Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia

SINDHU KIRANMAI ERNALA, Georgia Institute of Technology
ASRA F. RIZVI, Zucker Hillside Hospital, Psychiatry Research
MICHAEL L. BIRNBAUM, Zucker Hillside Hospital, Psychiatry Research
JOHN M. KANE, Zucker Hillside Hospital, Psychiatry Research
MUNMUN DE CHOUDHURY, Georgia Institute of Technology

Self-disclosure of stigmatized conditions is known to yield therapeutic benefits. Social media sites are emerging as promising platforms enabling disclosure around a variety of stigmatized concerns, including mental illness. What kind of behavioral changes precede and follow such disclosures? Do the therapeutic benefits of "opening up" manifest in these changes? In this paper, we address these questions by focusing on disclosures of schizophrenia diagnoses made on Twitter. We adopt a clinically grounded quantitative approach to first identify temporal phases around disclosure during which symptoms of schizophrenia are likely to be significant. Then, to quantify behaviors before and after disclosures, we define linguistic measures drawing from literature on psycholinguistics and the socio-cognitive model of schizophrenia. Along with significant linguistic differences before and after disclosures, we find indications of therapeutic outcomes following disclosures, including improved readability and coherence in language, future orientation, lower self preoccupation, and reduced discussion of symptoms and stigma perceptions. We discuss the implications of social media as a new therapeutic tool in supporting disclosures of stigmatized conditions.

CCS Concepts: • **Human-centered computing → Collaborative and social computing**; **Computer supported cooperative work**; **Social media**; • **Applied computing** → *Psychology*;

Additional Key Words and Phrases: social media; schizophrenia; mental health; self-disclosure; language; psychiatry; therapeutic benefits; Twitter

ACM Reference format:
Sindhu Kiranmai Ernala, Asra F. Rizvi, Michael L. Birnbaum, John M. Kane, and Munmun De Choudhury. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 41 (November 2017), 27 pages.
https://doi.org/10.1145/3134678

1 INTRODUCTION

> "Write something every day," she said
> "even if it's only a line,
> it will protect you."
> ... how then should it defend us?
> unless by strengthening
> our fierce and obstinate centers. – Elaine Feinstein [25]

2573-0142/2017/11-ART41 $15.00
https://doi.org/10.1145/3134678

Self disclosure, a process of “making the self known to others” [15], is identified as an important therapeutic element in the achievement of physical and mental well-being [35]. In individuals experiencing conditions associated with high stigma, like mental health challenges, self disclosure is a widely adopted mechanism for coping. Historically, “opening up” and disclosing about mental health experiences has been an established phenomena in psychotherapy, an activity that is situated between a therapist and client [21]. In stark contrast to such dyadic disclosures to a carefully selected receiver (the therapist), today, social media platforms have emerged as new arenas for *“broadcasting self-disclosures”* [36, 42]. The concept of broadcasting self-disclosures refers to sharing personal, sensitive information in public contexts, often to invisible audiences [43], and supported by the affordances of anonymity or semi-anonymity in these platforms [18]. Being quite distinctive from dyadic disclosures, whose prominent goal is relational development and deriving therapeutic benefits, broadcasting self-disclosures have impression management as a salient goal [8], including alleviating inhibitions [13], identifying confidants [45], building trust and intimacy [36], and finding a mechanism for emotional release [51].

Despite the pervasive adoption of these new broadcasting self-disclosure practices, particularly around stigmatized mental health concerns [14], how these disclosures lead to behavioral changes on social media platforms, and if they help an individual meet their therapeutic goals, are less explored. Recent research has studied mental health disclosures shared on social media platforms such as Reddit and Instagram, exploring the ways in which linguistic attributes such as affect, cognition and linguistic style may reveal cues about one’s psychological state [12, 18]. The attributes of support seeking nature of anonymous disclosures on Reddit among sexual abuse victims and depression sufferers has also been examined [2, 3]. Together, these works reveal how the unique needs around stigmatized mental health experiences can be met when one self discloses on a public platform like social media. We contribute to this line of research by examining how one of the most prominent goals of offline mental health disclosures, therapeutic benefits, that typically happen in offline therapist-client settings, translate to the context of online broadcasting self-disclosures. Specifically, we ask the question: *Can we identify specific linguistic markers that indicate behavioral changes following disclosures of schizophrenia on social media? Are these changes indicative of any therapeutic outcomes?*

We address these questions in this paper by drawing from the literature on psycholinguistics and the expressive writing paradigm [52]. We specifically focus on disclosures of schizophrenia diagnoses made by individuals on Twitter. Schizophrenia is one of the most stigmatized mental illnesses [23]. The psychopathology of the condition indicates that the sufferers are particularly known to benefit therapeutically from intimate self-disclosures, such as writing about their personal feelings and experiences about emotion-laden topics [61], as noted in Elaine Feinstein’s [25] quote above. Therapeutic benefits of schizophrenia particularly include improved emotional expressivity, social, and linguistic functioning, and lowered disorganized and paranormal thinking [38].

Towards our research goal, we leverage an expert (clinical psychiatrist) validated dataset of 146 disclosures of schizophrenia shared on Twitter over an year-long period. We adopt a theoretically grounded quantitative approach to first identify temporal phases around disclosure during which symptoms of schizophrenia are likely to be manifested. Then, we propose a variety of psycholinguistic, structural, and topic-based linguistic measures to characterize the behavioral changes preceding and following the disclosures. This allows us to identify linguistic markers whose changes precipitate the disclosures, as well as the changes that ensue the disclosure events.

Based on this analytical methodology, our results indicate significant behavioral differences before and after the disclosures, many of which align with known markers of reduction in the negative syndromes of schizophrenia. As a way to establish causation, we observe these differences to be minimal in a matched control group. Specifically, we find that following disclosures on Twitter, individuals express lowered stereotypy such as word repetitiveness, and demonstrate improved readability, linguistic complexity, and topical coherence in the content shared on Twitter. They also show greater future orientation and an increasing positive affect trend, as well as lowered attention to the self. Interestingly, following disclosures, the individuals tend to engage in reduced discussion of symptoms and stigma perceptions on social media. Situating our analyses within the socio-cognitive model of schizophrenia [47], the expressive writing paradigm [52], and feedback from clinician experts, we observe these observations to characterize the prominent effects of "opening up" about a stigmatized condition like schizophrenia. Summarily, our work signals therapeutic outcomes following disclosures made on a public platform like Twitter.

We discuss how naturalistic and unobtrusive data from social media like Twitter can be interpreted as forms of expressive writing, specifically around broadcasting self-disclosures on stigmatized conditions like mental illnesses. We also discuss the implications of social media as a therapeutic tool supporting candid self disclosures around mental illnesses.

2 RELATED WORK

2.1 Self Disclosure and Therapeutic Outcomes

Celebrated sociologist Erving Goffman emphasized the importance of "sympathetic others" in helping people cope with difficult experiences, as well in enabling self-disclosure [27]. Self-disclosure provides an opportunity to express one's thoughts and feelings, develop trust and build intimacy in personal relationships [36]. However, the act of self disclosure is a much more complex and critical process for people with a concealable, stigmatized identity such as mental illness [58]. On the one hand, the stigma around these conditions may risk unfavorable outcomes such as social rejection and discrimination and might be detrimental to well-being. Experimental manipulation studies found that participants do not experience the benefits of disclosure when confidant reactions are neutral or negative [60]. But on the other hand, positive outcomes of disclosure due to opening up, include a wide range of therapeutic benefits leading to both physical and mental well-being, such as lowered psychological distress [52]. For instance, studying the post-traumatic stress (PTSD) experiences of rape and sexual assault victims, Ullman and Filipas found that disclosures led to more positive and fewer negative social reactions [66]. This complex nature of both possibilities is nested within an ongoing process of *"stigma management"*—coping with the psychological and social consequences of their identity [27].

The Expressive Writing Paradigm. Among the several forms of self disclosure in the context of stigmatized or traumatic experiences, expressive writing has led to a new paradigm in psychotherapy. The seminal work of Pennebaker and colleagues includes several studies involving participants writing about traumatic or emotional experiences over consecutive days [52, 62]. These studies observe long term benefits of expressive writing in the form of both health outcomes (fewer stress related visits to the doctor, improved mood, lowered blood pressure, fewer post-traumatic intrusion and avoidance symptoms) and social/behavioral outcomes (reduced absenteeism from work, improved working memory, improved sporting performance, altered social and linguistic behavior) [6]. There are several hypotheses as to why self-disclosure through writing has these therapeutic benefits. Emotional catharsis, confrontation of previously inhibited emotions, development of a coherent narrative over time, reflecting increasing cognitive processing of the experience and repeated exposure are some of the notable explanations from prior literature [53, 67]. In comparison to the above studies, which explored the effects of writing in a laboratory setting, social media provides an unprecedented, unobtrusive form of expression through writing. Therefore, in this paper, we explore the therapeutic nature of broadcasting self disclosures of schizophrenia shared on Twitter. We draw from prior literature and examine the re-purposing of Twitter as a platform for expressive writing.

Language and Psychological Well-being. Language is a powerful form of expression, including of self-disclosures. It is recognized that language shapes and drives one's thoughts, actions, social relationships, and emotional processes. In fact, the words that people use in everyday lives convey information about their mental, social and physical states [55]. Several textual analysis techniques have been applied on the self disclosure texts from expressive writing mechanisms and they report language use as a marker of cognitive processes, personality style and social integration. For example, word choice in writing is identified to predict physiological changes and health [55]. Positive emotion words are also correlated with well-being, and increases in the use of causal and insight words (*realize*, *understand)* demonstrate heightened functioning. Taking inspiration from these prior works, in this paper, we present linguistic measures that quantify behavioral changes around social media disclosures of schizophrenia.

2.2 Self Disclosure Studies on Social Media

A rich body of work in the Computer Mediated Communication (CMC) literature has studied self disclosures and the socio-cognitive processes centered around them. Through several experimental and anecdotal evidence, CMC and other Internet-based behaviors have been characterized to exhibit high levels of self disclosure [34]. In fact, high self-disclosure has been recognized to lead to dis-inhibition on the Internet [33]. At the same time, self disclosure in CMC contexts is also argued to be beneficial, having been linked to trust and group identity [35], as well as playing an important role in social interactions by reducing uncertainty [15].

Turning to research on social media, an emergent line of research has investigated the nature of self disclosures on social media and online communities. Several quantitative studies have focused on identification, modeling and characterizing differences in multi-modal (textual, visual) forms of self disclosure on social media [7, 18, 20, 42, 70]. Similarly, from a qualitative perspective, prior work has studied how individuals undergoing gender transition appropriate Facebook for engaging in sensitive disclosures of their experiences [28]. Existing literature has also investigated unique design affordances of social media like "throwaway" accounts, in providing context-specific anonymity for first-time disclosures on abuse related posts on Reddit [2]. In another study, Andalibi et al. found that individuals struggling with negative emotions, such as that related to depression or self-harm, use Instagram to self-disclose and engage in social exchange and storytelling about their stigmatized experiences [3].

We situate our work analyzing social media disclosures of schizophrenia within this body of work. Specifically, we extend previous research by presenting methods to characterize linguistic markers leading to and following a self disclosure. Other works so far, have not exclusively studied changes around self disclosure. Further, the therapeutic outcomes of these disclosures are under-investigated in current literature. Our work aims to address these gaps and extend our understanding of the unique outcomes of social media disclosures of a stigmatized condition, schizophrenia.

2.3 Mental Health and Social Media

In recent years, a growing body of work has employed large scale social media data to model and infer mental well-being of individuals and populations [49]. In an early work, studying behavioral changes around major life events (childbirth), De Choudhury and colleagues [17] provided methods for detecting and predicting significant postpartum changes in behavior, language, and affect from Twitter data. This work serves as a formative step for our work, in order to systematically examine how behaviors are likely to change following a different event: public social media disclosure of one's schizophrenia experiences. De Choudhury et al. in another work also examined year-long Twitter postings of individuals suffering from major depressive disorder to build statistical models that predict the future occurrence of depression [19]. Recently, similar approaches have been adopted to identify and understand social media derived risk and psychological markers of other mental health conditions, ranging from postpartum depression [17], eating disorders [12, 13], post-traumatic stress [14], and other conditions [3, 42].

One of the main motivations in this emergent area of research has been to leverage naturalistic, unobtrusive data from social media to understand mental health states and related experiences. In contrast to the self-reported methodology in clinical therapy, where responses typically comprise of recollection of (subjective) health facts, social media captures behavior and language in a naturalistic setting. This provides access to real-time activity and psychological states that can be analyzed to discover and predict behavioral markers associated with a condition. Our work extends this research direction by employing naturalistically shared Twitter data as a mechanism to identify markers that precipitate and follow disclosure of a stigmatized condition: schizophrenia.

2.4 Background on Schizophrenia and Role of Social Media

Schizophrenia, a mental disorder characterized by abnormal social behavior and distorted perceptions of reality, is one of the most debilitating of mental illnesses [1]. The condition is often described in terms of positive and negative (or deficit) symptoms [38]. Positive symptoms are those that most individuals do not normally experience, but are present in people with schizophrenia. They can include delusions, disordered thoughts and speech, and hallucinations. Negative symptoms are deficits of normal emotional responses or of other thought processes. They commonly include flat expressions or little emotion, poverty of speech, inability to experience pleasure, lack of desire to form relationships, and lack of motivation. Therapeutic efforts, therefore, largely focus on remission of these positive and negative symptoms, and reducing the likelihood of a psychotic relapse [59]. In our work, we examine to what extent we can measure attributes of these symptoms linguistically in self-disclosing schizophrenic individuals, and thereafter, whether the disclosure event leads to noticeable changes in specific symptomatic expressions. With clinical collaborations, we seek to then examine if these changes align with therapeutic outcomes of the condition.

The clinical literature also identifies phases of manifestation of the symptoms of schizophrenia in an individual's course of illness. Two of these phases are particularly relevant in our work: the prodromal and the active phase [32, 50]. Prodrome is defined as the period before 'florid psychotic illness', where individuals are identified as being at 'clinically high risk' for developing a psychotic illness and may manifest sub-threshold symptoms. Active phase is when an individual transitions from a prodromal stage to a full blown psychotic state including delusions, grandiosity, and thought disorganization. Our expert assessments of schizophrenia disclosures draw from this literature—to understand the therapeutic outcomes of social media disclosures of the condition, we restrict our study to those individuals who self-disclose during either the prodromal or active phase.

Recent research has studied technology use by individuals suffering from schizophrenia and related psychotic disorders [44, 46]. Matthews et al. [44] found that technology use in this vulnerable population is often impacted by underlying mood, and that, these differential patterns in technology use may indicate incipient mood episodes. However, to our knowledge, research on the use of social media platforms by this population is lacking. Given that schizophrenia is a highly stigmatized condition, literature has recognized the value of candid disclosures resulting in improved well-being and therapeutic benefits among individuals challenged with this illness [41]. For instance, participation in offline self-help groups and advocacy organizations has been found to facilitate self-disclosure—such activities help challenge private shame about the illness, enhance self-esteem, enable persons to be more resilient in response to stigma experiences, and thereby support symptomatic coping [26]. Despite these known benefits, self disclosure goals of schizophrenic individuals in the online environment, and their resulting outcomes, are under-investigated. Our work in this paper seeks to close these gaps.

Table 1. Clinician-contributed key-phrases for Twitter data collection.

Diagnosed me with (schizophrenia | psychosis)
Diagnosed schizophrenic
I am diagnosed with (psychosis | schizophrenia)
I am schizophrenic
I have been diagnosed with (psychosis | schizophrenia)
I have (psychosis | schizoaffective disorder | schizophrenia)
I think I have schizophrenia
My schizophrenia
They told me I have schizophrenia
I was diagnosed with (psychosis | schizoaffective disorder | schizophrenia)
Told me I have (psychosis | schizophrenia)

Table 2. Descriptive statistics of acquired Twitter data, including statistics of clinician annotations.

Total number of disclosure tweets	21,254
Total number of unique users	15,504
Total number of unique users disclosing in 2014	3,338
Annotated users from 2014	671
Total tweets of all annotated users	12,272,534
Mean tweets per annotated user	18,289.92
Median tweets per annotated user	7247.0
Number of annotated ‘Yes’ users	146
Tweets of annotated ‘Yes’ users	1,940,921
Number of annotated ‘No’ users	424
Tweets of annotated ‘No’ users	8,829,775
Number of annotated ‘Maybe’ users	101
Tweets of annotated ‘Maybe’ users	1,501,838

3 DATA

3.1 Data Acquisition

We first present our methodology for obtaining data on self-disclosures of schizophrenia as expressed on Twitter. Our approach relied on compiling a list of key-phrases indicative of self-reported diagnoses of schizophrenia, which could eventually serve as search queries to discover potential sufferers on Twitter. Coppersmith et al. had leveraged this method; they observed that individuals engage in public social media self-disclosures of mental illnesses, often to seek support from others in their online social network, to fight the stigma of mental illness, or perhaps as an explanation of some of their behavior [14]. Due to our focus on a clinical condition, in our approach we consulted with two clinical psychiatrists to generate this list of key-phrases: Refer. Table 1.

Then, using the key-phrases as regular expressions, we constructed case-insensitive search queries, and then filtered data from the public Twitter stream. This resulted in a total of 21,254 posts authored by 15,504 unique users between 2012 and 2016. Since our research goal involves temporal analysis of Twitter content shared before and after the self-reported schizophrenia diagnoses, we selected disclosure posts and their authored users from the year 2014 (middle of time period of our collected data), so that, as a second step, we could crawl sufficient amount of their shared content prior to and following the disclosure event. For each filtered user in 2014, we extracted their Twitter timeline data from 2012 to 2016 using a web based Twitter crawler[1]. This timeline data for each filtered user included tweet text, username, posting time, hashtags, mentions, favorites, geo-location and tweet ID. We report basic descriptive statistics of this acquired data in Table 2.

3.2 Expert Annotation

Although the key-phrases involved first-person reports of schizophrenia experiences and diagnoses, several filtered tweets included noisy data in the form of disingenuous, inappropriate statements,

[1] https://github.com/Jefferson-Henrique/GetOldTweets-python

jokes, and quotes. For example, note the tweet: “*I wish I had schizophrenia. So I can escape reality*”. To obtain an accurate sample of genuine disclosures we design an annotation task for expert (psychiatrist) validation. However, since in many cases, the disclosure tweet by itself may not provide sufficient information to guide the experts, and because disclosure is often described as a process [24], with each filtered post as a self disclosure, we extracted 10 consecutive tweets before and 10 consecutive tweets after the disclosure post to form a set of contextual posts around disclosure. We refer to it as the *“disclosure context”* of a Twitter user. These disclosure contexts from 671 users were then passed on to two expert raters for annotation of authenticity. The raters are clinical psychiatrists at a large and renowned New York based hospital, with extensive expertise working with individuals with early stage psychosis/schizophrenia.

The raters devised a three class categorization of authenticity for the annotation task, where the disclosure context of each user was classified into one of the following three classes: “Yes”, “No” and “Maybe”. Class “Yes” contained users who appeared to have genuine disclosures. Class “No” contained users who had inauthentic posts including jokes, quotes, or were from accounts held by health related blogs. Finally, class “Maybe” contained users for whom the experts could not confidently appraise the authenticity of the disclosure. Paraphrased examples of posts belonging to each of these classes are presented in Table 3.

Table 3. Example clinician annotated tweets belonging to the “Yes”, “No” and “Maybe” classes (“Yes” indicates a genuine disclosure of schizophrenia).

Class ‘YES’
My mom took me to the doctor and he told me i have schizophrenia
I found out I have schizophrenia..was in a mental hospital for the last six days
Class ‘NO’
Twitter is an acceptable way to talk to yourself without being diagnosed schizophrenic
I wish I had schizophrenia. So I can escape the reality.
Class ‘MAYBE’
I am convinced that my schizophrenia is a better friend than you are.
Yes, I have schizophrenia. But, I’m not crazy.

Each rater annotated users separately and subsequently reviewed the annotations together to achieve consensus. The inter-rater reliability based on the Cohen’s κ measure between the two raters for the three class annotation task was 0.56. Upon inspection, although the disagreement was mainly caused by the “Maybe” class, between classes “Yes” and “No” the agreement was 0.81. The annotation task was conducted for a sample of 671 users resulting in 146 “Yes”, 101 “Maybe” and 424 “No” users. These three classes of users shared 1,940,921, 1,501,838 and 8,829,775 tweets respectively with a mean of 13293.98, 14869.68, and 20824.94 tweets per user. Additional descriptive statistics of this acquired Twitter data are listed in Table 2. The dataset that we use hereafter for our analysis is thus this sample of 146 users annotated to have made genuine disclosures (Class “Yes”).

3.3 Compiling Timeline Data on Genuine Disclosures

Towards addressing our research goal, which seeks to identify the linguistic changes around schizophrenia self-disclosures, we now compile Twitter timeline data of users who were annotated to have made genuine disclosures on Twitter, per the above expert annotation task. For each of the 146 users, the posting time of the disclosure tweet is taken as the *disclosure date*. Each user within the sample would have disclosed at different times during the year of 2014. Therefore, to analyze the data of all the users, we identified a fixed length time period preceding and succeeding the disclosures. For this, we adopted an empirical approach by first generating cumulative density functions (CDFs) of the number of Twitter posts shared by the users before and after their respective disclosure dates. These CDFs are shown in Figure 1(b) and (c). Based on these figures, we observe that most users (around 80%) have posts for at least 200 days before and 400 days after the disclosure dates, spread over 2014. Hence, going forward, we choose each of the 146 genuine disclosure users timeline data spanning 200 days before and 400 days after their disclosure dates as a fixed length time period with which we pursue the analyses.

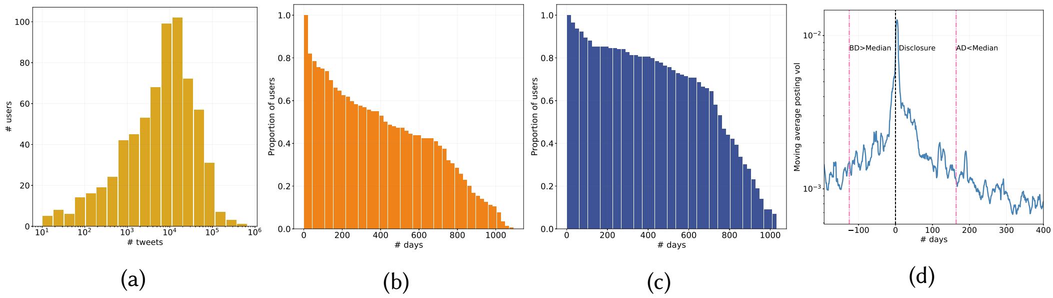


Fig. 1. (a) Distribution of number of users over number of tweets. (b) CDF of post distribution over the 146 genuine disclosure users preceding the disclosure dates. (c) CDF of post distribution over the 146 genuine disclosure users following the disclosure dates. (d) Temporal phases identified around disclosure using a moving average model of posting volume. The central vertical line indicates the disclosure event, while the vertical lines on its two sides indicate the boundaries of the *BD* and *AD* phases.

3.4 Matched Control Data

Parallelly, we obtain a control dataset, to allow robust statistical comparisons between Twitter users who choose to self-disclose regarding schizophrenia, and those who do not. This also allows us to establish causation between the schizophrenia disclosures and the linguistic changes we seek to see preceding and succeeding them—statistical matching is a established technique to demonstrate causation in observational data, like ours [63]. For each user who made a genuine disclosure on day d, as identified by the above expert annotation task, we identify a "matched control user" who had posted on Twitter, in the same year, on either of the days $d-1$, d, or $d+1$: this allows us to simulate a "control disclosure". Additionally, we ensure that the matched control user does not have any mentions of schizophrenia disclosures in their posts shared on their timeline. In this way, we compile the timeline data of 146 matched control users for the disclosure year 2014, and thereafter 200 days of pre- and 400 days post- control disclosure data for each of them. This resulted in 832,052 posts from the 146 matched controls, with a mean of 5699 posts (σ=6984.25) per control user.

4 METHODS

4.1 Identifying Temporal Phases around Disclosures

Analyzing the behavioral changes that surround self disclosures of schizophrenia necessitates identifying data spanning pre- and post-disclosure phases where the symptoms of schizophrenia are most likely to be manifested. For this purpose, we draw from findings given in the clinical literature, specifically around the prodromal and active phases of schizophrenia, as described in the Related Work section. Although self-disclosure on social media cannot be directly correlated to these phases due to the lack of information about the actual time of the users' diagnoses, we anticipate that the disclosures happen *after* a formal diagnosis of the condition – an observation that was confirmed by our experts during the annotation task. Therefore, we also expect that the prodromal and the active phases (which denotes the beginning or complete manifestation of symptoms) to occur before one self-discloses about their condition on social media. Similarly, we anticipate the active phase to persist for a short while after the self-disclosure events.

We leverage these clinically grounded observations of schizophrenia phases toward our temporal analysis of linguistic cues centering around self-disclosures of schizophrenia. We devise an approach to identify two phases in each (genuine) disclosing user's pre- and post-disclosure timeline data (compiled above) during which the symptoms of schizophrenia are most likely to be manifested: one preceding the disclosure (referred to as *"Before Disclosure"* or *BD*), and the other following it (*"After Disclosure"* or *AD*). Per clinical literature referred above [32], during these *BD* and *AD* phases, we expect the users to show markers of social withdrawal on social media. Abrupt declines in posting activity on social media are noted to be a sign of social withdrawal in prior work [19]. E.g., in individuals challenged with postpartum depression, changes in sociality and behavior on social media, manifested through patterns of posting is known to be a notable risk marker [19]. Therefore, we utilize measures of changes in posting volume of an individual (normalized number of posts per day) as a way to identify these *BD* and *AD* phases around the genuine disclosures.

Our phase identification approach includes the following steps:

(1) Drawing from the time series analysis literature [29], we first calculate the daily posting volumes of each of the 146 genuine disclosure users spanning their 200 days pre-disclosure and the 400 days post-disclosure timeline data. On this posting volume time series data of each user, we then computed the rates of change throughout the pre- and post-disclosure period, by employing a weekly moving average model. This model allows us to smooth out small local fluctuations, diurnal variations and seasonality while allowing comparison between the posting volume at day t and that during the 7 days preceding it: days $t - 1$ through $t - 7$.

(2) Having computed weekly rates of change in posting volume of the users, we now present the next step in the identification of the *BD* and *AD* phases. We conjecture that the first time point (day) of significant rate of change in posting volume would demarcate the boundaries of the *BD* and *AD* phases. Therefore, we compute the medians of the time series of weekly rates of changes in pre-disclosure data as well as that of the weekly rates of changes in post-disclosure data of every user. Since median is a reliable and robust measure of central tendency, we use it as a cutoff to define the *BD* and *AD* phase boundaries. Specifically, the first time point (day) in pre-disclosure data when the rate of change of posting volume becomes *greater* than the pre-disclosure median rate of change is taken to indicate the start of the *BD* phase. We assume the *BD* phase to end on the day before the day of schizophrenia disclosure. Similarly, the first time point (day) in post-disclosure data when the rate of change of posting volume becomes *lower* than the post-disclosure median rate of change is taken to indicate the end of the *AD* phase. Extending our previous logic, we assume the *AD* phase to begin a day after the disclosure is made by the corresponding user.

As shown in Figure 1(d), the median rate of change of posting volume for all the 146 users based on pre-disclosure data was 0.00329, and day -132 indicated the first point in time when the rate of change *surpassed* this median[2]. Similarly, the figure also shows that the first day on which the rate of change in posting volume was *lower* than the post-disclosure median (= 0.00181) was day 157. Since the rates of changes in posting volume are computed weekly, we adopt the following day demarcations to define the *BD* and *AD* phases: $BD = d_{-137}$ to d_{-1}; and $AD = d_1$ to d_{156}, assuming Disclosure = d_0.

To allow meaningful comparison, we mapped these *BD* and *AD* phases to the extracted data of the matched control cohort as well, to obtain control *BD* and *AD* phases.

[2] We assume day 0 as the day of schizophrenia disclosure.

4.2 Linguistic Markers Around Schizophrenia Disclosures

In this subsection, we present methods to quantify the linguistic markers of Twitter users around their disclosures, i.e. the BD and AD phases. To define the measures, we follow a theoretically grounded approach, drawing from prior work in clinical psychology and psycholinguistics.

4.2.1 Psycholinguistic Measures. A rich body of literature in psycholinguistics has identified the association of linguistic usage to emotion and behavior, including mental health states of individuals [55]. To quantify such psycholinguistic changes in the phases around disclosure, we use three categories of measures: (1) Affective attributes, (2) Cognitive attributes and (3) Linguistic style attributes. All of the above measures are calculated based on the well-validated psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) and have been employed extensive in prior social computing work [54].

To measure affective attributes from language, we consider the categories *positive and negative affect, anger, anxiety, sadness and swear* from LIWC. These categories help in characterizing the emotional expression around self-disclosure. Next, to quantify cognitive and perceptive attributes, we use the categories *cognitive mechanisms, discrepancies, inhibition, negation, causation, certainty, and tentativeness, see, hear, feel, percept, insight, and relative*. Quantifying these cognitive and perceptive attributes as manifested in language can lead to insightful markers of one's mental stability and cognitive complexity, functioning or impairment, especially in relation to their diagnosis of schizophrenia. Finally, to quantify linguistic style, we use the following four measures: (a) Function words: consisting of the lexicons in the categories *verbs, auxiliary verbs, adverbs, prepositions, conjunctions, articles, inclusive, and exclusive* (b) Temporal references: comprising of the categories *past, present and future tense* (c) Social and Personal concerns: consisting of words belonging to categories *family, friends, social, work, health, humans, religion, bio, body, money, achievement, home, sexual, and death* and (d) Interpersonal awareness and focus: comprising *1st person singular, 1st person plural, 2nd person, and 3rd person pronouns*. Literature has indicated that pronoun use, in particular, can quantify an individual's self and social awareness and can reveal mental well-being, including that manifested in social media [11]. Taken together, linguistic style captures the changes in psychological processes, awareness, personality, and social environment around self disclosure.

To calculate the above mentioned psycholinguistic attributes, we use the textual content of the posts of each user during the *BD* and *AD* phases respectively. The LIWC scores for each category are normalized by dividing the category word count by the post length, and an average value for each category is computed per user during *BD* and *AD* phases.

4.2.2 Linguistic Structures. Beyond just the usage of functional words and informational content, which is captured by the LIWC categories presented above, sentence structures and boundaries form an important aspect of written language [47]. In this subsection, we define measures of change characterizing linguistic structural attributes of Twitter posts spanning the BD and AD phases of the schizophrenia disclosing users.

Readability. This measure seeks to quantify the readability of the language used in the post of the disclosing users. The relation between thought or meaning and forms of grammatical organization have been extensively studied as symptoms of schizophrenia [41]. Specifically, the socio-cognitive model demonstrates that individuals with schizophrenia use simpler grammatical forms in spoken and written communication, as well as exhibit a lack of spontaneity and fluency [30].

To capture this linguistic structural measure, we use the Coleman-Liau Index (CLI). CLI is a readability assessment test based on character and word structure within a sentence [57]. It approximates a U.S. grade level required to understand the text and is calculated using the formula: $CLI = 0.0588L - 0.296S - 15.8$, where, L is the average number of letters per 100 words of content and S is the average number of sentences per 100 words. In our case, the CLI is calculated from the day-wise aggregated content of posts by each user during the BD and AD phases respectively.

Stereotypy. Next, we consider two measures of stereotypic thinking in the posts shared by users during the BD and AD phases: (1) *Word repeatability*, and (2) *Word complexity*. Per the socio-cognitive model [47], sufferers exhibit signs of impoverished speech and content, word repetitions, decrease in usage of complex words or sentence verbosity, in favor of a greater number of simple ones. These attributes are typically characterized as stereotypy, an individual with schizophrenia is likely to host repetitive thoughts that interfere with their ability to think and communicate [4].

In our data, we measure *word repeatability* by calculating the normalized count of non-unique words (or unigrams) in a Twitter post of a user during the *BD* or the *AD* phase, while *word complexity* is computed by estimating the normalized length of a word (or a unigram) in a disclosing user's posts during the *BD* or the *AD* phases.

4.2.3 Domain-Specific Content Measures. Twitter is largely used as a microblogging platform where people share a wide range of everyday experiences and happenings. This likely applies to the self-disclosed user group in this study as well. However, beyond the everyday experiences, individuals challenged with schizophrenia are likely to share content specific to their experiences of symptoms of the condition. E.g., over-representation of abstract and metaphysical termini or verbal abuse of death, power and hostility themes are known to have a strong bearing with the schizophrenic vision of the world [4].

To understand linguistic usage specific to the diagnosis or experiences of schizophrenia, we build a domain-specific lexicon. For this, we use Reddit as our data source. Unlike Twitter, Reddit, through its community (or subreddit) feature, offers an online space for seeking and receiving support on several mental health topics, including schizophrenia. The discussions within these communities are mainly around experiences of symptoms of schizophrenia, medical services, emotional, and informational support—providing a rich data source to compile markers of schizophrenia specific content and symptoms. Therefore, we leverage data from three of the largest support communities for schizophrenia on Reddit: */r/psychoticReddit, /r/Schizophrenia* and */r/Schizophrenic* to build a domain-specific lexicon—these subreddits were identified in consultation with two clinical experts. Combining the post and comment data from the three subreddits, an n-gram ($n = 2$) language model is built on the data. Then, to identify the n-grams that are most relevant to the case of schizophrenia, two clinical psychiatrist practitioners assigned binary "Yes"/"No" annotations to the top 5000 unigrams and the top 5000 bigrams given by the Reddit language model. Here, an n-gram with a "Yes" label would imply that it is directly relevant to the topic, symptoms and experiences of schizophrenia. The annotation task resulted in 1981 unigrams and bigrams relevant to schizophrenia which provide a schizophrenia lexicon we used to quantify differences in domain-specific content around disclosure. For each token in this lexicon, normalized occurrence is finally calculated per user during the *BD* and *AD* phases.

4.2.4 Topical Measures. One of the most emphasized category of language disturbances in schizophrenia sufferers has been discourse coherence disturbances: tangential responses, derailments and non sequitur responses [4]. The socio-cognitive model describes a lack of syntactic, semantic and pragmatic coherence in the language of schizophrenics [47]. Additionally, interview transcript data of schizophrenia sufferers indicates thematic and semantic variation in expression [9]. Our final set of measures thus includes measures of thematic variation and topical coherence.

To quantify both of these measures in the Twitter content of users shared before and after their disclosure, we employ topic modeling [10], which is a useful, established approach to identify themes in data that are not captured by textual analysis at the level of tokens and sentences. To build a topic model, we run Latent Dirichlet Allocation using MALLET: MAchine Learning for

LanguagE Toolkit[3], which has been an established method in prior work on mental health and social media [12]. After preprocessing the textual data in posts by removing stopwords and URLs, we use the default hyper-parameters of MALLET to extract 30 topics. Thereafter, for each BD and AD post from every user, we compute the post-specific topic posterior distributions.

Theme Variation. Now, to identify thematic variation manifested in the Twitter posts of users around the disclosure event, we employ a qualitative technique to identify semantically interpretable, broader themes from the LDA generated 30 topics. The same two human raters as above (clinical psychiatrist practitioners) were employed to perform semi-open coding on the extracted topics, drawing from their knowledge of the condition and interactions with schizophrenia sufferers. For each topic, by analyzing the top contributing keywords to the topic and the Twitter posts with the highest probability for the topic, the raters built a set of topical descriptors for each topic, and then agglomeratively combined topics into themes. Finally, we calculate the z-scores of the average probability of each theme per day across all users; this allows us to identify theme-specific variation manifested in the *BD* and *AD* phases. Since z-scores reveal relative differences in the values of a given distribution, it qualifies as a suitable metric to study theme variation over time.

Topical Coherence. Finally, we characterize topical coherence as a similarity measure between consecutive day's topic distributions. For calculating the topical coherence measure during the BD and AD phases, we consider the topic distribution of a user's posts on day t, and compare it with the mean topic distribution over all posts shared by the same user in the previous week, i.e., days $t - 1$ through $t - 7$. Since we are comparing distributions, we employ the cosine similarity metric. Thus, a higher cosine similarity would indicate that the content shared on day t is topically coherent with respect to the same in the week before.

5 RESULTS

5.1 Changes in Psycholinguistic Measures

We first present the behavioral differences revealed by the psycholinguistic measures during the *AD* period compared to the *BD* phase. Table 4 gives a summary; we indicate the mean value of each measure in the *BD* and *AD* phases, their mean difference across the Twitter posts of all self-disclosing users, as well as the results of Wilcoxon signed-rank tests comparing the measures across the *BD* and *AD* phases.

Affective attributes. To begin with, for the affective attributes, there is a significant increase in overall *negative affect* (mean difference 5.6%) and decrease in overall *positive affect* (mean difference 2.4%) right after the disclosure. This relates to the expressive writing literature, which associates an immediate increase in negative affect and decrease in positive affect after opening up about emotionally distressing topics rather than immediate relief of emotional tension [6]. The exposure to distress and confrontation of stigmatized conditions like schizophrenia might also implicate the increase in anger, sadness (mean differences 4.7%, 2.9% and 7.1% respectively). E.g., consider the paraphrased tweet: *"I'm sad sad sad sad"*.

Cognitive attributes. Among the cognitive attributes, we observe an increase in *certainty* words after disclosure, demonstrating heightened emotional stability indicative of the therapeutic nature of self disclosure. On the other hand, there is also an increase in *inhibition* (7.1% increase) which relates to the restraint and self-consciousness around disclosing about a stigmatized condition (schizophrenia diagnosis) on a public social platform. For instance: *"Awkwardly waiting"*. Moving to the set of perception attributes, we observe that a majority of the measures show a decrease (e.g., *hear, feel, percept, insight*), characteristic of the emergence of a personal narrative writing style following a disclosure [52].

[3] http://mallet.cs.umass.edu/

Table 4. Differences in psycholinguistic measures between the *BD* and *AD* phases, based on Wilcoxon signed rank tests. Only significant measures, following Bonferroni correction, are included.

LIWC	BD	AD	*t*	*p*	Mean diff
Affective attributes					
Positive Affect	0.0410	0.0400	3048.0	***	-0.0243
Negative Affect	0.0095	0.0100	1832.0	***	0.0569
Anger	0.0124	0.0130	3086.0	***	0.0475
Anxiety	0.0029	0.0029	2086.0	***	0.0043
Sadness	0.0046	0.0048	1591.0	***	0.0298
Swear	0.0083	0.0077	904.0	***	-0.0715
Cognitive attributes					
Cognition					
Cognitive mech	0.0986	0.0963	3007.0	***	-0.0232
Inhibition	0.0035	0.0037	335.0	***	0.0710
Causation	0.0098	0.0093	508.0	***	-0.0497
Certainty	0.0094	0.0098	1003.0	***	0.0412
Negation	0.0162	0.0150	2888.0	***	-0.0696
Tentativeness	0.0150	0.0139	1836.0	***	-0.0769
Perception					
See	0.0076	0.0076	894.0	***	0.0023
Hear	0.0054	0.0051	1711.0	***	-0.0604
Feel	0.0058	0.0055	1864.0	***	-0.0538
Percept	0.0203	0.0194	3249.0	**	-0.0434
Insight	0.0150	0.0144	282.0	***	-0.0413
Relative	0.0869	0.0924	136.0	***	0.0637
Temporal References					
Past Tense	0.0194	0.0187	3600.0	*	-0.0338
Present Tense	0.0774	0.0739	436.0	***	-0.0453
Future Tense	0.0067	0.0068	1546.0	***	0.0034

LIWC	BD	AD	*t*	*p*	Mean diff
Lexical Density and Awareness					
Auxiliary Verbs	0.0681	0.0670	2205.0	***	-0.0158
Preposition	0.0725	0.0780	138.0	***	0.0750
Adverbs	0.0346	0.0357	905.0	***	0.0322
Verbs	0.1095	0.1076	1930.0	***	-0.0171
Article	0.0335	0.0335	566.0	***	0.0014
conjunction	0.0317	0.0313	2167.0	***	-0.0127
Inclusive	0.0212	0.0217	2801.0	***	0.0241
Exclusive	0.0173	0.0160	1800.0	***	-0.0739
Social/Personal Concerns					
Money	0.0037	0.0037	378.0	***	0.0125
Humans	0.0078	0.0076	618.0	***	-0.0299
Home	0.0030	0.0028	1535.0	***	-0.0582
Religion	0.0021	0.0022	1459.0	***	0.0344
Health	0.0081	0.0069	1792.0	***	-0.1478
Bio	0.0302	0.0286	592.0	***	-0.0551
Social	0.0670	0.0639	2826.0	***	-0.0469
Body	0.0109	0.0100	419.0	***	-0.0856
Death	0.0027	0.0029	3120.0	*	0.0887
Friends	0.0017	0.0015	1570.0	***	-0.1073
Achievement	0.0096	0.0098	458.0	***	0.0159
Work	0.0267	0.0273	2863.0	***	0.0246
Interpersonal focus					
1st p. singular	0.0347	0.0318	2096.0	***	-0.0853
2nd p.	0.0173	0.0165	3267.0	**	-0.0492
3rd p.	0.0064	0.0062	1414.0	***	-0.0288
1st p. plural	0.0029	0.0030	1369.0	***	0.0324
indefinite pronoun	0.0311	0.0306	3371.0	**	-0.0142

Linguistic style attributes. Finally, among the linguistic style attributes, we see several significant changes around the disclosure. Notably, the variations in pronoun usage before and after disclosure (*1st person singular, 1st person plural, 2nd person* and *3rd person*) reflect a transformation in the way people think about themselves in relation to others and the world. Following disclosure, individuals tend to show reduced self-attentional focus (mean difference for 1st pp. singular is -8.5%) as well as lowered social interactivity and orientation, as indicated by reduced usage of 2nd and 3rd pp (4.9% and 2.8% decreases respectively). Reduction in self preoccupation is a known attribute of improved psychological functioning [53]. Through greater use of 1st p. plural (mean difference 3.2%), we observe the emergence of a collective identity succeeding disclosures, which prior work has observed to be linked to therapeutic outcomes following psychological crises [64].

Among the lexical density and awareness attributes, there is an increase in the usage of *articles* (mean difference .14%) which characterizes categorization or concrete thinking. In turn, this change is known to bear therapeutic implications—in schizophrenia sufferers disorder in processing concepts obstructs concrete thinking [41]. In general, other measures within this category also indicate a significant increase, such as *prepositions, adverbs and inclusive* words, which together demonstrate increase in lexical density of the language of Twitter posts in the disclosing users. Language framing limitations are linked to poor cognitive functioning and coherence in individuals with mental illness [38]; therefore an increase in lexical density is likely to be indicative of therapeutic outcomes. Considering the measures of temporal references, we notice increased future orientation through the use of *future tense*; at the same time, reduced focus on the here and now, as revealed in the use of *present tense* (4.5% decrease).

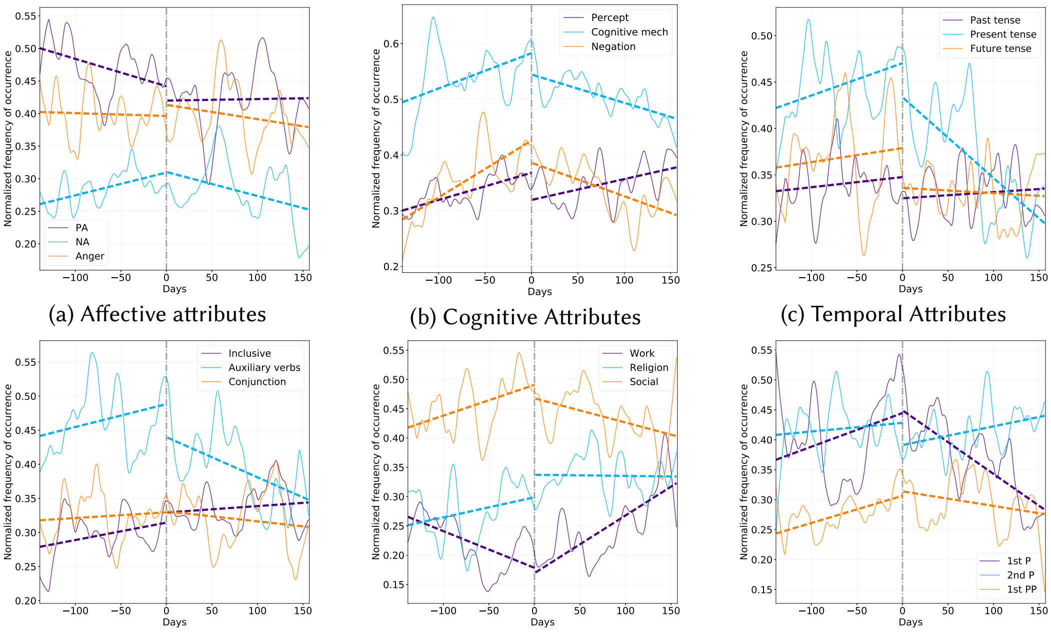
Social/ Personal concerns. Among the attributes of Social and Personal concerns, an increase in usage of *achievement* words (mean difference 1.5%) indicates improved self esteem following engaging in disclosure of a stigmatized illness like schizophrenia. E.g., consider the tweet: *"I tried"*, *"Mission success!"*. Additionally, according to the social cognitive behavioral model [47], "mastery experience", which involves providing an individual with ample opportunities to succeed, is an underlying positive health behavior change mechanism. This change may also show a tendency of the users to work towards goal-oriented activities following disclosure, which is often attributed to be a reduction in the symptoms of schizophrenia [41]. Further, there is a decrease in the *health*, *body* and *bio* categories (mean differences -14.7%, -8.5%, -5.5% respectively), signaling reduced self-consciousness of their wellness status or perceptions of their physical health. Reduced cognition of these topics is linked to improved therapeutics in individuals with mental illnesses [59]. Next, reduced use of *death* words indicates improved self-efficacy and the evolution of more positive attitudes towards life [68]. Finally, a negative mean difference in the usage of *social*, *home* and *friends* words (mean difference = -0.0469, -5.8%, -0.1073 respectively) reflects a detachment and isolation from the social realm after the disclosure, as also revealed earlier in the lowered use of 2nd and 3rd person pronouns. This may indicate a desire for the users to engage in solitude perhaps due to disclosing a stigmatized condition.

Temporal Changes. Next, we provide finer grained temporal analyses of these psycholinguistic measures around the disclosure events. In Figure 2, we show the mean time series distribution of three psycholinguistic measures from each category; we pick a sample of the statistically significant measures. We overlay these time series data with their respective linear trends (based on a fitting polynomial models of degree 1).

Despite an overall decrease and overall increase in *positive affect* and *negative affect* respectively after disclosure, the temporal analysis shows an increasing trend in *positive affect* and decreasing trend in *negative affect* over time. This improvement in affect over time is identified as one of the long-term health benefits of self-disclosure. Additionally, schizophrenia sufferers are characterized by the inability to experience pleasure [4], and therefore an increasing trend in *positive affect* may indicate a reduction in anhedonia, and improvement in overall functioning. Similarly, the increasing trend in usage of *work* related words also finds place as a long term behavioral outcome of expressive writing—following a sensitive disclosure, reference to multifaceted topics spanning one's everyday life is a known feature [4]. Further, we observe an increasing trend in first person plural pronouns, which relates to prior findings that among members of stigmatized groups, writing about being a group member changes the sense of self worth one derives from group membership. Finally, the decreasing trend in *auxiliary verbs* and self referential pronouns is indicative of lower self preoccupation.

5.2 Changes in Linguistic Structures

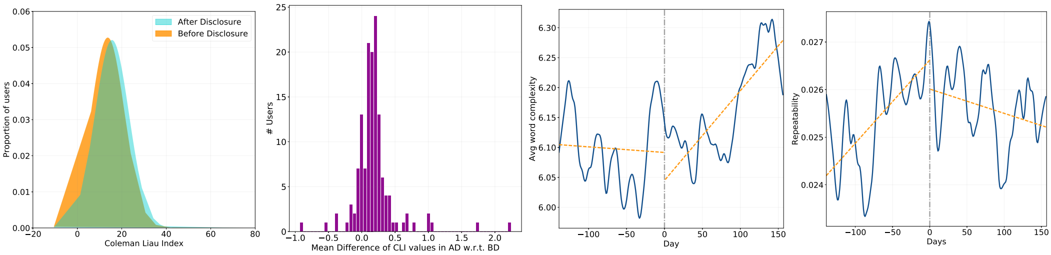
To characterize structural differences in the language of Twitter posts before and after the disclosure (i.e., the *BD* and *AD* phases), we present an analysis of the three measures, *readability, word complexity*, and *word repeatability*. Figure 3(a) shows a distribution of the mean difference in readability scores (AD compared to BD) over number of users, as measured by the Coleman-Liau index; individual distributions of users over CLI scores in the *BD* and *AD* phases are also shown in Figure 3(b). An overall positive mean difference with an average of 0.18 (σ = 0.32) is observed



(a) Affective attributes (b) Cognitive Attributes (c) Temporal Attributes

(d) Lexical density (e) Social/Personal Concerns (f) Interpersonal focus

Fig. 2. Time series distribution and trend (based on fitting a linear model) of psycholinguistic attributes spanning the *BD* and *AD* phases. Selected statistically significant measures per Table 4 are shown. 0 indicates disclosure date.



(a) (b) (c) (d)

Fig. 3. (a) Distribution of CLI scores (readability) over number of users in the *BD* and *AD* phases. (b) Distribution of mean differences in CLI index (readability) in the *AD* phase, compared to the *BD* period. (c) Temporal changes and linear trend in the word complexity measure in the *AD* phase, compared to the *BD* phase. (d) Temporal changes and linear trend in word repeatability in the *BD* and *AD* phases. 0 indicates disclosure date.

indicating an increase in readability after disclosing regarding one's diagnosis of schizophrenia. More elaborately, we find that overall 80% of the users show a mean CLI difference value greater than 0, indicating that largely, disclosing users show an improvement in the language framing limitations characteristic of people with schizophrenia, which is considered to be a therapeutic change [4]. In other words, literature suggests that reorganizing and structuring traumatic memories or experiences helps in developing a complex and coherent narrative [56], suggesting that with these observed changes in our readability measure, disclosing users show evidence of reduced sentence framing limitations in the *AD* period.

Next, there is an increasing trend observed in word complexity as measured by the normalized length of words in the users Twitter posts (mean difference 0.01). This signifies an increase in usage of complex words and sentence verbosity moving away from the stereotypy symptoms of schizophrenia. Prior work says that people with the schizophrenia illness are limited in their ability to think with any degree of complexity. They are able to think in very simple terms, but generally are unable to solve complex problems, plan ahead or organize their thoughts [4]. The evidence of emergence of complexity in the language of Twitter users, therefore, reveals a reduction of an important negative symptom of schizophrenia. Finally, repeatability in terms of the proportion of non-unique words in Twitter posts has a decreasing trend after disclosure. Individuals suffering from schizophrenia often have repetitive thoughts that interfere with their ability to think, per the socio-cognitive model of schizophrenia [47]. Reduction of word repeatability is, therefore, likely indicative of lesser word repetitions and better articulation via language, as well as more concrete thinking and functioning among the disclosing users, a finding also observed in the case of the psycholinguistic measures [37].

5.3 Changes in Domain-Specific Content Measures

We observe considerable differences in the usage of domain relevant lexicon words as shown in Table 5. Among the 1981 tokens which were extracted based on the technique described in the Methods section, the most distinctive ones are tokens that were used only before or only after the disclosure. From Table 5, we draw several interesting observations. The usage of tokens related to symptoms and medication for schizophrenia such as *'experience hallucinations'*, *'voices really'*, *'meds work'* primarily appear only during the BD phase. Hallucinations, delusions, and paranoia are among the most distinctive negative symptoms of schizophrenia [41]. Usage of these tokens in the posts of the disclosing users, such as, *"i never sleep alone, hallucinations are troubling me"* and *"I miss hearing voices that tell me to stand in the rain at five in the morning"* reveal that prior to the disclosure, the users in our dataset were appropriating social media to engage in discourse on these topics and personal experiences.

Table 5. Domain-specific (Reddit) token usage during the *BD* and *AD* phases. The second column from the left lists tokens found more frequently in the *BD* phase than the *AD* period; the third column from the left lists the converse.

Tokens used only *BD*	Tokens *BD* > *AD*	Tokens *AD* > *BD*	Tokens used only *AD*
anxiety medication, coping mechanisms, experience hallucinations, formally diagnosed, meds work, really struggling, voices really, trouble remembering, believe mental, need medication, mind racing, police called	suicide attempt, seek professional, people watching, visual hallucination, voice inside, mentally unstable, new meds, cameras, feel lonely, triggering, really anxious, withdrawal symptoms	therapy, socially awkward, survived, fighting, isolation, doctor appointments, warning signs, drugs, socialize, counseling, hospitalization, cope	mentally healthy, john nash, going doctor, inpatient, new therapist, rehabilitation, self care, seeking help, depersonalization, positive symptoms, feel scared, prognosis

Whereas, tokens related to treatment or help (*'going doctor'*, *'inpatient'*, *'seeking help'*), self-care (*'rehabilitation'*, *'self care'*) appear only after the disclosure. This indicates that, following disclosure, the users feel comfortable and less restrained in talking about their treatment experiences around schizophrenia; e.g., *"Everyone who was an inpatient with me at the hospital has moved on.."*. At the same time, tokens around help seeking and self-care show a shifted attitude of the disclosing

Table 6. Theme keywords derived from topic modeling and human annotation analysis.

Theme	Topics	keywords
Mental illnesses	Topic 5, 12	mental, fighters, stigma, people, health, hospital, crazy, problems, doctor, illness, schizophrenia, donate, pndchat, pndhour, support, amazing, work, pnd
Symptoms	Topics 15,2,22,3,20	r/paranormal, r/ufos, r/creepy, ufo, house, ghost, ass, shit, fuck, lol, dick, fuck, shit, people, hate, stop, life, stupid, talking, god, hell, damn, holy, friends, anymore, love, jesus, hell, world, real, angel, christ, heaven, lord, soul, trust, fight, bless, sleep, night, bed, day, tomorrow, morning, work, time, tonight, hours, asleep, tired, today, sleeping, wake
Functioning	Topics 4, 7, 9, 10,	time, day, years, today, happy, week, times, past, months, minutes, eyes, face, back, hand, head, dear, lips, touch, felt, smile, deep, pain, care, anymore, people, time, whats, wrong, make, feel, hurt, wanna, love, talk, life, live, world, die, heart, mind, time, real, thing, true, words, end, rest, dream
Stigma	Topic 5	today, mental, fighters, stigma, people, drugs, health, days, hospital, left, bad, crazy, lot, real, problems, doctor, profit, illness, schizophrenia, donate

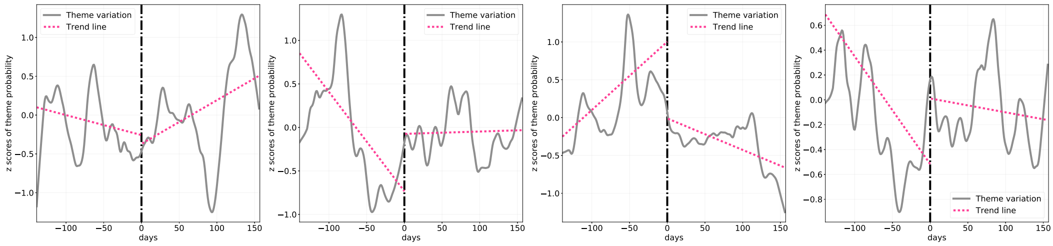
users to appropriate the Twitter platform for support as well as to share their desire for improved self-efficacy and recovery from the debilitating effects of schizophrenia: *"today was a nice self care day tbh i slept a lot and for the most part kept my mind off of stressful things"*.

Considering the tokens that appear during both the *BD* and *AD* phases, those related to therapy, relapse, counseling and coping have an increased usage after the disclosure. After having disclosed about a sensitive condition like schizophrenia on Twitter, users might be appropriating the platform as a way to document their and self-initiated strategies on coping, as well as information about their treatment and therapy experiences with a broader and richer audience: *"I think treatment is important in providing better coping skills."*, *"I need today's therapy session badly. Yay Medicare"*. On the other hand, tokens relevant to symptoms (*'cameras'*, *'voice inside'*, *'visual hallucinations'*) are used more often during the BD phase than the AD phase. For instance, we observe the following post excerpts in our BD phase data: *"my visual hallucinations take form of random shadows, usually person- or insect-like!"*. Overall, the usage of symptoms words largely during the BD phase and the usage of hospitalization words extensively during the AD phase affirms our expectation that the premorbid and active phases occur before an individual self discloses their formal diagnosis on social media. Finally, the increased focus on well-being as observed by the usage of coping, psychotherapy and treatment related tokens portrays a transformation in the health behavior of disclosing individuals.

5.4 Changes in Topical Measures

To understand the broader themes revealed by language around the disclosure events, we now present the results of the topic modeling approach, including the two measures derived from it—*thematic variation* and *topical coherence*. To start with, Table 6 shows four major themes that appeared from the semi-open coding task involving two clinical psychiatrist annotators, the set of topics that define the theme and the most contributing words in each theme. The themes primarily appear to revolve around the clinical attributes of schizophrenia and are prevalent in both the *BD* and *AD* phases. However, from Figure 4 we observe significant shifts across the two time periods around the disclosure event.

First, the theme "Symptoms" includes words such as */r/paranormal*, */r/ufos* and */r/creepy* which are Reddit communities for discussions about paranormal thoughts and activities. Together with terms like *ghost*, *ufo* and *house*, these words capture disorganized thinking and delusional attitudes



(a) Symptoms (b) Mental Illness (c) Functioning (d) Stigma

Fig. 4. Temporal variation (z-scores) of theme probabilities during the *BD* and *AD* phases. The time series are derived per each day by computing the average posterior probability of topics that are contained in a theme, spanning all posts of all disclosing users. To each time series, linear trends are obtained by a fitting a linear mode. 0 indicates disclosure date.

which are notable markers of schizophrenia; e.g., as demonstrated in this tweet: *"An orange 'UFO' story from /r/Paranormal"*. Additionally, terms related to sleeplessness like *tired*, *sleep*, *waking* appearing in tweets like *"I'm tired, in pain, and cranky. Someone please make it stop, I swear to God."* are also established negative symptoms of schizophrenia. In fact, sleep disturbances and exhaustion, fatigue have significant impact on quality of life in individuals with schizophrenia [38]. Next, words like *jesus*, *god*, *holy*, *angel*, *heaven* reveal spirituality and religiousness, which are notable in schizophrenia sufferers [37]. Figure 4(a) gives some interesting insights into this theme's trends over the *BD* and the *AD* phases. We notice that shortly prior to the disclosure event, there is a reduction in the discussion of the theme, and it persists to be low for sometime in the *AD* phase as well. It may indicate improved functioning of the disclosing users; absence of the negative symptoms of schizophrenia are known to be linked to therapeutic outcomes [41].

The next theme given by our topic modeling and human annotation method is on "Functioning". This includes words like *happy, touch, smile, pain, hurt, die, care, wrong*; an example tweet says: *"im in a bad mood like im ready to either hurt myself or someone else"*. Typically, the socio-cognitive model of schizophrenia [47] indicates that reduced functioning, as is indicated by the words in this theme, is an important attribute of the schizophrenia experience, including behaviors such as neglect of social, emotional, physical, and cognitive aspects of life, as well as a lack in overall sense of purpose in an individual. We note that, from Figure 4, during the *BD* period, the presence of these functioning related topics shows a monotonically increasing trend. However, following disclosure, we observe that the disclosing users tend to share less about content related to functioning, resulting in a decreasing trend. This shows improved therapeutic outcomes in the cohort of the disclosing individuals. It may also indicate that the users feel lowered inhibition and restraint following self disclosure, a finding that also aligns with the results from our psycholinguistic analysis above.

More generally, beyond schizophrenia related themes, we also observe the disclosing users to share content about other mental illnesses on Twitter such as around hospital visits, both during the *BD* and the *AD* phases. This might demonstrate comorbidity in the user group or expression of group identity related to stigmatized mental health conditions. For example, the words *pnd*, *pndhour*, *pndchat, stigma* point to a support community of people affected by postnatal depression. This theme shows a noticeable dip prior to the disclosure event, and then shows a stable, but overall reduced activity in the *AD* period. The higher values of this theme in the *BD* period might be due to the presence of the premorbid/active phases of schizophrenia prior to the disclosure event.

Finally, "Stigma" appears as a theme in itself comprising terms related to fighting the stigma of experiences of schizophrenia—*fighters, hospital, bad, doctor, illness*, also demonstrated in tweets like:

"my schizophrenia has gotten worse ever since I started living alone". From Figure 4 we observe lowered stigma leading up to the disclosure event, although there is a peak right after the disclosure, which likely conveys the difficulty in disclosing about one's stigmatized conditions like schizophrenia. However, in general, the Stigma theme shows persistent but lower activity in the *AD* phase indicating that the users are less concerned about these issues, and also characteristic of "opening up" as observed in the expressive writing literature [53].

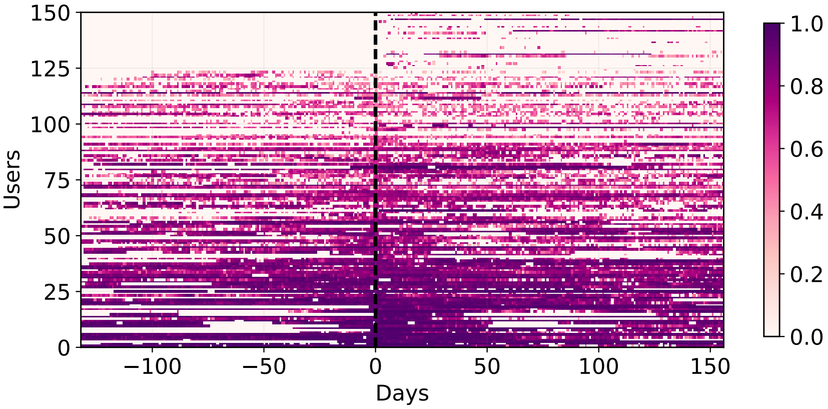


Fig. 5. Changes in topical coherence in the *BD* and *AD* phases. Higher intensity cells in the heatmap indicate higher topical coherence. 0 indicates disclosure date.

Finally, to examine changes in topical coherence before and after disclosure, the average coherence value per user, per day is plotted as a heatmap in Figure 5, where time is indicated along the x-axis while the y-axis indicates unique disclosing users. The color intensity of each cell on the heatmap indicates the topical coherence by user y on day x i.e. how topically coherent the posts made by user y on day x were with respect to posts made by y from days $x - 1$ through $x - 7$. Recall that topical coherence was characterized by the cosine similarity measure. Therefore, a higher value of topical coherence (revealed by higher intensity cells in the heatmap) indicates that content shared on day x is coherent with respect to the same spanning the week before. The figure reveals a gradual increase in topical coherence for most users following their disclosure of the schizophrenia diagnosis or experience. This result is also found in psychotherapy literature [65, 68] and the expressive writing paradigm [52], where self disclosure is known to help people organize and remember events in a coherent fashion while integrating their thoughts and feelings. The therapeutic nature of self disclosure are noted by Joinson [33] also identifies an increase in coherency and articulation after disclosing and confronting difficult experiences.

5.5 Assessing Psychiatric Significance of Findings

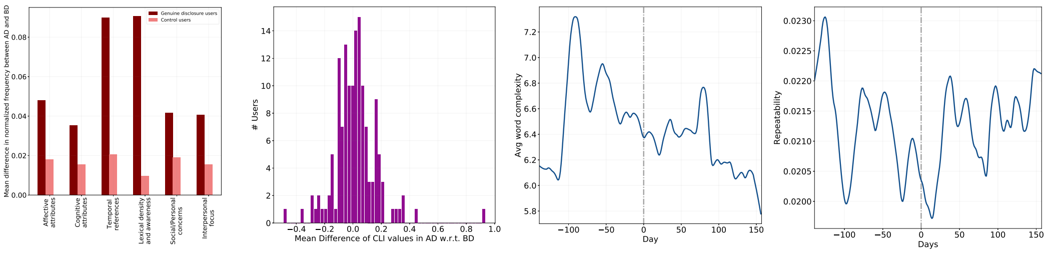
In the absence of ground truth assessments on whether our observed linguistic changes indicate *psychiatrically valid* therapeutic outcomes, we employed a qualitative assessment framework. Specifically, following each analysis described above, we shared the results with our psychiatrist partners to gain qualitative grounding on their significance, per the theoretical (e.g., the socio-cognitive model [47]) as well as psychiatry services literature [59]. Then, we iteratively adapted our analytical approach to ensure that the observations gleaned from the analyses corroborate what is known about the manifestation and treatment of the illness. Our clinical collaborators indicated the linguistic changes to indeed be indicative of therapeutic outcomes they would expect to see in an individual who was on initial stages of recovery or symptomatic remission. Such methods, that engage evaluators as collaborators in the analytical process itself, have been suggested in the clinical literature [31] to validate outcomes of field trials where obtaining post-hoc objective evaluations may not be possible.

5.6 Comparison with Matched Controls

Finally, in order to validate and to establish causality surrounding the above observed linguistic changes observed in genuine disclosures before and after the disclosure, we present comparisons to our matched control user group. Figure 6(a) shows the mean relative differences of psycholinguistic attributes for the genuine disclosure and control users, spanning the BD and AD phases. Over all the categories, we observe a greater change in psycholinguistic measures for the disclosed user group as compared to the control group. For example, lexical density and awareness attributes (e.g., adverbs, auxiliary verbs, prepositions), affective attributes, and attributes of interpersonal focus (e.g., first person singular) showed the largest change in the *AD* phase compared to the *BD* phase for the genuine disclosure group; however for the control group, the change across the time phases was minimal. In fact, based on independent sample t-tests (that adopted Bonferroni correction), we observe that the changes in the case of the genuine disclosure group were statistically significant across the different attributes, compared to the control cohort ($p < 10^{-15}$).

Further, we notice minimal changes in the linguistic structure measures for the control group. The mean difference in readability (CLI measure) between *AD* and *BD* phases was 0.18 for the genuine disclosure group; whereas, the control group had a difference of 0.018. This is shown in Figure 6(b). Similarly, changes in word complexity, and word repeatability measures for genuine disclosure group were 5%, 22% greater in magnitude when compared to the control group.

Together, these observations show that the patterns of linguistic differences we observe in the case of the individuals disclosing schizophrenia on Twitter, can be attributed to the disclosure event itself, since such changes are absent in individuals who do not engage in a similar disclosure.



(a) (b) (c) (d)

Fig. 6. (a) Comparison of LIWC differences BD, AD of genuine disclosure users with matched control users. (b) Distribution of mean differences in CLI index (readability) in the *AD* phase, compared to the *BD* period (c) Temporal changes in the word complexity measure in the *AD* phase, compared to the *BD* phase (d) Temporal changes in word repeatability in *BD* and *AD* phases. 0 indicates disclosure date for the matched users.

6 DISCUSSION

Our findings provide some of the first insights into the therapeutic outcomes of self disclosures made on Twitter, particularly around schizophrenia. With these insights, we are able to draw implications for multiple areas, spanning theoretical and therapeutic ones, as well as design considerations for developing tools that can support this new form of broadcasting self-disclosure practices on social media. We discuss these in the following subsections.

6.1 Theoretical and Therapeutic Implications

With increasing number of individuals adopting social media platforms for support seeking and disclosure of stigmatized conditions, we discuss the implications of these broadcasting disclosures with respect to existing dyadic disclosures in a therapist-client setting. We focus on the broader question: How do the disclosure goals transform when dyadic disclosures become broadcast disclosures? According to the functional theory of self disclosure, proposed by Derlega and Grzelak, disclosure goals or subjective reasons for self-disclosing activate disclosure decision-making process and shape its content. They say, "[...] if we wish to understand and predict individuals' self-disclosing behavior, we must identify (and measure) the major sources of value that self-disclosure has for individuals" [22]. As also noted before, social validation, self-expression, relational development, identity clarification, therapeutic benefits and social control are some of the prominent goals identified by prior work [8, 34].

In our study, we focused on one of these prominent goals i.e. therapeutic benefits. By analyzing the linguistic content shared around schizophrenia disclosures on Twitter, we have found that therapeutic benefits, are apparent even in broadcasting disclosures shared via a public social media platform like Twitter. For instance, we found that there are long term trends indicative of reduction in the negative syndromes of the condition, such as decreasing negative affect, increasing positive affect, greater future orientation and reduced self preoccupation. Moreover, as characterized by the linguistic structural measures, we observed a considerable improvement in readability of language manifested in the disclosers' Twitter posts. These results find place in several prior studies on the "opening up" phenomenon [52], providing ground for their interpretation as therapeutic outcomes.

Expressive Writing on Twitter: Venting out Emotions. To discuss this thought-provoking question as an implication of our work, we draw on the literature on the expressive writing paradigm [53] , which suggests that venting out negative emotions, confronting inhibited thoughts are among the reasons why disclosure, beyond the dyadic therapist-client setting has therapeutic benefits. Moreover, giving an experience structure and meaning is known to make it more manageable and facilitate a sense of resolution [52] . This leads us to ponder whether social media platforms like Twitter are being adopted by certain individuals as an expressive writing platform, to vent out and give structure to stigmatized experiences. This is supported by the empirical findings of our work: we observe that "venting out" via self-disclosures of schizophrenia are followed by improved topical coherence, increased focus on well-being and self-care, and reduced expression of symptoms and stigma perceptions. In fact, recent research in social computing supports this conjecture of the "venting out" phenomenon on social media; by disclosing one's deepest thoughts and feelings on social media, one can suppress and inhibit dysfunctional negative thoughts. For instance, it has been found that individuals struggling with negative emotions like depressive thoughts appropriate Instagram for emotional release [3]. People with pro-eating disorder behaviors engage with sensitive communities on Instagram and Tumblr to find a "safety valve" of negative behavior, and thereby to engage in dis-inhibiting discourse that can eventually prevent dangerous or adverse health outcomes [13, 51]. Thus, despite not having the emotional closeness and intimacy of a dyadic disclosure setting, social media, by providing a space for articulation of thoughts and emotions around stigmatized experiences, appears to be supporting unintended therapeutic outcomes.

Expressive Writing on Twitter: Connecting with the Invisible Audience. Further, recall one of the annotated Twitter posts given in Table 3: "*Twitter is an acceptable way to talk to yourself without being diagnosed schizophrenic*". The post was annotated by the two clinical psychiatrist raters to be a disingenuous statement and therefore not included as a genuine schizophrenia disclosure. Nevertheless, it points to the general perception about the nature of Twitter as a space for self-expression, even in the face of an "invisible" audience, which is considerably larger than offline or dyadic audiences prevalent in the therapist-client setting, as well as likely to be anonymous or pseudonymous [40]. Interestingly, the audience quality or the stark distinctions in *who* listens to these broadcasting disclosures seems to make little difference to the disclosers. This may explain why we see them to be associated with improved therapeutic outcomes— such as, improved mood, increased readability and coherence in language, or lower self preoccupation. Moving from identified, dyadic, private settings that have characterized traditional therapeutic disclosures, some individuals thus seem to have adapted Twitter's social norms of publicly broadcasting goings on of daily lives to invisible audiences, to derive therapeutic benefits.

An alternative explanation could also be that the disclosers draw unique benefits from this invisible audience: It is known that communicating with an invisible audience underlies the "disinhibition effect", a tendency to engage in more candid disclosures around sensitive topics than is possible in front of an identified audience [34]. And that online disclosure may be framed as an interpersonal process, in which people regulate their disclosures based on what the invisible audience chooses to disclose about them [39]. Our conjecture is supported by the observation that after the disclosure events, individuals tend to discuss more frequently about stigma related issues and consistently about mental illness topics, both of which are known indicators of reduced inhibition or self-restraint [23].

Expressive Writing on Twitter: A New Model of Writing Therapy. Despite the empirical evidence we gather from this work, surrounding the appropriation of Twitter for schizophrenia disclosure, we note a sharp contrast in the affordances provided by Twitter as a platform of self-expression, and the paradigm adopted in expressive writing therapy. In the traditional expressive writing paradigm, clients engage in long-form writing privately, pouring down their thoughts vividly in an elaborate manner, where often multiple diverse topics blend together [62]. On the other hand, Twitter is spontaneous, real-time, and largely a public place for sharing everyday musings in somebody's life, and it lets people do so in 140 characters or fewer at a time. Appropriating Twitter for expressive writing implies that self-disclosers' have identified mechanisms to adapt these in-the-moment writing forms to meaningfully communicate a variety of simple or complex, and abstract or concrete thoughts and emotions that characterize the experience of a stigmatized illness like schizophrenia [30]. In fact, "poverty of speech" is a known negative symptom of the illness [41]: therefore, we hypothesize that the brevity of expression enforced by Twitter might be supporting individuals to communicate those underlying feelings that might otherwise be challenging to express in long-form writing therapy settings.

The presence of the observed therapeutic benefits on Twitter, despite the stark affordances and norms of use of the platform, extends existing discussions in recent research [69]: that the dichotomy between offline and online expression, and its role in enabling candid self-disclosures, whether in the dyadic (private) or the broadcasting (public) form, might be blurring after all. Or that, the disclosers' are inventing new opportunities to derive therapeutic benefits from short-form, spontaneous blurbs shared on public social media platforms, going beyond the structure of offline dyadic therapist-client settings.

Nevertheless, our work also raises important questions around the therapeutic efficacy of social media as an expressive writing paradigm. In this work, while we found that self-disclosures of schizophrenia on Twitter can lead to distinct linguistic changes indicative of therapeutic benefits, we did not assess the responses that these disclosures elicited from the disclosers' social networks. Future work could examine such responses in order to further assess the utility of social media platforms as promising tools for writing therapy. Importantly, *why* individuals might be appropriating public social media as an expressive writing paradigm, and if there are conscious purposes and therapeutic goals underlying their decisions remain interesting directions for future work.

Finally, one of the reasons the expressive writing phenomenon leads to improved mental well-being is because it enables disclosers' develop intimacy and pursue relationship building goals with the audience [53], which, in most cases, is a therapist. Given the distinct affordances and norms of use of social media platforms, how do disclosers' appropriate these goals, where the audience constitutes simply lay individuals whose identities are largely unknown? Future work could explore these open questions.

6.2 Design Implications

Building on our findings that indicate Twitter's function and appropriation as a therapeutic platform to individuals with schizophrenia, we now discuss some design directions. In an interpersonal setting of dyadic disclosures, situational cues like personality characteristics, intimacy, and non verbal communication are identified as activators of disclosure goals [48]. In the absence of such cues in a social media setting, design affordances of an online platform can assist as activators and supporters of disclosure goals, around therapeutic benefits. We present two design directions:

6.2.1 Journaling Tools. As discussed previously, one of the potential reasons behind the observed therapeutic outcomes following disclosures on social media appears to be its affordance as a platform of self expression, whether for venting negative emotions or for connecting with an invisible audience. We therefore envision that journaling tools may be built and integrated directly or indirectly with social media platforms, wherein posts preceding and succeeding disclosures could be logged voluntarily, serving as a timestamped and archive of one's thoughts, feelings, and experiences around conditions like schizophrenia. The psychotherapy literature has identified unique benefits of such archival of writing. For instance, it can help an individual develop a logical narrative of events, experiences, and mental health challenges, and thereby enable them to meet self-care and coping goals [56]. Journaling tools integrated with social media can further allow disclosers' to be self-reflective and more empowered: since the archives are likely to draw upon imagination and creativity over a period of time, they can help individuals become more knowledgeable about themselves and to increase their sense of agency.

The archives logged through this journaling tool can also complement counseling efforts. For instance, they can be augmented with the various linguistic measures we utilized in this paper: linguistic structural attributes, topical coherence, psycholinguistic attributes and so on, which can reveal longitudinal trends related to specific markers of well-being around disclosure. When shared with a therapist, the archives can aid them "by entering the client's mental constructs via the written word" [5], or by understanding those thoughts and feelings which the client might be "unable to vocalise" [16].

6.2.2 Social Support Recommendations. Recall that one of the most contrasting aspects between dyadic disclosures in a therapist-client setting and broadcasting disclosures in a social media setting, is the lack of intimacy and inter-personal closeness in the latter. Unlike support communities like on Reddit, microblogging platforms like Twitter lack explicit design features to foster communal relationships. Features like hashtag and mentions are being adopted to indicate solidarity or group membership (like *#pndchat* observed in our data analysis). However, more directed support around disclosure can help individuals with mental health challenges to cope and identify with peers with similar conditions. For instance, tools can be built which align with the social media disclosure events of different individuals, and then algorithmically recommend social connections. These recommendations can further be boosted by leveraging our linguistic measures, such that disclosing individuals with similar mental health experiences (symptoms, stigma perceptions) are more likely to be connected. Additionally, the recommendations can also include pointers to help resources, such as ways to avail online therapy, pop-ups to reach out to a friend, or a self-care expert. We conjecture that with these support recommendation tools, mental illness disclosers' on social media will be able to gain therapeutic benefits collaboratively [39] and reciprocally [3].

6.3 Privacy and Ethical Considerations

Given the sensitive nature of the topic of investigation in this paper, we discuss our privacy and ethical considerations. First, the data used for our study is publicly available and we do not interact with the users; therefore it did not qualify for approval from our respective Institutional Review Boards. However, without the users' consent, knowledge, or awareness, we are cognizant of the ethical limitations that occur in the absence of consent and feedback from the study population. To reduce risk of users' identity and data being revealed inadvertently, we paraphrased quotes in the paper, obfuscated any personally identifiable information, a method that has been used in other similar social computing work [51].

We also note that disclosing about stigmatized concerns like schizophrenia might call upon negative impacts such as social discrimination and rejection, which are detrimental to well-being. Therefore, the support recommendations discussed under design implications need to be cognizant of the boundary regulation choices of the disclosing individuals [69], e.g., restricting recommendations to a chosen audience of the discloser, to prevent unintended negative consequences. Ideally, they also need to adapt to the responses that disclosures may elicit from an individual's social network, so that the amount of disclosure information revealed is adequate to gather support, however does not divulge excessive details about the user. Similarly, for the journaling tool, we indicated the possibility of sharing of archives with a therapist. These design approaches need to factor in boundary regulation issues in the patient-therapist interpersonal relationship, and need to develop adequate data and informational abstractions and curation methods. This would allow balancing the disclosers' clinical needs and their privacy expectations, attending to their privacy concerns at the forefront.

6.4 Limitations and Future Work

There are some limitations to our work. We acknowledge that our findings are limited by our data acquisition capabilities. For our study, we relied on a set of hand curated key phrases to assist in data collection. Although these phrases are clinician validated, they do not include all possible ways in which Twitter users disclose their diagnosis of schizophrenia.

Further, we note that the nature of self disclosures on a micro-blogging platform like Twitter in itself will be very different from that made on online communities platforms like Reddit, social networks like Facebook and so on. Future work can seek to unravel these nuances of platform-specific disclosures. In terms of our findings, we also acknowledge the small size of our data sample of users with genuine disclosures. However, since our findings on this sample, closely align with observations in the literature of self disclosure and the clinical characteristics of schizophrenia, it shows promise about the generalizability of our methods to larger populations.

We also note that while comparison with a matched control group allows us to establish a weak causation, further work is needed to examine to what extent the linguistic changes and their therapeutic attributes, that ensue social media disclosures, align with clinical assessments of improvement in an individual's well-being or actual remission of the illness. Complementary information on clinical treatment or psychotherapy received by the disclosed individuals in the offline world could help investigate these questions in future work.

7 CONCLUSION

In this paper, we provided some of the first empirical insights into the therapeutic outcomes of social media disclosures of schizophrenia. Starting with an expert validated list of 146 schizophrenia disclosures shared on Twitter, we first presented a clinically grounded quantitative approach to identifying temporal phases around disclosure during which symptoms of schizophrenia are likely to be significant. Then, we defined linguistic measures to quantify behavioral changes around the disclosures. Alongside significant linguistic differences before and after the disclosure, we observed markers of therapeutic outcomes: characterized by increase in positive affect, enhanced readability, improved topical coherence, future orientation and reduced discussion on schizophrenia symptoms and associated stigma. Our work reveals the potential of social media platforms as new therapeutic tools supporting broadcasting self-disclosures.

8 ACKNOWLEDGMENTS

We thank the members of the Social Dynamics and Well-Being Lab for their valuable feedback in preparing this manuscript. Ernala and De Choudhury were partly supported by a National Institutes of Health grant #1R01GM11269701.

REFERENCES

[1] Xavier F Amador, David H Strauss, Scott A Yale, and Jack M Gorman. 1991. Awareness of illness in schizophrenia. *Schizophrenia bulletin* 17, 1 (1991), 113.

[2] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3906–3918.

[3] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: the case of# depression. In *CSCW*.

[4] Nancy C Andreasen. 1982. Negative symptoms in schizophrenia: definition and reliability. *Archives of general psychiatry* 39, 7 (1982), 784–788.

[5] K Anthony. 2000. Information Technology. Counselling in cyberspace. *COUNSELLING-RUGBY-* 11, 10 (2000), 625–627.

[6] Karen A Baikie and Kay Wilhelm. 2005. Emotional and physical health benefits of expressive writing. *Advances in psychiatric treatment* 11, 5 (2005), 338–346.

[7] Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1373–1378.

[8] Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64, 4 (2014), 635–657.

[9] Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia* 1 (2015), 15030.

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

[11] R Sherlock Campbell and James W Pennebaker. 2003. The secret life of pronouns flexibility in writing style and physical health. *Psychological science* 14, 1 (2003), 60–65.

[12] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *CSCW*. ACM, 1171–1184.

[13] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *CSCW*. ACM, 1201–1213.

[14] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. *NAACL HLT 2015* (2015), 1.

[15] Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.

[16] D Cullen. 2000. *A byte size study of online counselling: who is doing it and what is lit like.* Ph.D. Dissertation. MSc dissertation, University of Bristol.

[17] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *CSCW*. ACM, 1431–1442.

[18] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.. In *ICWSM*.

[19] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media.

[20] Munmun De Choudhury, Sanket Sharma, Tomaz Logar, Wouter Eekhout, and René Nielsen. 2017. Quantifying and Understanding Gender and Cross-Cultural Differences in Mental Health Expression via Social Media. In *CSCW*.

[21] Valerian J Derlaga and John H Berg. 2013. *Self-disclosure: Theory, research, and therapy*. Springer Science & Business Media.

[22] Valerian J Derlega, Janusz Grzelak, et al. 1979. Appropriateness of self-disclosure. *Self-disclosure: Origins, patterns, and implications of openness in interpersonal relationships* (1979), 151–176.

[23] Faith B Dickerson, Jewel Sommerville, Andrea E Origoni, Norman B Ringel, and Frederick Parente. 2002. Experiences of stigma among outpatients with schizophrenia. *Schizophrenia bulletin* 28, 1 (2002), 143–155.

[24] Barry A Farber, Kathryn C Berano, and Joseph A Capobianco. 2004. Clients' Perceptions of the Process and Consequences of Self-Disclosure in Psychotherapy. *Journal of Counseling Psychology* 51, 3 (2004), 340.

[25] E Feinstein. 1993. Muse: for ET In Sixty Women Poets. (1993).

[26] Frederick J Frese. 1998. Advocacy, recovery, and the challenges of consumerism for schizophrenia. *Psychiatric Clinics* 21, 1 (1998), 233–249.

[27] Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and Schuster.

[28] Oliver L Haimson, Jed R Brubaker, Lynn Dombrowski, and Gillian R Hayes. 2015. Disclosure, stress, and support during gender transition on Facebook. In *CSCW*. ACM, 1176–1190.

[29] James Douglas Hamilton. 1994. *Time series analysis*. Vol. 2. Princeton university press Princeton.

[30] Wolfram Hinzen and Joana Rosselló. 2015. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. *Frontiers in psychology* 6 (2015), 971.

[31] John Hunsley. 2003. Introduction to the special section on incremental validity and utility in clinical assessment. *Psychological Assessment* 15, 4 (2003), 443.

[32] Assen Jablensky, Norman Sartorius, Gunilla Ernberg, Martha Anker, Ailsa Korten, John E Cooper, Robert Day, and Aksel Bertelsen. 1992. Schizophrenia: manifestations, incidence and course in different cultures A World Health Organization Ten-Country Study. *Psychological Medicine Monograph Supplement* 20 (1992), 1–97.

[33] Adam Joinson. 1998. Causes and implications of disinhibited behavior on the Internet. (1998).

[34] Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European journal of social psychology* 31, 2 (2001), 177–192.

[35] Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the Internet. *The Oxford handbook of Internet psychology* (2007), 237–252.

[36] Sidney M Jourard. 1971. Self-disclosure: An experimental analysis of the transparent self. (1971).

[37] Stanley R Kay, Lewis A Opler, and Jean-Pierre Lindenmayer. 1988. Reliability and validity of the positive and negative syndrome scale for schizophrenics. *Psychiatry research* 23, 1 (1988), 99–110.

[38] Richard SE Keefe, Philip D Harvey, Mark F Lenzenweger, Michael Davidson, Seth H Apter, James Schmeidler, Richard C Mohs, and Kenneth L Davis. 1992. Empirical assessment of the factorial structure of clinical symptoms in schizophrenia: negative symptoms. *Psychiatry Research* 44, 2 (1992), 153–165.

[39] Airi Lampinen, Vilma Lehtinen, Asko Lehmuskallio, and Sakari Tamminen. 2011. We're in it together: interpersonal management of disclosure in social network services. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 3217–3226.

[40] Alex Leavitt. 2015. This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *CSCW*. ACM, 317–327.

[41] Jeffrey A Lieberman, Diana Perkins, Aysenil Belger, Miranda Chakos, Fred Jarskog, Kalina Boteva, and John Gilmore. 2001. The early stages of schizophrenia: speculations on pathogenesis, pathophysiology, and therapeutic approaches. *Biological psychiatry* 50, 11 (2001), 884–897.

[42] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. (2017).

[43] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[44] Mark Matthews, Elizabeth Murnane, Jaime Snyder, Shion Guha, Pamara Chang, Gavin Doherty, and Geri Gay. 2017. The double-edged sword: A mixed methods study of the interplay between bipolar disorder and technology use. *Computers in Human Behavior* 75 (2017), 288–300.

[45] Lynn C Miller, John H Berg, and Richard L Archer. 1983. Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology* 44, 6 (1983), 1234.

[46] Elizabeth L Murnane, Dan Cosley, Pamara Chang, Shion Guha, Ellen Frank, Geri Gay, and Mark Matthews. 2016. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health. *Journal of the American Medical Informatics Association* 23, 3 (2016), 477–484.

[47] John A Naslund, Kelly A Aschbrenner, Sunny Jung Kim, Gregory J Mchugo, Jürgen Unützer, Stephen J Bartels, and Lisa A Marsch. 2017. Health Behavior Models for Informing Digital Technology Interventions for Individuals With

Mental Illness. *Psychiatric rehabilitation journal* (2017).

[48] Julia Omarzu. 2000. A disclosure decision model: Determining how and when individuals will self-disclose. *Personality and Social Psychology Review* 4, 2 (2000), 174–185.

[49] Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-Depressed Users in Twitter.. In *ICWSM*.

[50] J Parnas, L Jansson, LA Sass, and P Handest. 1998. Self-experience in the prodromal phases of schizophrenia: A pilot study of first-admissions. *Neurology Psychiatry and Brain Research* 6, 2 (1998), 97–106.

[51] Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger hurts but starving works": characterizing the presentation of eating disorders online. In *CSCW*. ACM, 1185–1200.

[52] James W Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science* 8, 3 (1997), 162–166.

[53] James W Pennebaker. 2004. Theories, therapies, and taxpayers: On the complexities of the expressive writing paradigm. *Clinical Psychology: Science and Practice* 11, 2 (2004), 138–142.

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

[55] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[56] James W Pennebaker and Janel D Seagal. 1999. Forming a story: The health benefits of narrative. *Journal of clinical psychology* 55, 10 (1999), 1243–1254.

[57] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 186–195.

[58] Diane M Quinn and Stephenie R Chaudoir. 2009. Living with a concealable stigmatized identity: the impact of anticipated stigma, centrality, salience, and cultural stigma on psychological distress and health. *Journal of personality and social psychology* 97, 4 (2009), 634.

[59] Delbert G Robinson, Margaret G Woerner, Marjorie McMeniman, Alan Mendelowitz, and Robert M Bilder. 2004. Symptomatic and functional recovery from a first episode of schizophrenia or schizoaffective disorder. *American Journal of Psychiatry* 161, 3 (2004), 473–479.

[60] Robert R Rodriguez and Anita E Kelly. 2006. Health effects of disclosing secrets to imagined accepting versus nonaccepting confidants. *Journal of Social and Clinical Psychology* 25, 9 (2006), 1023–1047.

[61] Algimantas M Shimkunas. 1972. Demand for intimate self-disclosure and pathological verbalization in schizophrenia. *Journal of Abnormal Psychology* 80, 2 (1972), 197.

[62] Joshua M Smyth and James W Pennebaker. 1999. Sharing one's story: Translating emotional experiences into words as a coping tool. *Coping: The psychology of what works* (1999), 70–89.

[63] Elizabeth A Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25, 1 (2010), 1.

[64] Richard G Tedeschi and Lawrence G Calhoun. 2004. " Posttraumatic growth: Conceptual foundations and empirical evidence". *Psychological inquiry* 15, 1 (2004), 1–18.

[65] Douglas Turkington, David Kingdon, and Peter J Weiden. 2006. Cognitive behavior therapy for schizophrenia. *American Journal of Psychiatry* 163, 3 (2006), 365–373.

[66] Sarah E Ullman and Henrietta H Filipas. 2001. Predictors of PTSD symptom severity and social reactions in sexual assault victims. *Journal of traumatic stress* 14, 2 (2001), 369–389.

[67] Bessel A Van der Kolk. 1996. The complexity of adaptation to trauma: Self-regulation, stimulus discrimination, and characterological development. (1996).

[68] Roland Vauth, Birgit Kleim, Markus Wirtz, and Patrick W Corrigan. 2007. Self-efficacy and empowerment as outcomes of self-stigmatizing and coping in schizophrenia. *Psychiatry research* 150, 1 (2007), 71–80.

[69] Jessica Vitak and Jinyoung Kim. 2014. You can't block people offline: Examining how Facebook's affordances shape the disclosure process. In *CSCW*. ACM, 461–474.

[70] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *CSCW*. ACM, 74–85.

Received June 2017; revised August 2017; accepted November 2017