving factor behind different gender and cultural subgroups' mental health disclosures on social media.

## 7 Conclusion

Psychology research recognizes considerable data gaps in gender based and cultural dimensions of global mental health. In this paper, we provided some of the first quantitative insights into gender and cultural differences associated with individuals who self-disclose to suffer from a mental health concern on Twitter. To identify genuine disclosures, we developed a semi-supervised learning based framework, which yielded high accuracy (96%) based on expert feedback. Thereafter, we explored the range of differences in the content shared by female and male users, and users who report on Twitter to be from US, UK, India or South Africa. We observed male users to express higher negativity and lower desire for social support, whereas majority world users (India and South Africa) demonstrated more inhibition in their expression. Our findings help validate, via use of social media data, a number of known characteristic differences in mental health experience of gender and cultural subgroups. We believe our work can encourage re-thinking of privacy-honoring health interventions to be more gender and culture aware, so that they could bring appropriate and personalized help and support to vulnerable individuals on social media.

## 8 Acknowledgements

We would like to thank Bapu Vaitla and Rebecca Furst-Nichols from the United Nations Foundation for helping frame the problem statement and for fruitful discussions. For gender inferences we would like to thank Thomas Baar from Leiden University?s Centre for Innovation, Parisa Zahedi from Risa-IT, Pim van de Pavoordt and Gerard Simons from Qualogy, and Maral Dadvar who helped through UN Volunteers. This work was supported by a United Nations Foundations grant to De Choudhury.

## References


[1] Mustafa Afifi. Gender differences in mental health. *Singapore medical journal*, 48(5):385, 2007.

[2] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI)*, 2016.

[3] Steven Bird. NLTK: the natural language toolkit. In *COLING/ACL Interactive presentation sessions*, 2006.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[5] Lera Boroditsky, Lauren A Schmidt, and Webb Phillips. Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought*, pages 61–79, 2003.

[6] Ana Mari Cauce, Melanie Domenech-Rodríguez, Matthew Paradise, Bryan N Cochran, Jennifer Munyi Shea, Debra Srebnik, and Nazli Baydar. Cultural and contextual influences in mental health help seeking: a focus on ethnic minority youth. *Journal of consulting and clinical psychology*, 70(1):44, 2002.

[7] Stevie Chancellor, Zhiyuan (Jerry) Lin, Erica Goodman, Stephanie Zerwas, and Munmun De Choudhury. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on Computer supported cooperative work & social computing*.

[8] Cindy Chung and James W Pennebaker. The psychological functions of function words. *Social communication*, pages 343–359, 2007.

[9] Aaron V Cicourel. Cognitive sociology: Language and meaning in social interaction. 1974.

[10] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.

[11] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in twitter. In *International Conference on Weblogs and Social Media (ICWSM)*, 2014.

[12] Aron Culotta. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1335–1344. ACM, 2014.

[13] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

[14] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from facebook data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 2014.

[15] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *International Conference on Weblogs and Social Media (ICWSM)*, 2014.

[16] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *AAAI Conference on Weblogs and Social Media*, 2013.

[17] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2016.

[18] Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[19] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[20] Simone Freitag, Anna Grimm, and Silke Schmidt. Talking about traumatic events: A cross-cultural investigation. *Europe's Journal of Psychology*, 7(1):40–61, 2011.

[21] James J Gross and Ricardo F Muñoz. Emotion regulation and mental health. *Clinical psychology: Science and practice*, 2(2):151–164, 1995.

[22] Francis Guillemin, Claire Bombardier, and Dorcas Beaton. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *Journal of clinical epidemiology*, 46(12):1417–1432, 1993.

[23] Kira Hall and Mary Bucholtz. *Gender articulated: Language and the socially constructed self*. Routledge, 2012.

[24] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.

[25] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 2010.

[26] Christopher M Homan, Naiji Lu, Xin Tu, Megan C Lytle, and Vincent Silenzio. Social structure and depression in trevorspace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 615–625. ACM, 2014.

[27] Yuheng Hu, Kartik Talamadupula, Subbarao Kambhampati, et al. Dude, srsly?: The surprisingly formal nature of twitter's language. In *ICWSM*, 2013.

[28] Mutsumi Imai and Dedre Gentner. A cross-linguistic study of early word meaning: Universal ontology and linguistic influence. *Cognition*, 62(2):169–200, 1997.

[29] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320. ACM, 2010.

[30] Max Kaufmann and Jugal Kalita. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*, 2010.

[31] Ichiro Kawachi and Lisa F Berkman. Social ties and mental health. *Journal of Urban health*, 78(3):458–467, 2001.

[32] Zoltán Kövecses. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press, 2003.

[33] Morton A Lieberman and Benjamin A Goldstein. Not all negative emotions are equal: The role of emotional expression in online support groups for women with breast cancer. *Psycho-Oncology*, 15(2):160–168, 2006.

[34] Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2015.

[35] Anthony J Marsella and Michael A Christopher. Ethnocultural considerations in disasters: An overview of research, issues, and directions. *Psychiatric Clinics of North America*, 27(3):521–539, 2004.

[36] Elizabeth L Murnane and Scott Counts. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1345–1354. ACM, 2014.

[37] Fran H Norris, Julia L Perilla, Gladys E Ibañez, and Arthur D Murphy. Sex differences in symptoms of posttraumatic stress: Does culture play a role? *Journal of Traumatic Stress*, 14(1):7–28, 2001.

[38] World Health Organization. *The World Health Report 2001: Mental health: new understanding, new hope*. World Health Organization, 2001.

[39] Johan Ormel, Michael VonKorff, T Bedirhan Ustun, Stefano Pini, Ailsa Korten, and Tineke Oldehinkel. Common mental disorders and disability across cultures: results from the who collaborative study on psychological problems in general health care. *Jama*, 272(22):1741–1748, 1994.

[40] Michael Paolisso and Joanne Leslie. Meeting the changing health needs of women in developing countries. *Social Science & Medicine*, 40(1):55–65, 1995.

[41] Vikram Patel, Ricardo Araya, Mauricio de Lima, Ana Ludermir, and Charles Todd. Women, poverty and common mental disorders in four restructuring societies. *Social science & medicine*, 49(11):1461–1471, 1999.

[42] Vikram Patel and Mark Winston. " universality of mental illness" revisited: Assumptions, artefacts and new directions. *The British Journal of Psychiatry*, 1994.

[43] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 2011.

[44] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[45] Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*, 2008.

[46] RJ Senter and EA Smith. Automated readability index. Technical report, DTIC Document, 1967.

[47] Robin W Simon. Gender, multiple roles, role meaning, and mental health. *Journal of Health and Social behavior*, pages 182–194, 1995.

[48] Rachel E Spector. Cultural diversity in health and illness. *Journal of Transcultural Nursing*, 13(3):197–199, 2002.

[49] Shelley E Taylor and Jonathon D Brown. Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 103(2):193, 1988.

[50] Harry C Triandis. The self and social behavior in differing cultural contexts. *Psychological review*, 96(3):506, 1989.

[51] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM, 2015.

[52] T Bedirhan Üstün and Norman Sartorius. *Mental illness in general health care: an international study*. John Wiley & Sons, 1995.

[53] Bapu Vaitla. The landscape of big data for development. *Data2x*, 2014.

[54] Xiangdong Wang, Lan Gao, Naotaka Shinfuku, Huabiao Zhang, Chengzhi Zhao, and Yucun Shen. Longitudinal study of earthquake-related ptsd in a randomly selected community sample in north china. *American Journal of Psychiatry*, 2000.

[55] Benjamin Lee Whorf, John B Carroll, Stephen C Levinson, and Penny Lee. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. Mit Press, 2012.

[56] Peter D Yeomans and Evan M Forman. Cultural factors in traumatic stress. *Sociocultural influences on mental health. Boston: Blackwell*, page 221Á44, 2009.

[57] Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.