Measuring the Stigmatizing Effects of a Highly Publicized Event on Online Mental Health Discourse

Anna Fang
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Haiyi Zhu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

ABSTRACT

Media coverage has historically played an influential and often stigmatizing role in the public's understanding of mental illness through harmful language and inaccurate portrayals of those with mental health issues. However, it is unknown how and to what extent media events may affect stigma in online discourse regarding mental health. In this study, we examine a highly publicized event – the celebrity defamation trial between Johnny Depp and Amber Heard – to uncover how stigmatizing and destigmatizing language on Twitter changed during and after the course of the trial. Using causal impact and language analysis methods, we provided a first look at how external events can lead to significantly greater levels of stigmatization and lower levels of destigmatization on Twitter towards not only particular disorders targeted in the coverage of external events but also general mental health discourse.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

online communities,mental health,mental health stigma

ACM Reference Format:
Anna Fang and Haiyi Zhu. 2023. Measuring the Stigmatizing Effects of a Highly Publicized Event on Online Mental Health Discourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581284

1 INTRODUCTION

Stigma – negative attitudes that devalue a group of people in social contexts [62] – is a central challenge for those suffering from mental health issues. Mental health stigma can take the form of prejudice from the public as well as self-stigma, which is when people internalize these stigmatizing attitudes of the public [26, 37, 83]. Common stigmatizing stereotypes about people with mental illness include that they are dangerous, incompetent, and at fault for their conditions, often leading to discrimination in society such as limited social or career opportunities and isolation [26]. As a result, both public and self-stigma are primary reasons why many people choose not to seek mental health treatment, avoid diagnosis and labeling of their mental illness, and suffer from worsened personal self-esteem and social support [26].

While there are many potential causes of mental health stigma, a key influence throughout history in creation and perpetuation of mental health stigma has been the media. News coverage has profound contributions to societal attitudes and knowledge about health [53]; in recent years, online forms of media, such as online social networking platforms, have likewise come to be an important influence and reflection of society's perspectives towards mental well-being [3]. Both traditional and online news coverage of mental health has often stereotyped and showed misleading portrayals of people with mental illness as dangerous and a risk to the public [6, 14, 27], often through using stigmatizing language [37, 57, 83]; however, at its best news coverage also has shown potential for increasing mental health awareness [69].

Although past work has explored the stigmatizing language within traditional media towards mental illness [74], it remains unclear to what degree media impacts stigmatizing language use in online mental health discourse. It is vital to understand stigmatization of mental health in the online context, given that social media has become one of the primary sources of information and platform for discussion about a wide variety of mental health issues in recent years [66]. Online communities are a key resource for those with mental health issues to gain emotional and informational support as well as express their own experiences [31, 67]. The language present in online communities towards mental health has been shown to have drastic effects on people's well-being [33]. Not only can social media platforms be reflective of existing stigmatizing attitudes in society regarding mental illness, but they can also exacerbate the issue of mental health stigma by even further trivializing mental health problems and its treatment [18, 29, 52, 78].

In this paper we investigate how highly publicized events in media affect the use of stigmatizing language against mental health conditions on social media. We use a case study of an event of high public interest and viewership – the defamation trial between celebrities Johnny Depp and Amber Heard in 2022 that was publicly aired and live-streamed. The Depp-Heard trial discussed domestic abuse allegations from both parties, and included multiple diagnoses and discussion around personality disorders; in particular, the Depp-Heard trial drew causality between both parties' suspected diagnoses of mental illness and their alleged abusive behaviors. The Depp-Heard trial is particularly notable as a media event given its considerable social media attention and mass impact; the Depp-Heard trial was one of the most covered events in recent history on social media and news, reaching billions of viewership numbers across social media platforms like YouTube and TikTok as well as garnered greater public interest than over significant news topics at the time of its airing (discussed further in Section 3.1). We also

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581284

note the long history of case studies in HCI research for analyzing online community activity [55, 68, 94] and providing in-depth analyses of important experiences and events [38]; we later discuss the implications of this work for wider consideration about online interventions for mental health discourse. As a result, our study used the Depp-Heard trial to investigate how highly publicized events can affect stigmatizing (and destigmatizing) language on social media towards both personality disorders and general mental health.

In particular, our work answers the central research question:

RQ: To what extent do highly publicized events affect the stigmatization of mental health on social media?

In order to answer this question, we investigate three sub-research questions below using the case study of the Depp-Heard trial and its discussion around personality disorders. We answer the questions:

RQ1: What is the effect size and lasting impact (if any) of how the Depp-Heard trial affected stigmatization, destigmatization, and sentiment towards the mental illnesses discussed in its coverage?

RQ2: What kinds of stigmatizing and destigmatizing language were used regarding mental illnesses in the Depp-Heard trial during its coverage?

RQ3: How did the Depp-Heard trial affect general mental health discourse at large?

To answer the research questions, we used a text analysis tool (the Linguistic Inquiry and Word Count, or LIWC [89] to analyze language used in online discourse discussing general mental health as well as disorders particularly mentioned during the event. Existing literature shows that mental health stigma is often associated with danger and treating those with mental illness as an out-group [28, 82]. Based on [81], we measured stigmatizing languages using the LIWC dictionaries of *Risk* and *Power*, existing dictionaries on *Mechanistic Dehumanization* and *Animalistic Dehumanization* [76], as well as the dictionary *Inappropriate Labels for Mental Illness* . We measured the destigmatizing effect of language using the LIWC dictionaries of personal pronouns (*I* and *We*) to represent personal sharing and disclosure [31], as well as wellness-related words (*Health*, *Wellness*, and *Well-being*). We studied how online discourse surrounding the personality disorders highlighted in the trials (borderline, histrionic, and narcissistic) change on social media from the Depp-Heard trial using OLS regression and causal impact analysis [16], and compare these results to changes in discourse around the seven personality disorders that were not highlighted during the trial (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, and obsessive-compulsive). Additionally, we explored how there were further effects for discussion about general mental health from the event.

Our findings include that there was significantly greater use of stigmatizing language in Twitter posts regarding personality disorders as well as referencing mental health generally both during the time period of the trial and even after the trial's conclusion, as well as significantly less destigmatizing language including personal disclosure and well-being language. We found that tweets about personality disorders and general mental health during the Depp-Heard trial had significantly greater stigmatizing language use both during and after the trial; for example, analysis using OLS regression found that tweets mentioning disorders highlighted in the Depp-Heard trial showed 7% greater stigmatization, over 15% less destigmatizing language, and a striking 48% less personal pronoun usage during the trial. Additionally, our causal impact analysis found that the Depp-Heard trial had over 97% probability of causing a roughly 17% increase in stigmatizing language use. We also contribute knowledge of how stigmatizing language towards mental illness appears on social media as we found stigmatizing language was largely attributed to words about *dehumanization* and *power*, rather than labels that are societally known to be inappropriate labels towards people with mental illness (e.g. slurs against people with mental illness). As far as we know, our work is the first analysis to investigate how media coverage affects the stigmatization of online mental health discourse. Our work has direct implications for designing new tools for the measurement and surveillance of online stigmatization targeting mental health. Therefore, our work not only contributes to new understanding of mental health discourse in social media, but also has key implications for designing anti-stigma campaigns for greater public understanding and acceptance of mental health issues, creating interventions for more effective social support exchange, and opening up new opportunities for supporting constructive mental health discourse in social media .

2 RELATED WORK

We first review past literature about media impact on public perception of mental health. We then discuss related work regarding mental health discourse online, including stigmatization of disorders that were specifically highlighted in the Depp-Heard trial and will be analyzed in our study.

2.1 Media and Public Perception of Mental Health

Media and press contribute to the public's understanding of mental health, and can both stigmatize or destigmatize mental health issues [74]. At its best, public events and coverage around mental health can promote support for mental health issues and promote greater understanding towards mental health in both traditional media (e.g. television) and online social media [57, 85]. Parrott et al. found that public disclosure of personal mental health issues by male professional basketball players led to overwhelmingly supportive responses from fans online and helped challenge gender norms around mental illness [71]. Public attention from the American hip-hop music artist Logic's song titled after the U.S. National Suicide Prevention Lifeline was also found to coincide with reduction in suicides and a rise in calls to the suicide prevention hotline [69].

However, media coverage about mental illness has usually painted a harmful image of mental illness. Whitley and Berry conducted a 6-year analysis on over 11,000 newspaper stories covering mental illness, finding that dangerousness and criminality accounted for direct themes in 40% of news coverage, while treatment for mental illness was discussed less than 20% of the time and over 80% of articles lacked perspective from someone actually experiencing mental illness; there were no significant changes to this finding over time [92]. Similarly, work by Bowen investigated a 10-year period of United Kingdom tabloid coverage about personality disorders and found frequent use of violent language that were not

found when news discussed other physical health conditions (e.g. diabetes) [13]. These linguistic choices often construct unrealistic images of violence around people with mental illness, and can lead to higher levels of self-stigma among those with a diagnosis; this can be especially harmful for those suffering from disorders like personality disorders, which our study focuses on, that are already characterized by negative self-concept [13]. Past work studying newspaper articles similarly found reinforced stereotypes of people with mental illness as violent and unpredictable [93], and over two decades of reviewing mass media coverage found mental illness depicted negatively as peculiar and dangerous [57]. Media sources have also made online social media particularly vital to their spread of news about mental health diseases due to its increasing interest in online platforms [3].

2.2 Mental Health Discourse on Social Media

As the internet and social media have become pivotal for sharing knowledge, there is growing reliance on the internet as a source of information and health advice [3]. In fact, mainstream and traditional media, such as television and news channels, have begun to turn towards social media for their coverage in order to influence large groups of people at once [49]. Social media platforms has thus become a primary avenue for mental health discourse, allowing users to freely share and discuss information on a wide variety of mental health issues. Research done on online communities and well-being has shown that online communities provide users with both informational and emotional support [66], provide a safe space in anonymity for individuals to share their experiences and receive support [23, 31, 34], and lead to reduced suicidal ideation and improved well-being [33].

Although social media can be a key platform for people to find mental health support [42, 47, 91], it has also been found to be significantly more stigmatizing towards mental health issues compared to physical health issues [80]. Stigma can take many forms in the context of mental health including, but not limited to, blaming people for their illness, incorrectly labeling people with mental health conditions as dangerous, and promoting avoidance or withdrawal away from people who are ill [46]. Stigmatized individuals suffer from high levels of isolation and internalized self-stigma, which is when people accept societal prejudices and incorporate these ideas into their own self-concept; Self-stigma causes lower self-esteem and reduced self-efficacy, less social support, social maladaption, and is linked to greater self-harm and suicide attempts [46]. Certain illnesses, such as bipolar disorder and obsessive-compulsive disorder, have been found to face greater stigmatization and trivialization as well as less support on online social media compared to content about general mental health [18, 80]. As a result, many people with mental health issues fail to seek treatment due to feelings of hopelessness or fear of being seen as crazy [46]. Studies such as ours that investigate the stigmatizing language towards mental illness in online communities can lend to greater understanding of evolving public perspective on mental health and countering the immensely harmful effects of public stigma on those with mental health issues.

Online discourse about mental illness has often been shown to contain misguided beliefs and misinformation, as well as promotion of unhealthy behaviors and stigmatizing attitudes towards specific mental illnesses such as eating disorders and schizophrenia [20, 52, 78]. Even treatment for mental health disorders, such as antidepressants to treat depression and anxiety, are subject to negative attitudes including trivialization on social media [29]. There is some work within the HCI community that has specifically targeted dealing with stigmatizing mental health experiences online, including work that has created a social bot for social media users to practice reacting to mental illness disclosure [56] and investigating the dynamics of disclosing on social media experiences of pregnancy loss [5]. Past work has also studied the stigmatizing experiences of online social media disclosure such as regarding users' LGBTQ+ identity [35, 84], socioeconomic and class struggles [79], and sexual violence [45]. However, although there is acknowledgement in much of mental health research that stigmatization is a central challenge to help-seeking, little work has specifically focused on stigmatizing language use towards mental health specifically.

We particularly highlight prior work done on analyzing language and conversation patterns in the online well-being context, which has been one of the most effective and naturalistic ways to understand people's mental health conditions, perspectives, and differences in people's psychological state [21]. Our work is situated in and motivated by prior work that has shown how largescale changes in a population's psychological and emotional state are reflected in their language use after a significant event [24]; Kumar et al. also found that posting activity, emotional expression, and topic content regarding mental health were affected by suicides of high-profile figures, including increased posting activity, increased suicidal ideation in posts, and negative and anxious emotion [60]. Intervention for online mental health have been studied regarding awareness for eating disorders, showing the application of causal impact analysis to health discussion on social media and quantifying engagement with online awareness campaigns on Twitter [88]. Past work in the online mental health space specifically includes findings that language use in mental health discussion online has been shown to reflect a user's mental health conditions [25] and indicate diagnosis of depression or even suicidal ideation [30, 32, 34].

2.3 Stigma Against Personality Disorders

Our work will primarily investigate change in language around personality disorders highlighted during the Depp-Heard trial. Given that the Depp-Heard trial included lengthy discussion and multiple diagnoses of personality disorders, we review below past work about personality disorder stigma in particular. Although there has been research about general discourse around common mental health disorders such as depression and anxiety, very little work has studied online discourse about personality disorders despite them being some of the most stigmatized of all mental health disorders [12].

Personality disorders are defined by the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), the principal authority for psychiatric diagnoses, as "an enduring pattern of inner experience and behavior that deviates markedly from the expectations of the individual's culture, is pervasive and inflexible, has

an onset in adolescence or early adulthood, is stable over time, and leads to distress or impairment" [39]. There are ten specified personality disorders: paranoid, schizoid, schizotypal, antisocial, borderline, histrionic, narcissistic, avoidant, dependent, and obsessive–compulsive personality disorder[1]. Evidence suggests that personality disorders are among the most stigmatized of all mental health disorders [2, 12]. personality disorders are susceptible to greater stigma and less recognition when compared to other common mental health disorders, such as depression, anxiety, and eating disorders [19, 87]; for example, Furnham and Dadabhoy's 2012 study found that out of 102 non-expert participants, only 5% were able to identify a case vignette as borderline personality disorder (borderline personality disorder) compared to 74% accurately identifying depression [44]. Harmful language has perpetuated personality disorder stigma further. Past work has found that those with borderline personality disorder are often referred to as "manipulative", which inaccurately suggests that those with borderline personality disorder are consciously displaying their behaviors [2]. Negative language around personality disorders often ignore the origin of these disorders as well, which are often linked to childhood trauma and experiencing previous devaluation of emotions [2]. Stigma against those with personality disorders is even prevalent within the healthcare industry, with past studies finding that workers view patients as difficult, nuisances, frustrating, and "bad" rather than "ill" [12, 58].

3 METHODS

In order to study how events in media affect online discourse of mental health, we focused on a highly publicized and widely viewed event involving discussion around mental illness – the celebrity trial between Johnny Depp and Amber Heard. We examine how language used in online discourse around mental illness changed both during and after the trial, compared to before the trial's start. Below, we describe the trial and its discussion around mental illness (Section 3.1), the collection method for social media posts on Twitter about illnesses related to the Depp-Heard trial as well as general mental health discourse (Section 3.2), our language analysis methods using the psycho-linguistic tool Linguistic Inquiry and Word Count (Section 3.3), and our application of causal impact analysis to this language change over time (Section 3.4).

3.1 The Depp-Heard Trial

John C. Depp, II v. Amber Laura Heard was a defamation trial held in Virginia starting on April 11, 2022 and concluding with jury verdict on June 1, 2022. Both John C. Depp (commonly known as Johnny Depp) and Amber Heard are celebrity actors; Depp and Heard were married in early 2015 and divorced by early 2017. The trial involved Depp as plaintiff filing a defamation complaint against defendant Heard while Heard filed counterclaims against Depp, tracing back to domestic abuse allegations from the dissolution of their marriage.

The Depp-Heard trial was publicly televised and live-streamed (a rare occurrence in Virginia courts) and widely covered by both traditional and new media. News coverage and social media posts regarding the Depp-Heard trial had extremely high viewership. For example, one of the largest channels airing the trial called Law&Crime Network has over a billion views on its social media content related to Depp-Heard trial coverage to date [2] and reported having one million viewers per hour of the trial on their YouTube channel alone [3]. By April 29, 2022, videos on social media platform TikTok having the hashtag #justiceforjohnnydepp had over 5 billion views (and 18 billion views by the trial's conclusion) while TikTok videos with the hashtag #JusticeForAmberHeard had over 21 million views. Many media outlets even found coverage about the trial was of greater public interest than other significant news topics of the time such as international affairs or abortion rights in the U.S. [3].

Mental illness was an important topic of discussion during the trial as both Amber Heard and Johnny Depp were diagnosed (or suspected of having a diagnosis) with personality disorders by experts on the stand during the live-streamed trial. Depp's team's forensic psychologist publicly diagnosed Amber Heard with both borderline personality disorder and histrionic personality disorder, while one of Heard's team's trial experts stated that Depp exhibited traits of narcissistic personality disorder. These diagnoses and discussions of personality disorders were said by experts to have been "weaponized" by both parties during the trial [4], which drew causality between the opposing party's personality disorder and accused abusive behavior [5].

Thus, as a result of the highly publicized nature of the Depp-Heard trial and its direct discussions of mental illness, we use the Depp-Heard trial as our case study of how media coverage can affect language use about mental health on Twitter.

3.2 Twitter Dataset

Twitter is a popular social media platform that allows users to post messages (called "tweets") that are up to 280 characters in length. Twitter users can mark their tweets to be private, but otherwise tweets are public and are available through Twitter's public API [6]. Twitter also allows users to interact with each other's tweets through (1) like, (2) "retweet" (re-share the tweet on a user's own timeline), and/or (3) reply. Below, we describe our criteria for collecting tweets using the Twitter API.

First, we established the time window for a tweet's creation in order to be included in our study. Since the trial's airing lasted 51 days from opening arguments on April 11, 2022 until the jury decision on June 1, 2022, we collected tweets during the 51 days before the trial, 51 days during the trial, and 51 days after the trial. Second, we established keywords to include when collecting tweets of interest between 51 days prior to the trial to 51 days past the trial's conclusion. Given that the Depp-Heard trial specifically highlighted three personality disorders – borderline, histrionic, and narcissistic (as described in Section 3.1) – we collected all tweets that mentioned any one of *"borderline personality disorder"* or *"BPD"* (a common acronym for the disorder), *"histrionic personality disorder"*, or *"narcissistic personality disorder"*. We also collected tweets

[1]Note that obsessive-compulsive disorder, commonly known as OCD, is not the same as obsessive-compulsive personality disorder (OCPD)

[2]https://www.youtube.com/c/LawCrimeNetwork
[3]https://www.axios.com/2022/05/17/amber-heard-johnny-depp-trial-social-media
[4]https://www.psycom.net/mental-health-wellbeing/johnny-depp-amber-heard-trial
[5]https://www.standard.co.uk/comment/amber-heard-johnny-depp-trial-borderline-personality-mental-health-b998112.html
[6]https://developer.twitter.com/en/docs/twitter-api

referencing any one of the other seven DSM-5-recognized personality disorders (*"paranoid personality disorder", "schizoid personality disorder", "schizotypal personality disorder", "antisocial personality disorder", "avoidant personality disorder", "dependent personality disorder", "obsessive-compulsive personality disorder"*) as a comparison point. In other words, to answer RQ1 and RQ2 we compare language change seen for tweets referencing specific disorders highlighted during the trial versus similar disorders that were not mentioned in trial. In addition to tweets mentioning personality disorders, we collected tweets referencing the keyword *"mental health"* in order to answer RQ3 about any significant language effects from the trial on mental health discourse generally. For all three of the above sub-datasets, we also did a manual pass of 100 tweets, qualitatively finding they were all relevant to the mentioned mental disorders or mental health issues. We note that many tweets also concern other topics (i.e. domestic violence, drugs) but remain relevant to personality disorders and/or mental health, such as tweets promoting avoidance of romantic relationships with people with personality disorders for fear of abuse. Descriptive statistics of our Twitter dataset, including total numbers of tweets and average like/retweet/reply counts, are in Table 1. For all tweets, we organize them into three time periods depending on the tweet's creation time: before, during, or after the Depp-Heard trial time period. We then measure how stigmatizing language use changes during each of these time periods for each of the three sub-datasets described.

3.3 Language Analysis

After collecting all tweets of interest, we conducted language analysis to compare and contrast language use over the course of the trial. Past work has found that stigmatization and prejudice are majorly rooted in language. For example, speaking about mental health in a stigmatizing way (i.e. mocking others by saying they are "mentally challenged") and using pejorative terms can intensify negative and prejudicial attitudes towards specific groups of people [36, 72]. Emotional expression is also a strong driving force for public discourse, and social media texts alone can be powerful for reflecting and driving others' emotional and mental stances towards topics [9, 72]. As a result, we created linguistic categories to score and evaluate tweets on their stigmatizing and destigmatizing language use as well as their emotional tone.

We used automated text analysis tool Linguistic Inquiry and Word Count (LIWC) [89]. LIWC is a dictionary-based text analysis software that counts words using psychologically meaningful dictionaries. LIWC scores a given text (in our case, a tweet's contents) by calculating a percentage of total words in the text matching the list of words in a given dictionary. LIWC has been widely used and validated in prior research studies about online language use [10, 22, 89] and has its own in-house dictionaries measuring categories such as sentiment, emotion, and context; LIWC has also been used in past work to measure attitudes about health [1, 72, 90, 95]. In addition to LIWC's existing in-house dictionaries, users can create custom dictionaries. We rely on both LIWC's in-house dictionaries and custom dictionaries created by prior research studies to create aggregated scores in linguistic categories of (1) Stigmatizing, (2) Destigmatizing, and (3) Emotional Tone to score tweets. We then compare scores of stigmatizing, destigmatizing, and emotional tone language usage changed over time when comparing before, during, and after the Depp-Heard trial. Additionally, we used causal impact analysis [16] to directly estimate the effect of the trial on the language use, using data from exactly a year prior (February to July of 2021) to compare how stigmatizing and destigmatizing language use is predicted to have evolved had the trial not happened (described further in Section 3.4).

We summarize below each LIWC dictionary used in our study broken down by higher level categorizations, and provide example keywords for each dictionary in Table 2.

3.3.1 Measures of stigmatizing language use. To create an overall measure of stigmatizing language use, we used work by Pavlova and Berkers that studied mental health discourse online [72] and drew from their use of LIWC in-house dictionaries as well as a custom LIWC dictionary created in their study. We used in-house dictionaries of **Risk** and **Power** that have been used in past literature to study mental health stigma based on stigmatizing language around mental illness often revolving around risk, danger, and violence [72]. We also included existing dictionaries from past work about linguistic **Mechanistic Dehumanization** and **Animalistic Dehumanization**, as dehumanizing language has been used in media to express distrust and establish an "out-group" of people [76]. Lastly, we use the existing user-made LIWC dictionary from Pavlova and Berkers that we call **Inappropriate Labels for Mental Illness**, which includes words or phrases that have been deemed inappropriate specifically for describing mental illness as identified by [81]. We excluded words in all LIWC dictionaries that may be understood in a non-stigmatizing way or that are part of our study's clinical terms for personality disorders (i.e. "ill", "histrionic", "schizophrenia") [72, 81]. Thus, for each tweet we average the output LIWC scores of Risk, Power, Mechanistic Dehumanization, Animalistic Dehumanization, and Inappropriate Labels for Mental Illness to create an overall **Stigmatizing** score. Then, we conduct OLS regression to explore how trial time and tweet content affects the overall Stigmatizing score of tweets; we also include analysis results for these dictionaries broken down individually for analysis in Section 4.2.

3.3.2 Measures of destigmatizing language use. Self-disclosure is one primary way to destigmatize mental illness; hearing about others' experiences having mental illness has been shown to effectively address both public stigma and self-stigma, and further understanding of mental illness symptoms [7, 43]. Additionally, discussion around the causes, medical symptoms, and treatment for mental illness are important frames for destigmatizing mental illness [92]; speaking about mental health as essential to physical and medical well-being also has destigmatizing effects [7, 73]. Health-related words online, indicating sharing of health and well-being information, has also been shown to receive greater social support [31].

As a result, we used the LIWC dictionaries of **I** and **We** to indicate personal sharing and self-disclosure, as has been studied in past literature [31]. We also used the LIWC in-house dictionaries of **Health**, **Wellness**, and **Well-being** as health and wellness-related words that consist of informational-related words and hopeful well-being words. Lastly, in addition to self-disclosure and health-related

Table 1: Descriptive statistics of our Twitter dataset, separated into three sub-datasets: tweets containing (1) personality disorders mentioned in trial (borderline, histrionic, narcissistic), (2) all other personality disorders, which were not highlighted in trial (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive), and (3) "mental health". We show the mean number of likes, retweets (RTs), and replies (repl.) for each sub-dataset.

	Before Trial Feb. 18, 2022 - April 10, 2022				**During Trial** April 11, 2022 - June 1, 2022				**After Trial** June 2, 2022 - July 22, 2022			
Tweets with:	N	Avg Likes	Avg RTs	Avg Repl.	N	Avg Likes	Avg RTs	Avg Repl.	N	Avg Likes	Avg RTs	Avg Repl.
personality disorders in trial	54.1k	9.6	1.2	0.7	94k	12.4	1.6	0.8	68.6k	6.9	1	0.7
all other personality disorders	4.3k	4.4	0.6	0.6	5.2k	9.5	1.8	0.7	5.2k	5.5	0.9	0.7
"mental health"	1.29m	11.8	2.1	0.8	1.68m	11.9	2.3	0.9	1.45m	12.5	2.6	0.8

words, we included dictionaries to indicate positive prosocial behavior, which has been shown to mitigate negative emotions in everyday life and create greater feelings of belonging and trust [4, 77]. We included the LIWC in-house dictionary **Prosocial** and a custom LIWC dictionary studied in past work called the **Self-transcendent** dictionary that has words related to "greater human connectedness, prosociality, and human flourishing" [50]. Thus, we average the LIWC scores of Personal Pronouns, Health, Wellness, Well-being, Self-transcendent, and Prosocial to create an overall **Destigmatizing** score. Then, we conduct OLS regression to explore how trial time and tweet content affects the overall Destigmatizing score of tweets; we also include analysis results for these dictionaries broken down individually for analysis in Section 4.2.

3.3.3 Measures of sentiment. In addition to measuring stigma and destigma around mental illness, we also investigated overall sentiment in tweets through a linguistic category of Emotional Tone. Social and cultural attitudes can both influence and be influenced by emotional expression in the online context [11]. Sentiment analysis in social media posts has often been used in research to evaluate users' opinions about health topics in particular [51, 96]. Additionally, positive tone and sympathetic portrayals of people with mental illness can also be important for support of mental health [72, 92]. As a result, to evaluate the emotionality of tweets we used the existing in-house LIWC dictionaries of **Negative Tone** and **Positive Tone**.

3.4 Causal Impact Using Time-Series Models

In addition to conducting OLS regression, we also conducted **causal impact analysis using Bayesian structural time-series models** to estimate the probability that changes in tweets' stigmatizing/destigmatizing scores was actually caused by the trial, by comparing to a control dataset of tweets with the same content but exactly one year prior (February to July 2021). Descriptive statistics of our collected 2021 tweets are shown in Table 3.

We conducted causal inference analysis using Bayesian structural time-series models based on prior work by Broderson et al. [16]; prior work has used causal impact analysis to determine effectiveness of interventions in the online health context [88]. Causal impact analysis takes time series data and attempts to estimate the effects of some event or intervention, identified by a particular point in time splitting the data into a "pre-intervention period" versus "post-intervention period). Specifically, causal impact analysis uses a control time series unaffected by an intervention (e.g. tweets from 2021) to construct a Bayesian structural time-series model for a response time series (e.g. tweets in our study's dataset in 2022); this time-series model is then used to predict the counterfactual if the intervention (e.g. the Depp-Heard trial) had not occurred, and compares this to the actual outcome in the post-intervention period. This model relies on a few assumptions, including that the control time series was not directly affected by the intervention; in our case, the data from 2021 was indeed not affected by the Depp-Heard trial that took place in 2022. Additionally, the model assumes that the relationship between covariates and response time series as established during some defined "pre-period" would remain stable throughout the "post-period". We apply the causal impact model for our study through the R package [16] and show our results, including the estimated average causal effect of the Depp-Heard trial on stigmatizing/destigmatizing language as well as the posterior probability of the trial having a causal effect, in Section 4.1.3.

4 RESULTS

In this section, we present the results of our analysis (both OLS regression and causal impact analysis) on each of the datasets described in Section 3.2 to show how trial time (before, during, or

Table 2: Each individual LIWC dictionary used in our study along with example keywords in each dictionary. Note that these keywords are not exhaustive, and the majority of these dictionaries contain hundreds of words. Stigmatizing and destigmatizing categories are compared in Section 4.1.2, while tone and emotion results are presented in Section 4.1.3.

LIWC Dictionary	Categorization	Keywords (examples)
Inappropriate Labels for Mental Illness	Stigmatizing	attention-seeking, deranged, halfwit lunatic, nutcase, sick in the head, psycho
Mechanistic Dehumanization	Stigmatizing	alien, hardwired, calculating, heartless, incompatible, machine, object, unemotional
Animalistic Dehumanization	Stigmatizing	creature, dangerous, infectious, pig, subhuman, uncivil, vermin
Power	Stigmatizing	advantage, armed, assault, beaten, killing, mighty, nuclear, obey, weak
Risk	Stigmatizing	avoid, careful, threat, unprotected, unsafe
Self-transcendent	Destigmatizing	aspire, believe, benevolent, compassion, loyal, overcome, virtuous, wholehearted
Health	Destigmatizing	ache, faint, healed, hospital, nutrients, pain, psychiatry, tired, wounded
Wellness	Destigmatizing	diet, gym, healthful, herbal tea, nutrition, physical activity, yoga
Well-being	Destigmatizing	ambition, dream, happiness, growth, prosocial, understanding, value
Prosocial	Destigmatizing	care, empower, friendly, helpful, inspiring, kindness, solidarity, virtuous
Personal Pronouns (I and We)	Destigmatizing	i, i'm, my, my kid, let's, our, us, we
Tone - Positive	Emotional Tone (Positive)	accuser, fool, frantic, furious, harmful, pitiful, problematic, wrong, your fault
Tone - Negative	Emotional Tone (Negative)	brave, calm, dear, favorite, happy, love, thank, triumph, valuable, well

after trial) impacted the language of tweets. We first report in Section 4.1 our regression results for how stigmatizing/destigmatizing language and tone (aggregated from LIWC dictionaries as described in Section 3.3) changed over the course of the trial timeline, and the likelihood this change was caused by the trial. We used the statistical software package Stata [7] to run regressions on each dataset (tweets including trial-mentioned personality disorders, tweets including all other personality disorders, tweets including "mental health" keywords) individually as independent variables while utilizing chat rating as our dependent variable; all qualities were organized into categorical variables (i.e. gender groups, age groups, experience level groups). Then, we investigate further what types of words (i.e. individual LIWC dictionaries) are used to stigmatize or destigmatize mental health from the trial in Section 4.2. Finally, we discuss in Section 4.3 the significant effects on tweets mentioning "mental health" to explore how events targeted to specific illnesses also impact general mental discourse.

[7] www.stata.com

Table 3: Similar to Table 1, shown are descriptive statistics of our study's Twitter dataset but one year prior in 2021. Similarly, the statistics are separated into three sub-datasets: tweets containing (1) personality disorders mentioned in trial (borderline, histrionic, narcissistic), (2) all other personality disorders, which were not highlighted in trial (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive), and (3) "mental health".

	Feb. 18, 2021 - April 10, 2021				April 11, 2021 - June 1, 2021				June 2, 2021 - July 22, 2021			
	N	Avg Likes	Avg RTs	Avg Repl.	N	Avg Likes	Avg RTs	Avg Repl.	N	Avg Likes	Avg RTs	Avg Repl.
Tweets with:												
personality disorders in trial	6.4k	7	1.1	0.7	7.5k	6.8	1.1	0.7	7.6k	6.4	0.9	0.6
all other personality disorders	1.5k	4.4	0.8	0.7	1.5k	6.4	0.9	0.7	1.4k	3.8	0.7	0.6

4.1 Stigma in Tweets Mentioning Personality Disorders

For RQ1, we investigated the effect size and lasting impact of how the Depp-Heard event affected stigmatizing language towards mental health.

4.1.1 Volume of Tweets. Before investigating our study's metrics for stigmatizing, destigmatizing, and tone language change, we briefly review any changes in tweet volume as well as their counts of likes, retweets, and replies.

As charted in Figure 1, there was a noticeable rise in the number of tweets for all three sub-datasets (tweets with personality disorders in trial, tweets with all other personality disorders, and tweets with "mental health") during the trial. Tweets referencing trial-highlighted disorders (borderline, histrionic, narcissistic) saw a 74% rise in the total number of tweets during the trial compared to before the trial's start date. Tweets referencing any of the other personality disorders, also saw a rise in the total number of tweets, although to a lesser degree (increase of 21% during trial compared to before trial). For all of these sub-datasets, the volume of tweets lowered after the trial, but still remained at a similar or higher volume than before the trial's start.

As for the number of interactions per tweet, we ran an OLS regression for each dataset (tweets with personality disorders in trial vs. tweets with all other personality disorders), predicting the number of likes, retweets (RTs) and replies per tweet using time period relative to the trial dataset. Full results are available in our supplementary material. We did not see any significant effects for the volume of likes or retweets per tweet as shown in our full results. However, we did see a statistically significant rise during the trial in the number of replies per tweet with a 10.7% rise and 11.7% rise during the trial's duration for disorders mentioned in trial and all other personality disorders, respectively.

4.1.2 OLS Regression on Stigmatizing and Destigmatizing Language. To answer RQ1, we measured how tweets before, during, and after the trial vary in their LIWC scores of stigmatizing and destigmatizing language. We examined tweets that reference personality disorders specifically highlighted in trial, and contrast their language changes to tweets with personality disorders not mentioned in trial as a comparison point. Our results are shown in Table 4.

As seen in Table 4, we saw overall significantly greater use of stigmatizing language and significant less destigmatizing language from the trial for tweets that included personality disorders in trial. **There was a nearly 7% increase in stigmatizing language use during the trial alongside a 15.6% decrease in destigmatizing language use.** We saw no significant changes for tweets that do not mention these personality disorders. As a result, our findings seem to support the idea that the Depp-Heard trial had an effect of greater stigmatization towards illnesses that were diagnosed and highlighted during trial, which we did not see for very similar disorders that were not mentioned in the trial.

Additionally, we note a potentially lasting impact of the Depp-Heard trial. **There continued to be significant effects even during the nearly two months after the trial's conclusion as seen by significant values of a 2.3% rise in stigmatizing language use and a 5.3% decrease in destigmatizing language use after the trial's conclusion for tweets mentioning borderline, histrionic, or narcissistic personality disorder.** This effect size was to a smaller degree (for example, stigmatizing language rising 2.3% after the trial versus 6.9% increase during the trial, compared to our baseline of before the trial's start); this may support the intuition that the Depp-Heard trial's discussion of mental illness triggered a surge of mental health stigma that "leveled out" but continued to be significant even beyond just the trial's duration.

4.1.3 Causal Impact Analysis Results. Our causal impact analysis results are shown in Table 5. As described in Section 3.4, causal impact analysis takes input of a control time series unaffected by an intervention, a pre-intervention period, a post-intervention period, and a response time-series. We used 2021 data of February 18, 2021 until July 22, 2021 as the control time series (which was unaffected

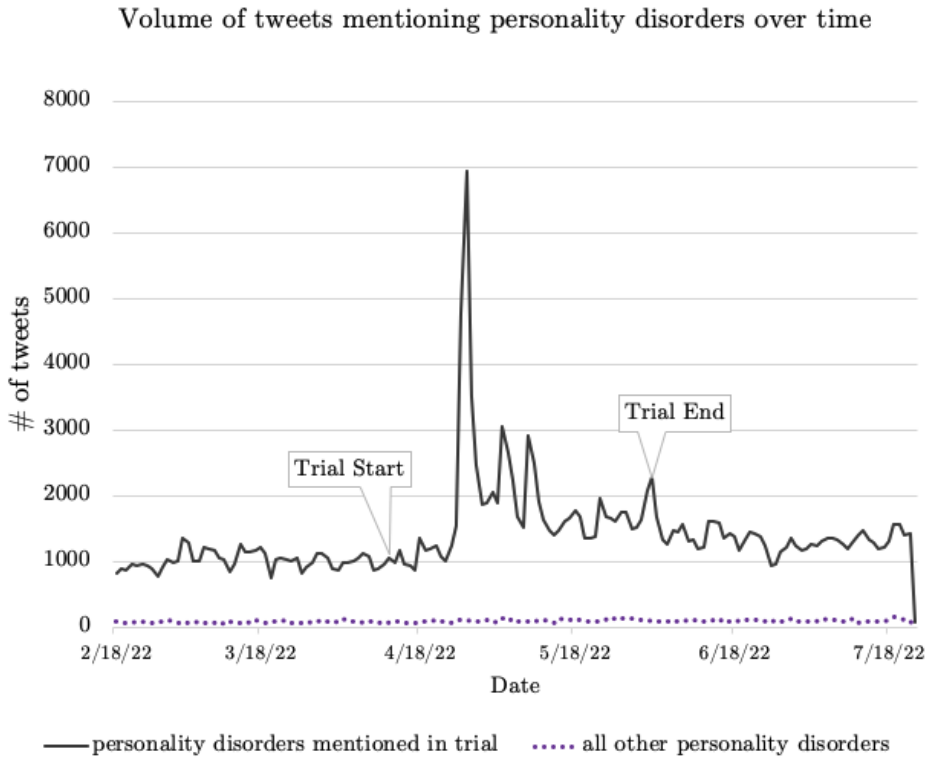

Figure 1: Volume of tweets mentioning personality disorders from the 51 days prior to 51 days after the Depp-Heard trial

Table 4: OLS regression results for Stigmatizing, Destigmatizing, Negative Tone, and Positive Tone language use in tweets containing either (1) personality disorders mentioned in the Depp-Heard trial (borderline, histrionic, narcissistic) (2) all other personality disorders (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive). The dependent variable for all models is the LIWC dictionary score for their respective categories; the predictor for all models is the time period relative to the trial. We use the Before Trial as our baseline for each regression.

		Model 1 Stigmatizing *coeff, std.err*	**Model 2 Destigmatizing** *coeff, std.err*	**Model 3 Negative Tone** *coeff, std.err*	**Model 4 Positive Tone** *coeff, std.err*
Tweets containing:	**Time period relative to trial**				
personality disorders in trial	before *(baseline)*	0	0	0	0
	during	**0.069***, 0.003	**-0.156***, 0.008	**0.121***, 0.027	**-0.326***, 0.023
	after	**0.023***, 0.004	**-0.054***, 0.009	-0.012, 0.029	**-0.247***, 0.025
all other personality disorders	before *(baseline)*	0	0	0	0
	during	0.015, 0.016	0.045, 0.033	-0.21, 0.096	-0.10, 0.075
	after	0.009, 0.016	0.023, 0.033	-0.02, 0.097	**-0.15***, 0.076

* **p < 0.05**

by the Depp-Heard trial in 2022), a pre-period intervention of previous to the trial's start date (April 11, 2022), a post-period of after the trial's start date, and the response time series of our study's tweet dataset from February 18, 2022 until July 22, 2022.

The results from the causal impact analysis are overall consistent with the OLS regression results. Our results from Model 1 in Table 5 studying the effects on Stigmatizing language use in tweets mentioning personality disorders in trial found a striking

Table 5: Results for Causal Impact Time-Series Analysis, finding the probability of a causal effect of the Depp-Heard trial on post-period changes in Stigmatizing (Model 1) and Destigmatizing (Model 2) scores.

		Model 1 Stigmatizing		**Model 2 Destigmatizing**	
		average	cumulative	average	cumulative
personality disorders in trial	Actual:	0.3	289.6	2.4	2314.5
	Prediction (s.d.):	0.26 (0.02)	249.2 (20.4)	2.5 (0.06)	2375.9 (60.02)
	95% CI:	[0.2, 0.3]	[209.8, 290.4]	[2.3, 2.6]	[2259.7, 2494.3]
	Actual effect (s.d.):	0.04 (0.02)	40.44 (20.44)	-0.06 (0.06)	-61.36 (60.02)
	95% CI:	[-0.001, 0.1]	[-0.8, 79.9]	[-0.2, 0.1]	[-179.7, 54.8]
	Relative effect (s.d.):	17% (9.8%)	17% (9.8%)	-2.5% (2.5%)	-2.5% (2.5%)
	95% CI:	[0.28%, 38%]	[-0.28%, 38%]	[-7.2%, 2.4%]	[-7.2%, 2.4%]
	Posterior tail-area probability *p*:	**0.027**		**0.147**	
	Posterior prob. of causal effect:	**97.3%**		**85%**	
all other personality disorders	Actual:	0.46	363.27	1.9	1465.4
	Prediction (s.d.):	0.46 (0.1)	361.07 (79.1)	1.8 (0.22)	1395 (171.53)
	95% CI:	[0.3, 0.7]	[213.9, 519]	[1.4, 2.2]	[1075.3, 1738.2]
	Actual effect (s.d.):	0.003 (0.1)	2.20 (79.1)	0.09 (0.22)	70.70 (171.53)
	95% CI:	[-0.2, 0.2]	[-155.7, 149.4]	[-0.3, 0.5]	[-272.9, 390.1]
	Relative effect (s.d.):	6.1% (28%)	6.1% (28%)	6.7% (14%)	6.7% (14%)
	95% CI:	[-30%, 70%]	[-30%, 70%]	[-16%, 36%]	[-16%, 36%]
	Posterior tail-area probability *p*:	**0.47**		**0.33**	
	Posterior prob. of causal effect:	**53%**		**67%**	

97.3% probability that Stigmatizing language increased by around 17% (seen by the relative effect output). Destigmatizing language in the same dataset saw an 85% probability of a causal effect of a smaller -2.5% change. On the other hand, similar to our regression results, we see little to no evidence of a causal effect for personality disorders that were not directly mentioned in the trial; Stigmatizing scores only had a 53% probability of a causal effect while Destigmatizing had a slightly higher 67% probability of causal effect.

4.1.4 Emotional Tone. Similar to our previous analysis, we used LIWC scores for dictionaries Negative Tone and Positive Tone as dependent variables predicted by trial time in two separate OLS regressions, one for tweets containing disorders highlighted in trial and one for all other personality disorders. Again, we ran regressions separately for tweets mentioning borderline, histrionic, or narcissistic personality disorders and for tweets mentioning any other personality disorder (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive) to compare differences in trial effects. Regression results are shown in Table 4.

Tweets with personality disorders in the Depp-Heard trial saw a significantly more negative tone during and after the trial. In particular, LIWC scores for Negative Tone saw a significant 12.1% rise in use during the trial alongside a 32.6% drop in Positive Tone during the same period. Importantly, results showed a significantly lower amount of Positive Tone even during the nearly two months after the trial compared to before the trial; Positive Tone scores were 24.7% lower after the trial compared to before the trial. Interestingly, we also saw a significant 15% drop of Positive Tone usage after the trial even for tweets containing any other personality disorder not mentioned in the Depp-Heard trial (but no significance during the trial's airing).

4.1.5 Replies to Tweets. In addition to analyzing the language use in tweets containing personality disorder keywords, we collected all replies to these tweets. Similarly to the above analyses, we measured Stigmatizing and Destigmatizing LIWC scores, as well as both

Positive Tone and Negative Tone among all replies. We additionally analyzed how the continuous variables of stigmatizing and destigmatizing scores for a tweet containing personality disorder terms was predictive of the stigmatizing and destigmatizing scores for its replies.

Our findings, shown in Table 6, show that there is a statistically significant rise in the levels of stigmatizing language and lower levels of destigmatizing language during the trial solely for replies to tweets that contain personality disorders highlighted in trial. In other words, **not only are posts referencing borderline, histrionic, and narcissistic personality disorder more stigmatizing and less stigmatizing during the trial, but their replies were likewise more stigmatizing and less destigmatizing.** We do not see significance after the trial's conclusion or for tweets that include all other personality disorders.

However, unlike the original tweets containing personality disorder keywords, the replies to tweets containing borderline, histrionic, or narcissistic personality disorder keywords had a significantly greater negative tone during the trial (seen by a coefficient of 0.509 for Negative Tone and -1.117 for Positive Tone in Table 6) but we see positive tone go to a much higher level with a coefficient of 1.259 after the trial's conclusion. This is contrasted with a still significantly higher level of negative tone after the trial's conclusion, however. More investigation would be needed to investigate the nuances of this contrast; it is possible that greater support occurred after the trial's conclusion, perhaps as a response to the negative tone displayed during the trial and taking the form of "anti-stigma" [75].

Our results in Table 7 show our results for using the Stigmatizing and Destigmatizing scores of tweets containing personality disorder terms as a predictor for the Stigmatizing and Destigmatizing scores of the tweet's own replies. We saw that there was a significant positive relationship between the stigma and destigma scores of a reply with the tweet it replied to. In other words, a tweet that is stigmatizing towards personality disorders had replies that were likewise stigmatizing, rather than these tweets being more likely to receive anti-stigma responses.

4.2 Language Choice within Stigmatizing and Destigmatizing Tweets

To answer RQ2 regarding what *kinds* of stigmatizing and destigmatizing language are used during and after the Depp-Heard trial, we identified the effect size and significance of each LIWC dictionary individually; results are shown in Table 8 for Stigmatizing dictionaries and Table 9 for Destigmatizing dictionaries.

For stigmatizing language use in tweets about personality disorders, we found that the only significant effects for rise in language use was for Power and Dehumanization. Interestingly, language regarding inappropriate labels for mental illness (Model 1 in Table 8) did not see any significance and in fact this language was rarely ever used throughout the tweets regarding any personality disorder. Instead, **language that was stigmatizing was much more heavily reliant on words related to Power as well as Mechanistic and Animalistic Dehumanization.** This may follow past literature finding that dehumanizing words are especially prevalent in judgments towards groups of people (particularly immigrants and those seen as different) and have been researched as causing social harm [63, 64]. Additionally, the Power dictionary in LIWC contains many words of violent and dangerous natures, which are often cited as a primary depiction of those with mental illness [72]. While Inappropriate Labels for Mental Illness is particularly mental health context-specific, it may contain words that are not often used by users to describe those with mental illness either due to societal knowledge that these words are not appropriate or lack of frequency of these words in everyday or online language. As for destigmatizing language, there were significant effects across the board for LIWC dictionaries and a clear negative relationship between during and after trial time with these dictionaries. In other words, during and after the trial, all of Self-transcendent, Wellness, Well-being, Prosocial, and Personal Pronoun use decreased significantly (Health had a significant decrease in use after the trial compared to before the trial, but no significance during the trial's duration). We also note an increase in the usage of personal pronouns for tweets referencing personality disorders not mentioned in trial, but no significant changes for any other dictionary.

4.3 Effects on General Mental Health Discourse

In addition to evaluating how the Depp-Heard trial may have influenced stigmatizing versus destigmatizing language around personality disorder discourse, we also investigated whether there may have been influence over general mental health discourse on Twitter as well.

We show our regression results using the similar techniques from previously in Table 10.

We find that there is a surprisingly significant effect on stigmatizing language and a decrease in destigmatizing language over the trial's time period and afterwards, similar to our findings for tweets containing trial-mentioned personality disorders. Although we cannot confirm that the trial caused these language changes (for example, it is possible that mental health discourse was skewing towards stigmatization regardless of the trial taking place), we do note the coefficients following a similar pattern to trial-mentioned disorders where there continues to be a significant effect even after the trial's conclusion but noticeably to a smaller coefficient that during the trial.

4.4 Summary of Findings

In sum, our key findings were:

- Tweets mentioning disorders highlighted in the Depp-Heard trial had significant increases in stigmatizing language and decreases in destigmatizing language. In particular, regression results showed a 7% increase in the stigmatizing language and a 16% decrease in destigmatizing language, while causal impact analysis showed over 97% probability that the trial caused a 17% rise in stigmatizing language. In contrast, all other disorders not mentioned in the trial had no changes in stigmatizing or destigmatizing language.
- We found a lasting impact of the Depp-Heard trial: OLS regression results showed was a 2.3% rise in stigmatizing language use and a 5.3% decrease in destigmatizing language use during almost two months after the trial's conclusion.

Table 6: OLS regression results for Stigmatizing/Destigmatizing LIWC score for *replies* to tweets (1) personality disorders mentioned in the Depp-Heard trial (borderline, histrionic, narcissistic) (2) all other personality disorders (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive). The dependent variable for Model 1 is the average LIWC score for Stigmatizing, Model 2 is the average LIWC score for Destigmatizing, Model 3 is LIWC score for dictionary Negative Tone, and Model 4 is LIWC score for dictionary Positive Tone; the predictor for all models is the time period relative to the trial. We use Before Trial as our baseline for each regression.

		Model 1 Stigmatizing *coeff, std.err*	**Model 2 Destigmatizing** *coeff, std.err*	**Model 3 Negative Tone** *coeff, std.err*	**Model 4 Positive Tone** *coeff, std.err*
Replies to tweets with:	**Time period relative to trial**				
personality disorders in trial	before (*baseline*)	0	0	0	0
	during	**0.083***,0.006	**-0.179***, 0.011	**0.509***, 0.043	**-1.117***, 0.066
	after	0.008, 0.007	0.015, 0.012	**0.103***, 0.049	**1.259***, 0.074
all other personality disorders	before (*baseline*)	0	0	0	0
	during	0.021, 0.030	0.064, 0.043	0.077, 0.186	0.382, 0.240
	after	0.027, 0.030	0.017, 0.043	0.093, 0.187	0.176, 0.240

* $p < 0.05$

Table 7: OLS regression results for how stigma of a reply is predicted by the stigma of a tweet with (1) personality disorders mentioned in the Depp-Heard trial (borderline, histrionic, narcissistic) (2) all other personality disorders (paranoid, schizoid, schizotypal, antisocial, avoidant, dependent, obsessive-compulsive). The dependent variable for Model 1 is the LIWC score for Stigmatizing for a reply to a tweet while the independent variable is the LIWC score of Stigmatizing for the tweet that received replies. The dependent variable for Model 2 is the LIWC score for Destigmatizing for a reply to a tweet while the independent variable is the LIWC score of Destigmatizing for the tweet that received replies.

	Model 1 Stigma of Reply *coeff, std.err*	**Model 2 Destigma of Reply** *coeff, std.err*
Tweets with:		
personality disorders in trial	**0.152***, 0.004	**0.137***, 0.003
all other personality disorders	**0.158***, 0.019	**0.166***, 0.126

* $p < 0.05$

- In terms of specific languages used to stigmatize personality disorders, we found increases in the use of Animalistic Dehumanization words (e.g., creature, dangerous, infectious, pig) and Power-related words (e.g., armed, assault, beaten, killing). In terms of specific destigamtizing languages, we found a significant decrease (48%) in using pronounces of I or We, which suggests a significant decline in self-disclosure.
- In terms of general mental health discourse on Twitter, we saw a surprisingly significant effect on the increase of stigmatizing language and a decrease in destigmatizing language over the trial's time period (+11% and -3%) and afterward (+6% and -2%).

5 DISCUSSION

This study was a first look at how media can impact and potentially exacerbate stigmatizing language online regarding mental health. In our work, we focused on the Depp-Heard defamation trial that was of high public interest and conducted language analysis techniques, grounded in existing knowledge about language used in mental health stigma, to uncover how this event affected online stigma towards mental health. We found significantly greater levels of stigmatizing language along with significantly lower levels of destigmatizing language in tweets referencing personality disorders both during and after the Depp-Heard trial. Our regression analyses showed that external events can cause significantly greater prevalence of stigmatizing language relating to dehumanization and power, greater use of negative language, and decreased amounts of personal disclosure and health-related language. We also note the possibility of causes for this language change other than the Depp-Heard trial, such as general mental health discourse patterns; however, our analysis also used the stigmatizing and destigmatizing language use patterns of exactly one year prior to the trial in our causal impact analysis to control for other effects such as seasonal factors. Our causal impact analysis to find the probability that the Depp-Heard trial had a causal effect on increases in stigmatizing language. Our findings include over 97% probability that the event caused a significant increase in stigmatizing language use on Twitter about personality disorders. The Depp-Heard trial affected language towards mental health discourse at large as well;

Table 8: OLS regression results broken down from the aggregated Stigmatizing score, which included individual dictionaries of Inappropriate Labels for Mental Illness (Model 1), Mechanistic Dehumanization (Model 2), Animalistic Dehumanization (Model 3), Power (Model 4), and Risk (Model 5).

		Model 1 Inappr. Labels Mental Illness *coeff, std.err*	**Model 2 Mech. Dehuman.** *coeff, std.err*	**Model 3 Animal. Dehuman.** *coeff, std.err*	**Model 4 Power** *coeff, std.err*	**Model 5 Risk** *coeff, std.err*
Tweets containing:	**Time relative to trial**					
personality disorders	before *(baseline)*	0	0	0	0	0
in trial	during	0, 0	0.003, 0.002	0.002, 0.004	**0.336***, 0.014	0.004, 0.006
	after	0.001, 0	0.002, 0.002	**0.010***, 0.004	**0.105***, 0.015	-0.003, 0.006
all other personality disorders	before *(baseline)*	0	0	0	0	0
	during	0, 0	0.010, 0.008	0.019, 0.016	-0.005, 0.060	0.049, 0.046
	after	0,0	**0.032***, 0.008	0, 0.016	-0.092, 0.061	0.112, 0.046

* *p < 0.05*

Table 9: OLS regression results broken down from the aggregated Destigmatizing score, which included individual dictionaries Self-transcendent (Model 1), Health (Model 2), Wellness (Model 3), Well-being (Model 4), Pro-social (Model 5), Personal Pronouns (Model 6).

		Model 1 Self-transc. *coeff, s.e.*	**Model 2 Health** *coeff, s.e.*	**Model 3 Well-ness** *coeff, s.e.*	**Model 4 Well-being** *coeff, s.e.*	**Model 5 Pro-social** *coeff, s.e.*	**Model 6 Person. Pron.** *coeff, s.e.*
Tweets with:	**Time relative to trial**						
personality disorders	before *(baseline)*	0	0	0	0	0	0
in trial	during	**-0.05***, 0.01	-0.17, 0.05	**-0.01***, 0.01	**-0.03***, 0.01	**-0.02***, 0.01	**-0.48***, 0.02
	after	**-0.02***, 0.01	**0.21***, 0.06	**-0.01***, 0.01	**-0.03***, 0.01	**-0.04***, 0.01	**-0.26***, 0.02
all other personality disorders	before *(baseline)*	0	0	0	0	0	0
	during	-0.01, 0.02	0.12, 0.25	-0.02, 0.01	-0.01, 0.02	-0.01, 0.04	**0.15***, 0.06
	after	0.03, 0.02	-0.19, 0.25	0.01, 0.01	0.01, 0.02	-0.02, 0.04	**0.17***, 0.06

* *p < 0.05*

our analyses found an effect size of over 32% decrease in destigmatizing language use for tweets just mentioning "mental health" rather than any specific illness, and over 75% decrease in positive tone expression. Given continued observation of significant effects for the period of almost two months post-trial conclusion, we found that external events may have lasting effects on general mental health discourse. Other quantitative approaches, such as studying the contagion of stigmatizing perspectives using network analysis or the likelihood of a user tweeting stigmatizing language towards mental health given exposure to other users' stigmatizing posts, may also prove important for future research to assess the widespread effects of stigmatizing attitudes on Twitter across mental health discourse.

Our work has direct implications for awareness and understanding of the significant impacts of offline media events for the stigmatization of mental health. Our analysis not only provided evidence that offline events can significantly increase stigmatizing language use in online mental health discourse, but also that these effects may have longer-lasting effects extending past an event's conclusion. We argue that these findings are important for developing interventions to destigmatize mental health on online platforms by

Table 10: OLS regression results for LIWC scores of Stigmatizing (Model 1), Destigmatizing (Model 2), Negative Tone (Model 3), and Positive Tone (Model 4) in tweets containing keyword "mental health". The dependent variable for all models is the LIWC score while the predictor is the time period relative to the trial. We use Before Trial as our baseline for each regression.

		Model 1 Stigmatizing *coeff, std.err*	**Model 2 Destigmatizing** *coeff, std.err*	**Model 3 Negative Tone** *coeff, std.err*	**Model 4 Positive Tone** *coeff, std.err*
	Time period relative to trial				
Tweets with "mental health"	before (*baseline*)	0	0	0	0
	during	**0.111***, 0.00	**-0.326***, 0.00	**0.070***, 0.01	**-0.758***, 0.01
	after	**0.055***, 0.001	**-0.166***, 0.002	0.159, 0.005	**-0.253***, 0.005

* $p < 0.05$

revealing the exacerbating effects that even one-time media events have on online mental health stigma. Our findings also have practical design implications for monitoring, visualizing, and designing for timed interventions for countering rises in stigma online in response to offline events and for monitoring online community language towards sensitive subjects; likewise, these findings can be expanded upon to study how external events may also have widespread positive (i.e. destigmatizing) effects for sensitive subjects. There have been anti-stigma campaigns both in offline and online contexts for mental health causes [40, 41, 86]; our study's findings may provide an important look at the timing of such interventions and the need for anti-stigma movements during media events that surge stigma against mental illness. This growth of stigmatizing language on social media may be leading to the growth of stigmatizing perspectives in offline society, or it may be simply a reflection of existing stigmatization in society; the direction of this influence remains to be unpacked in future work. Regardless, stigmatizing language use online has been repeatedly shown to be harmful to people's self-image and well-being [48, 61, 70]. One concrete future direction to explore the real-world impacts of online stigmatization is to conduct an analysis predicting the large-scale survey responses of mental health and well-being (e.g., [59]) using the stigmatization languages in social media. Other future work may also explore the real-world effects of online stigmatization on offline help-seeking behaviors as well; for example, it is unknown how online stigmatization may affect exposed persons in therapy, which may be especially important to study in the context of personality disorders given that people with personality disorders are already susceptible to sensitivity towards societal rejection [8].

There is also important future work in studying stigmatizing language in nuanced and context-specific ways. Stigma may take many different forms. For example, Kealy and Ogrodniczuk studied the marginalization of people with borderline personality disorder, finding that stigmatization of the disorder primarily fell into categories of: (1) the incorrect claim that borderline personality disorder is unresponsive to treatment, (2) borderline personality disorder is rare, (3) borderline personality is not a real illness, (4) language choice of "borderline" sidelining the illness [54]. Different types of stigma may have differing effects on society as well as manifest differently in online discourse; future avenues for research in this field may benefit from studying stigmatization in context- and disorder-specific ways. There are also several other dimensions to mental health stigma that our work did not explore, such as demographics. Evidence suggests that gender, for example, is a primary dimension of social stigma. For example, personality disorders and eating disorders are closely associated with women while substance use is associated with men [15, 65]; these associations affect the type of stigma associated with disorders as well, such as males with borderline personality disorder being seen as more dangerous while females with borderline personality disorder are viewed with greater pity [65]. Other analysis including account information, such as tweets from accounts that are news coverage versus non-news coverage or incorporating account following/follower numbers, could also provide greater detail into how language use may vary between populations and how stigma is spread through different types of channels.

Lastly, we acknowledge that work studying how external events can exacerbate harm in online discussions carries the risk of misuse. For example, just as our work could be used for timed interventions for countering online harm like stigma, it may also be used for designing targeted campaigns to take advantage of public events to exacerbate harmful language online. However, we believe our work to have significant positive influence alongside the recent developments in online mental health awareness and interventions [60, 61, 88], and a necessary look into how media affects stigma in online communities of people with mental illness.

5.1 Limitations

Our work has several limitations worth noting. First, Twitter may not be representative of all social media platforms. Twitter may be favored by users who present feelings of greater stigma due to its potential anonymity [3], and may not be representative of general public opinion nor social media attitudes on other platforms. Additionally, the Depp-Heard trial is just one of many events that may impact online stigmatization towards mental health. However, our analysis may provide a significant starting point and contribution for the research space of external events' effects for online mental health discussion. Second, we acknowledge that our analysis relied on just one language analysis tool LIWC. However, LIWC may be a limited approach in that it cannot determine whether language

used is towards a mentioned mental health condition versus another topic (e.g. domestic violence, a specific individual). Other language models and tools (such as the popular probabilistic topic model Latent Dirichlet Allocation, or LDA) may be additionally explored for analysis in future work. Third, we acknowledge that our analysis used course time periods of 51-day long periods before, during, and after the Depp-Heard trial; using more granular time periods, such as looking at day-to-day changes in language use, as well as data collection past the 51-day post-trial period can allow deeper investigation to how language use changes over time as well as the longer-lasting effects of the event. Fourth, our work is solely a quantitative look at how language has changed on Twitter from the Depp-Heard trial, and did not include other measures of perception that may be able to be measured in mixed-methods approaches to this topic. Our study can only perceive perception towards mental health from the language choices of users. Lastly, we also note key limitations in our keyword-based approach to sampling tweets, such that we could not capture discussions about mental health or personality disorders that did not use our exact keywords. However we note that our personality disorder keywords are the formal medical terms used for these disorders; the actual effects of stigmatization that we identified in our study may be even greater in reality, given that our study did not capture tweets with derogatory, outdated, or inaccurate terms used to reference these disorders [17, 81].

6 CONCLUSION

Our work investigated the question of how and to what extent highly publicized events affect language towards mental health on social media, particularly in stigmatizing or destigmatizing ways. Evaluating Twitter data from the widely covered news event of the Johnny Depp - Amber Heard trial and its discussion around personality disorders, we found significantly greater use of stigmatizing language in Twitter posts regarding personality disorders as well as referencing mental health generally. We found significant effects both during the time period of the trial and even for the nearly two months after the trial's conclusion. Our results indicate that offline media has significant effects on online mental health discourse, and are key for motivating effective online interventions to mitigate media's potential to exacerbate mental health stigma on social network platforms.

ACKNOWLEDGMENTS

We sincerely thank our colleagues from Social AI Group at Carnegie Mellon University for their feedback and support, including Jordan Taylor and Milo Fang. This work was supported by the Center for Machine Learning and Health at Carnegie Mellon University and the National Science Foundation (NSF) under Award No. 1939606, 2001851, 2000782, and 1952085.

REFERENCES

[1] Jo Ann A Abe. 2011. Positive emotions, emotional intelligence, and successful experiential learning. *Personality and Individual Differences* 51, 7 (2011), 817–822.
[2] Sania Ahmed, Desheane Newman, and M Yalch. 2021. The Stigma of Borderline Personality Disorder. *Advances in psychology research* 145 (2021), 59–78.
[3] Miguel Angel Alvarez-Mon, Angel Asunsolo Del Barco, Guillermo Lahera, Javier Quintero, Francisco Ferre, Victor Pereira-Sanchez, Felipe Ortuño, and Melchor Alvarez-Mon. 2018. Increasing interest of mass communication media and the general public in the distribution of tweets about mental disorders: observational study. *Journal of medical internet research* 20, 5 (2018), e9582.
[4] Lauren Alvis, N Shook, and B Oosterhoff. 2020. Adolescents' prosocial experiences during the covid-19 pandemic: Associations with mental health and community attachments. *PsyArXiv Preprints* (2020).
[5] Nazanin Andalibi, Margaret E Morris, and Andrea Forte. 2018. Testing waters, sending clues: Indirect disclosures of socially stigmatized experiences on social media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.
[6] Julio Arboleda-Flórez. 2002. What causes stigma? *World Psychiatry* 1, 1 (2002), 25.
[7] American Psychiatric Association. 2020. Stigma, Prejudice and Discrimination Against People with Mental Illness. https://www.psychiatry.org/patients-families/stigma-and-discrimination
[8] Ron B Aviram, Beth S Brodsky, and Barbara Stanley. 2006. Borderline personality disorder, stigma, and treatment implications. *Harvard review of psychiatry* 14, 5 (2006), 249–256.
[9] Christopher A Bail, Taylor W Brown, and Marcus Mann. 2017. Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. *American Sociological Review* 82, 6 (2017), 1188–1213.
[10] Erin O'Carroll Bantum and Jason E Owen. 2009. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment* 21, 1 (2009), 79.
[11] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 450–453.
[12] Oliver Bonnington and Diana Rose. 2014. Exploring stigmatisation among people diagnosed with either bipolar disorder or borderline personality disorder: A critical realist analysis. *Social Science & Medicine* 123 (2014), 7–17.
[13] Matt Bowen and Andy Lovell. 2019. Stigma: the representation of mental health in UK newspaper Twitter feeds. *Journal of Mental Health* (2019).
[14] Matt Laurence Bowen. 2016. Stigma: Content analysis of the representation of people with personality disorder in the UK popular press, 2001–2012. *International journal of mental health nursing* 25, 6 (2016), 598–605.
[15] Guy Boysen, Ashley Ebersole, Robert Casner, and Nykhala Coston. 2014. Gendered mental disorders: Masculine and feminine stereotypes about mental disorders and their relation to stigma. *The Journal of Social Psychology* 154, 6 (2014), 546–565.
[16] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* (2015), 247–274.
[17] Lauren M Broyles, Ingrid A Binswanger, Jennifer A Jenkins, Deborah S Finnell, Babalola Faseru, Alan Cavaiola, Marianne Pugatch, and Adam J Gordon. 2014. Confronting inadvertent stigma and pejorative language in addiction scholarship: a recognition and response. , 217–221 pages.
[18] Alexandra Budenz, Ann Klassen, Jonathan Purtle, Elad Yom Tov, Michael Yudell, and Philip Massey. 2020. Mental illness and bipolar disorder on Twitter: implications for stigma and social support. *Journal of Mental Health* 29, 2 (2020), 191–199.
[19] Kirsten Catthoor, Dine J Feenstra, Joost Hutsebaut, Didier Schrijvers, and Bernard Sabbe. 2015. Adolescents with personality disorders suffer from severe psychiatric stigma: evidence from a sample of 131 patients. *Adolescent health, medicine and therapeutics* 6 (2015), 81.
[20] Patricia A Cavazos-Rehg, Melissa J Krauss, Shaina J Costello, Nina Kaiser, Elizabeth S Cahn, Ellen E Fitzsimmons-Craft, and Denise E Wilfley. 2019. "I just want to be skinny.": A content analysis of tweets expressing eating disorder symptoms. *PloS one* 14, 1 (2019), e0207506.
[21] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* 1 (2007), 343–359.
[22] Cindy K Chung and James W Pennebaker. 2012. Linguistic inquiry and word count (LIWC): pronounced "Luke,"... and other useful facts. In *Applied natural language processing: Identification, investigation and resolution*. IGI Global, 206–229.
[23] Sabrina Cipolletta, Riccardo Votadoro, and Elena Faccio. 2017. Online support for transgender people: an analysis of forums and social networks. *Health Soc. Care Community* 25, 5 (Sept. 2017), 1542–1551.
[24] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.
[25] 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. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 1–10.
[26] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American psychologist* 59, 7 (2004), 614.
[27] Patrick W Corrigan, Karina J Powell, and Patrick J Michaels. 2013. The effects of news stories on the stigma of mental illness. *The Journal of nervous and mental disease* 201, 3 (2013), 179–182.

[28] Patrick W Corrigan and Abigail Wassel. 2008. Understanding and influencing the stigma of mental illness. *Journal of psychosocial nursing and mental health services* 46, 1 (2008), 42–48.

[29] Laura de Anta, Miguel Angel Alvarez-Mon, Miguel A Ortega, Cristina Salazar, Carolina Donat-Vargas, Javier Santoma-Vilaclara, Maria Martin-Martinez, Guillermo Lahera, Luis Gutierrez-Rojas, Roberto Rodriguez-Jimenez, et al. 2022. Areas of interest and social consideration of antidepressants on english tweets: a natural language processing classification study. *Journal of personalized medicine* 12, 2 (2022), 155.

[30] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*. 47–56.

[31] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *ICWSM* 8, 1 (May 2014), 71–80.

[32] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[33] Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *Eleventh International AAAI Conference on Web and Social Media*.

[34] Munmun De Choudhury and Emre Kıcıman. 2017. The Language of Social Support in Social Media and its Effect on Suicidal Ideation Risk. *Proceedings of the International AAAI Conference on Web and Social Media* 2017 (May 2017), 32–41.

[35] Michael A DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook' Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.

[36] John F Dovidio, Peter Glick, and Laurie A Rudman. 2008. *On the nature of prejudice: Fifty years after Allport*. John Wiley & Sons.

[37] Amy L Drapalski, Alicia Lucksted, Paul B Perrin, Jennifer M Aakre, Clayton H Brown, Bruce R DeForge, and Jennifer E Boyd. 2013. A model of internalized stigma and its effects on people with mental illness. *Psychiatric Services* 64, 3 (2013), 264–269.

[38] Louisa M Drost and Gerard M Schippers. 2015. Online support for children of parents suffering from mental illness: A case study. *Clinical child psychology and psychiatry* 20, 1 (2015), 53–67.

[39] Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc* 21, 21 (2013), 591–643.

[40] Sara Evans-Lacko, Elizabeth Corker, Paul Williams, Claire Henderson, and Graham Thornicroft. 2014. Effect of the Time to Change anti-stigma campaign on trends in mental-illness-related public stigma among the English population in 2003–13: an analysis of survey data. *The Lancet Psychiatry* 1, 2 (2014), 121–128.

[41] Sara Evans-Lacko, Jillian London, Sarah Japhet, Nicolas Rüsch, Clare Flach, Elizabeth Corker, Claire Henderson, and Graham Thornicroft. 2012. Mass social contact interventions and their effect on mental health related stigma and intended discrimination. *BMC public health* 12, 1 (2012), 1–8.

[42] Anna Fang and Haiyi Zhu. 2022. Matching for Peer Support: Exploring Algorithmic Matching for Online Mental Health Communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2022).

[43] Kaelyn Forde. 2020. By sharing their own struggles, celebs help teens tear down mental health stigma. https://www.ozy.com/the-new-and-the-next/by-sharing-their-own-struggles-celebs-help-teens-tear-down-mental-health-stigma/253556/

[44] Adrian Furnham and Hina Dadabhoy. 2012. Beliefs about causes, behavioural manifestations and treatment of borderline personality disorder in a community sample. *Psychiatry Research* 197, 3 (2012), 307–313.

[45] Ryan J Gallagher, Elizabeth Stowell, Andrea G Parker, and Brooke Foucault Welles. 2019. Reclaiming stigmatized narratives: The networked disclosure landscape of# MeToo. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.

[46] Ales Grambal, Jan Prasko, Dana Kamaradova, Klara Latalova, Michaela Holubova, Marketa Marackova, Marie Ociskova, and Milos Slepecky. 2016. Self-stigma in borderline personality disorder–cross-sectional comparison with schizophrenia spectrum disorder, major depressive disorder, and anxiety disorders. *Neuropsychiatric disease and treatment* 12 (2016), 2439.

[47] Keith N Hampton, Lauren Sessions Goulet, Lee Rainie, and Kristen Purcell. 2011. *Social networking sites and our lives*. Vol. 1. Pew Internet & American Life Project Washington, DC, Washington, D.C.

[48] Chelsea A Heuer, Kimberly J McClure, and Rebecca M Puhl. 2011. Obesity stigma in online news: a visual content analysis. *Journal of health communication* 16, 9 (2011), 976–987.

[49] Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology* 60, 11 (2009), 2169–2188.

[50] Qihao Ji and Arthur A Raney. 2020. Developing and validating the self-transcendent emotion dictionary for text analysis. *PloS one* 15, 9 (2020), e0239050.

[51] Xiang Ji, Soon Chun, Zhi Wei, and James Geller. 2015. Twitter sentiment classification for measuring public health concerns. *Social Network Analysis and Mining* 5, 1 (2015), 1–25.

[52] Adam J Joseph, Neeraj Tandon, Lawrence H Yang, Ken Duckworth, John Torous, Larry J Seidman, and Matcheri S Keshavan. 2015. # Schizophrenia: Use and misuse on Twitter. *Schizophrenia research* 165, 2-3 (2015), 111–115.

[53] Pew Research Center: Journalism and Media Staff. 2008. *Health news coverage in the US media*. Vol. 1. Pew Internet & American Life Project Washington, DC, Washington, D.C.

[54] David Kealy and John S Ogrodniczuk. 2010. Marginalization of borderline personality disorder. *Journal of Psychiatric Practice®* 16, 3 (2010), 145–154.

[55] Jooho Kim and Makarand Hastak. 2018. Social network analysis: Characteristics of online social networks after a disaster. *International journal of information management* 38, 1 (2018), 86–96.

[56] Taewan Kim, Mintra Ruensuk, and Hwajung Hong. 2020. In helping a vulnerable bot, you help yourself: Designing a social bot as a care-receiver to promote mental health and reduce stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[57] Anat Klin and Dafna Lemish. 2008. Mental disorders stigma in the media: Review of studies on production, content, and influences. *Journal of health communication* 13, 5 (2008), 434–449.

[58] Stephanie Knaak, Andrew CH Szeto, Kathryn Fitch, Geeta Modgill, and Scott Patten. 2015. Stigma towards borderline personality disorder: effectiveness and generalizability of an anti-stigma program for healthcare providers using a pre-post randomized design. *Borderline personality disorder and emotion dysregulation* 2, 1 (2015), 1–8.

[59] Robert E Kraut, Han Li, and Haiyi Zhu. 2022. Mental health during the COVID-19 pandemic: Impacts of disease, social isolation, and financial stressors. *PloS one* 17, 11 (2022), e0277562.

[60] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*. 85–94.

[61] Sarah E Lageson and Shadd Maruna. 2018. Digital degradation: Stigma management in the internet age. *Punishment & Society* 20, 1 (2018), 113–133.

[62] Winnie WS Mak, Cecilia YM Poon, Loraine YK Pun, and Shu Fai Cheung. 2007. Meta-analysis of stigma and mental health. *Social science & medicine* 65, 2 (2007), 245–261.

[63] David M Markowitz and Paul Slovic. 2020. Social, psychological, and demographic characteristics of dehumanization toward immigrants. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9260–9269.

[64] David M Markowitz and Paul Slovic. 2021. Why we dehumanize illegal immigrants: A US mixed-methods study. *Plos one* 16, 10 (2021), e0257912.

[65] Sara R Masland and Kaylee E Null. 2022. Effects of diagnostic label construction and gender on stigma about borderline personality disorder. *Stigma and Health* 7, 1 (2022), 89.

[66] J A Naslund, K A Aschbrenner, L A Marsch, and S J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. , 113–122 pages.

[67] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122.

[68] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*.

[69] Thomas Niederkrotenthaler, Ulrich S Tran, Madelyn Gould, Mark Sinyor, Steven Sumner, Markus J Strauss, Martin Voracek, Benedikt Till, Sean Murphy, Frances Gonzalez, et al. 2021. Association of Logic's hip hop song "1-800-273-8255" with Lifeline calls and suicides in the United States: interrupted time series analysis. *bmj* 375 (2021).

[70] Kirsten Noack, Nusha Balram Elliott, Eugenia Canas, Kathleen Lane, Andrea Paquette, Jeanne-Michelle Lavigne, and Erin E Michalak. 2016. Credible, centralized, safe, and stigma–free: What youth with bipolar disorder want when seeking health information online. *UBC Medical Journal* 8, 1 (2016).

[71] Scott Parrott, Andrew C Billings, Samuel D Hakim, and Patrick Gentile. 2020. From# endthestigma to# realman: Stigma-challenging social media responses to NBA players' mental health disclosures. *Communication Reports* 33, 3 (2020), 148–160.

[72] Alina Pavlova and Pauwke Berkers. 2020. Mental health discourse and social media: Which mechanisms of cultural power drive discourse on Twitter. *Social Science & Medicine* 263 (2020), 113250.

[73] Geraldine S Perry, Letitia R Presley-Cantrell, and Satvinder Dhingra. 2010. Addressing mental health promotion in chronic disease prevention and health promotion. , 2337–2339 pages.

[74] Greg Philo, Lesley Henderson, and Katie McCracken. 2010. Making drama out of a crisis: Authentic portrayals of mental illness in TV drama. *London: Shift* (2010).

[75] Vanessa Pinfold, Graham Thornicroft, Peter Huxley, and Paul Farmer. 2005. Active ingredients in anti-stigma programmes in mental health. *International Review of Psychiatry* 17, 2 (2005), 123–131.

[76] S. Platten, R. Haji, and R. L. Boyd. 2022. Humanizing and dehumanizing themes of Muslims surrounding 9/11: Computerized language analysis.
[77] Elizabeth B Raposa, Holly B Laws, and Emily B Ansell. 2016. Prosocial behavior mitigates the negative effects of stress in everyday life. *Clinical Psychological Science* 4, 4 (2016), 691–698.
[78] Nicola J Reavley and Pamela D Pilkington. 2014. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. *PeerJ* 2 (2014), e647.
[79] Eugenia Ha Rim Rho, Oliver L Haimson, Nazanin Andalibi, Melissa Mazmanian, and Gillian R Hayes. 2017. Class confessions: Restorative properties in online experiences of socioeconomic stigma. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3377–3389.
[80] Patrick Robinson, Daniel Turk, Sagar Jilka, and Matteo Cella. 2019. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. *Social psychiatry and psychiatric epidemiology* 54, 1 (2019), 51–58.
[81] D. Rose, G. Thornicroft, V. Pinfold, and A. Kassam. 2007. 250 labels used to stigmatise people with mental illness. *BMC health services research* 97, 7 (2007).
[82] Wulf Rössler. 2016. The stigma of mental disorders: A millennia-long history of social exclusion and prejudices. *EMBO reports* 17, 9 (2016), 1250–1253.
[83] Nicolas Rüsch, Klaus Lieb, Martin Bohus, and Patrick W Corrigan. 2006. Self-stigma, empowerment, and perceived legitimacy of discrimination among women with mental illness. *Psychiatric services* 57, 3 (2006), 399–402.
[84] Koustuv Saha, Sang Chan Kim, Manikanta D Reddy, Albert J Carter, Eva Sharma, Oliver L Haimson, and Munmun De Choudhury. 2019. The language of LGBTQ+ minority stress experiences on social media. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–22.
[85] Gaia Sampogna, I Bakolis, S Evans-Lacko, E Robinson, G Thornicroft, and C Henderson. 2017. The impact of social marketing campaigns on reducing mental health stigma: Results from the 2009–2014 Time to Change programme. *European Psychiatry* 40 (2017), 116–122.
[86] G Schomerus, MC Angermeyer, SE Baumeister, S Stolzenburg, BG Link, and JC Phelan. 2016. An online intervention using information on the mental health-mental illness continuum to reduce stigma. *European Psychiatry* 32 (2016), 21–27.
[87] Lindsay Sheehan, Katherine Nieweglowski, and Patrick Corrigan. 2016. The stigma of personality disorders. *Current Psychiatry Reports* 18, 1 (2016), 1–7.
[88] Victor Suarez-Lledo and Yelena Mejova. 2022. Behavior Change Around an Online Health Awareness Campaign: A Causal Impact Study. *Frontiers in Public Health* 10 (2022).
[89] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.
[90] William Tov, Kok Leong Ng, Han Lin, and Lin Qiu. 2013. Detecting well-being via computerized content analysis of brief diary entries. *Psychological assessment* 25, 4 (2013), 1069.
[91] Barry Wellman, Anabel Quan Haase, James Witte, and Keith Hampton. 2001. Does the Internet Increase, Decrease, or Supplement Social Capital?: Social Networks, Participation, and Community Commitment. *Am. Behav. Sci.* 45, 3 (Nov. 2001), 436–455.
[92] Rob Whitley and Sarah Berry. 2013. Trends in newspaper coverage of mental illness in Canada: 2005–2010. *The Canadian Journal of Psychiatry* 58, 2 (2013), 107–112.
[93] Meryl Williams and Judy Taylor. 1995. Mental illness: Media perpetuation of stigma. *Contemporary Nurse* 4, 1 (1995), 41–46.
[94] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–23.
[95] Kathryn Schaefer Ziemer and Gizem Korkmaz. 2017. Using text to predict psychological and physical health: A comparison of human raters and computerized text analysis. *Computers in Human Behavior* 76 (2017), 122–127.
[96] Anastazia Zunic, Padraig Corcoran, Irena Spasic, et al. 2020. Sentiment analysis in health and well-being: systematic review. *JMIR medical informatics* 8, 1 (2020), e16023.

A FULL RESULTS FOR VOLUME OF TWEETS (SECTION 4.1.1)

Table 11: OLS regression results for the number of likes, retweets, and replies per tweet.

		Model 1 **# Likes per tweet** *coeff, std.err*	**Model 2** **# RTs per tweet** *coeff, std.err*	**Model 3** **# Replies per tweet** *coeff, std.err*
Tweets with:	**Time period relative to trial**			
personality disorders in trial	before *(baseline)*	0	0	0
	during	2.674, 4.249	0.377, 0.461	**0.107***, 0.043
	after	-2.708, 4.541	-0.221, 0.492	0.017, 0.046
all other personality disorders	before *(baseline)*	0	0	0
	during	5.062, 3.241	1.182, 0.810	**0.117***, 0.053
	after	1.169, 3.259	0.357, 0.815	0.108, 0.053