# Quantifying language changes surrounding mental health on Twitter

A Thesis Presented

by

Anne Marie Stupinski

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Master of Science
Specializing in Complex Systems and Data Science

August, 2021

Defense Date: June 24th, 2021
Thesis Examination Committee:

Christopher Danforth, Ph.D., Advisor
Peter Sheridan Dodds, Ph.D., Advisor
Matthew Price, Ph.D., Chairperson
Jeremiah Onaolapo, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

# ABSTRACT

Mental health challenges are thought to afflict around 10% of the global population each year, with many going untreated due to stigma and limited access to services. Here, we explore trends in words and phrases related to mental health through a collection of 1- , 2-, and 3-grams parsed from a data stream of roughly 10% of all English tweets since 2012. We examine temporal dynamics of mental health language, finding that the popularity of the phrase 'mental health' increased by nearly two orders of magnitude between 2012 and 2018. We observe that mentions of 'mental health' spike annually and reliably due to mental health awareness campaigns, as well as unpredictably in response to mass shootings, celebrities dying by suicide, and popular fictional stories portraying suicide. We find that the level of positivity of messages containing 'mental health', while stable through the growth period, has declined recently. Finally, we use the ratio of original tweets to retweets to quantify the fraction of appearances of mental health language due to social amplification. Since 2015, mentions of mental health have become increasingly due to retweets, suggesting that stigma associated with discussion of mental health on Twitter has diminished with time.

# DEDICATION

*To my pandemic roommates Vanessa, Megan, and Kaitlin, who have supported me immensely throughout this project, this past year in general, and who always remind me to take care of my own mental health.*

# Acknowledgements

I am extremely grateful to my advisors, Chris Danforth and Peter Dodds, who consistently inspire me to work on projects that are for the greater good (or so we hope), for their support and feedback throughout this project, and for modeling and encouraging work-life balance for the members of our lab. I would like to thank my committee chair Matt Price for his long-time support since my undergraduate years, and for his insightful feedback and perspective on this project. I am also grateful to committee member Jeremiah Onaolapo, who in the short time I've known him has inspired new ideas and changed the way I think about social media research. Thank you to every member of the Computational Story Lab for the feedback and support, and especially to Thayer Alshaabi and Michael Arnold, whose technical assistance, help with understanding difficult concepts, and friendship has helped ease the transition to graduate student life. To Kelly Gothard, with whom it has been wonderful to grow alongside throughout the past years. To Lincoln Pierce, thank you for your support and encouragement, and for providing joyful moments and mental breaks from academic work. And to my family, for providing a loving environment for me to be myself and thrive best I can. Your support throughout all of life's obstacles means the world to me. Finally, thank you to the Massachusetts Mutual Life Insurance Company for the financial support and opportunity to work on this project, and to the Vermont Advanced Computing Core for the resources to do so.

# TABLE OF CONTENTS

# List of Figures

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

Recent estimates place 1 in 10 people globally as experiencing from some form of mental illness [38], with 1 in 30 suffering from depression [25]. These rates put mental illness among the leading causes of ill-health and disability worldwide. Moreover, rates of mental health disorders and deaths by suicide have increased in recent years, especially among young people [28].

Since the beginning of the COVID-19 pandemic and the subsequent social isolation, there have been recordings of drastic declines in physical activity and time spent socializing, and coinciding increases in screen time and symptoms of depression [22]. Google searches for mental health related topics also increased in the first weeks of the pandemic, leveling out after more information regarding stay-at-home orders were released [24]. Following March 2020, there has also been a measured increase in suicidal ideation that is associated with elevated reports of isolation [20]. The service Crisis Text Line reported receiving a higher than average volume of messages for every day following March 16th in the year 2020, with the main topics being anxiety, depression, grief, and eating disorders [1]. Price *et al.* [34] also found that daily "doomscrolling"— repeatedly consuming negative news and media content online—was associated with same-day increases in depression and PTSD. These effects were larger among those with a prior history of psychopathology and trauma exposure. The pandemic also influenced what content people discussed on social media, with users shifting away from "self-focused" perspectives and towards more "other-focused" topics that used to be vulnerable or taboo to discuss [30]. A survey of American adults during the pandemic found that the depth of distressing self-disclosures posted online could be predicted by a user's perceived anonymity, visibility control, and closeness to their audience [42].

Historically, the availability of mental health treatment services has not meet the demand for such [15]. Mental health care also experiences a paradox of being over-diagnosed yet under-supported, with some symptoms and disorders being readily medicated despite not being understood and accepted socially [19]. Furthermore, many who would benefit from mental health services do not seek or participate in care, as they are either unaware of such services, are unable to afford them, or the stigma associated with seeking treatment proves too great a barrier [11]. In fact, two-thirds of people with a known mental disorder do not seek help from a health professional [32].

While stigma has proven to be a significant barrier to receiving treatment from formal (e.g., psychiatrists, counselors) and informal sources (e.g., family and friends), the COVID-19 pandemic and subsequent isolation have spurred awareness of mental illness and discussion on this topic in public forums such as social media. Measuring changes in this conversation, we aim to quantify the hypothesized increase in discussions and awareness, and the corresponding reduction in stigma around mental illness. Our findings suggest that the number of mental health conversations on Twitter have substantially increased in recent years, particularly on dates associated with either awareness campaigns or tragedies. We also examine social attention and expressed happiness in an attempt to piece together how this conversation has shifted in the past decade.

# Chapter 2

# Related Work

Many researchers have used social media platforms in order to explore and understand dynamics of healthcare discussion [23]. Several reviews have been done on mental health discussion in particular, finding that social media is a viable platform for users to discuss mental health and feel supported, although privacy risks and ethical concerns of research applications exist as well [8, 31]. A study by De Choudhury *et al.* [12] analyzed the Twitter activity of individuals diagnosed with depression, along with clinically validated measures, in order to predict users who may be at risk of the mental illness. Reece and Danforth [37] used tweets posted prior to a user's diagnosis date to identify social media content associated with the onset of depression.

De Choudhury *et al.* has also worked on predicting postpartum depression in new mothers, using Facebook activity, linguistic expression in status updates, and demographic survey data [13]. Using consenting Instagram users' photos, Reece *et al.* [36] identified distinct predictive markers of depression in the images posted by individuals previously diagnosed with depression by a psychiatrist. Work by Coppersmith *et al.* [9] classified online users who suffer from Post Traumatic Stress Disorder by using self-disclosing messages on Twitter.

Another study using self-disclosures proposes a classifier to distinguish between Twitter users suffering from mental illness from those who are not, using messages collected from individuals self-reporting ten various mental illnesses [10]. Using the self-disclosure classification methods proposed by Coppersmith *et al.* [9], Bathina *et al.* [6] reported that Twitter users with a diagnosis of depression show higher levels of distorted thinking in their posts when compared to a random sample of messages.

Researchers have used mental health support threads on Reddit to examine the shift of suicidal ideation on social media, identifying users who are more likely than

others to make this transition from the typical content in these threads [14]. Analysis of text-based crisis counseling conversations found actionable strategies associated with more effective counseling, such as adaptability, dealing with ambiguity, creativity, and change in perspective [4]. While developments in predicting mental health states provide an opportunity for early detection and treatment, they come with several ethical concerns, such as incorrect predictions, involvement of bad actors, and potential biases [7]. Social media users also hold negative attitudes towards the concept of automated well-being interventions prompted by emotion recognition, stating that any automated message could not have the personal, human attributes necessary for such an interaction to be successful [40]. They also view emotion recognition in general as invasive, scary, and a loss of their control and autonomy, as people view emotions as insights to behavior that are vulnerable and prone to manipulation [5].

Several other studies have more directly examined attitudes towards those with mental illnesses, attempting to measure the stigma towards these individuals that exists in social communities. Rose *et al.* [41] sought to investigate the extent of stigma and treatment avoidance in 14-year-old students in relation to how they refer to people with mental illness, finding that the majority of phrases used fit into the theme "popular derogatory terms".

Reavley and Pilkington took a qualitative approach to monitoring stigma on Twitter, collecting tweets over a 7-day period that contain the hashtags #depression or #schizophrenia and categorizing them [35]. These tweets were coded based on the attitude they indicated (stigmatizing, personal experience, supportive, neutral, or anti-stigma) and on their content (awareness promotion, research findings, resources, advertising, news media, or personal opinion). Their findings show that tweets related

to depression mostly contain resources or advertisements for mental health services, while tweets on schizophrenia contain awareness promotion or research findings. The percentage of tweets showing stigmatizing attitudes was 5%, and most of these showed inaccurate beliefs about schizophrenia being multiple personality disorder.

Another recent study by Robinson *et al.* [39] used Twitter to investigate attitudes towards a variety of mental and physical health conditions, finding that mental health conditions were more stigmatized and trivialized than physical ones, especially among mentions of schizophrenia and obsessive compulsive disorder. Examinations of Chinese social media posts on the platform Weibo find that roughly six percent of posts include stigmatizing attitudes towards depression, reflecting beliefs that depression is "a sign of personal weakness" as well as "not a real medical illness" [27].

The goal of our present study is to contribute to this growing body of work, previously largely focused on individuals, by using a data-driven approach to examine the collective conversation over a full decade. Using messages from Twitter, we analyze the conversation around mental health by examining the growth of public attention, the divergence of language from general messages and the associated happiness shifts, and the rise of ambient words or phrases.

# CHAPTER 3

# DATA AND METHODS

## 3.1  DATA

Twitter is a real-time source of information on a wide variety of topics. Since most tweets are public and the platform is commonly used by both adults and young people, estimates of public opinion based on the platform can complement survey-based measures. Complicating this effort, Twitter's user base is limited [33], skewing slightly younger and more politically left-leaning than the US population overall. For these reasons and many others, Twitter messages will fail to capture many aspects of human behavior.

In particular, mental health discourse is a sensitive, often personal topic that many individuals will avoid discussing publicly. Nevertheless, Twitter is a valuable social ecosystem from which we can sketch a rough portrait of the existing conversation around mental health. And given that social media lowers the barrier for individuals to join difficult conversations, especially with Twitter allowing users to sign up anonymously, it is a promising source of unstructured language data describing the changing experience of a stigmatized group.

The source of data for the present study is Twitter's Decahose API, filtered for English messages, from which we collect a 10% random sample of all public tweets between January 2012 and January 2021. This collection is separated into three corpora consisting of (a) all tweets, (b) tweets containing the phrase "mental health", and (c) tweets containing a small set of phrases related to mental health. Statistics and timeseries comparisons between corpora are made as follows.

## 3.2   $n$-GRAMS

### 3.2.1   GENERAL TWITTER

To explore trends in the appearance of words, we process messages into 1-, 2- and 3-grams, where a 1-gram is a one-word phrase, 2-gram is a two-word phrase, and so on, using the $n$-gram popularity dataset Storywrangler [2].

For each day, we count the number of times each unique $n$-gram appears in tweets, and determine usage frequencies relative to the appearance of other phrases on Twitter. We rank $n$-grams by descending order of count; $n$-grams with a low rank value assigned to phrases appear on Twitter very often, while those with a high rank value appear rarely. For example, the 1-gram 'a' has a median rank of 1, as it is typically the most commonly used word in the English language. Meanwhile, the 1-gram 'America' is less common, with a median rank of 990 [17]. In order to better visualize this concept of descending count in the figures to follow, we will plot rank on an inverted axis.

### 3.2.2   MENTAL HEALTH COLLECTION

To explore the specific language used when discussing mental health on Twitter, we compile a separate collection of $n$-grams from tweets related to this topic. Restricting to messages that contain the 2-gram "mental health", we create $n$-grams in the same fashion as previously described, determining their usage frequency within this anchor set and ranking phrases by descending order of counts. Summary statistics for key dates in this new dataset compared to the general 1-grams dataset are shown in

|  | 2012-02-08 | | 2014-01-28 | | 2018-01-31 | |
|---|---|---|---|---|---|---|
|  | MH | General | MH | General | MH | General |
| Unique 1-grams | $3.0 \times 10^3$ | $1.7 \times 10^7$ | $1.6 \times 10^3$ | $2.4 \times 10^7$ | $4.9 \times 10^4$ | $2.1 \times 10^7$ |
| Total 1-grams | $3.0 \times 10^4$ | $3.1 \times 10^8$ | $2.3 \times 10^4$ | $4.9 \times 10^8$ | $4.4 \times 10^6$ | $5.4 \times 10^8$ |
| Total 1-grams (no RTs) | $9.3 \times 10^3$ | $2.2 \times 10^8$ | $1.5 \times 10^5$ | $2.9 \times 10^8$ | $2.6 \times 10^5$ | $1.6 \times 10^8$ |

*Table 3.1:* **Summary statistics of the mental-health $n$-gram dataset compared to the general Twitter $n$-gram dataset on three individual days.** *Dates shown correspond to 'Bell Let's Talk' Day, an annual fundraising and awareness campaign, which is also coincident with the annual peak in conversation regarding mental health. Unique 1-grams enumerate the set of distinct words found in tweets on these dates, reflecting roughly 10% of all tweets. The total 1-grams row shows the sum of counts of each unique 1-gram, and total 1-grams (no retweets) is the sum of the counts of 1-grams in tweets not including any messages that were retweeted. In 2012, roughly 1 in 10,000 messages referenced mental health. In 2018, the rate increased to roughly 1 in 100 messages.*

Table 3.1. We also compute the aggregated frequency and rank of $n$-grams over each year.

Using these datasets, namely counts of phrases in all tweets (General) vs. counts of phrases in tweets containing "mental health" (MH), we analyze changes in the conversation surrounding mental health through time. The dynamics of several other phrases related to mental health are analyzed as well, but we focus primarily on "mental health" as a representative example of such phrases rather than attempting to exhaustively gather all related content.

# CHAPTER 4

# GROWTH OF COLLECTIVE ATTENTION

## 4.1 Mental Health Discourse

Public awareness and education on an issue is an important step in reducing negative attitudes, as a major component of stigma is a lack of knowledge [11]. In order to understand the general public's level of awareness of mental health issues, we quantify the frequency at which people on Twitter have discussions about the topic of mental health. Using Twitter $n$-gram data, we construct a rank timeseries of the 2-gram "mental health" on a logarithmic axis, which can be seen in Fig. 4.1.

We find that this 2-gram increased in rank by nearly two orders of magnitude between 2012 and 2018. For the first four years, only a handful of dates resulted in ranks for "mental health" more popular than the overall median, while for the final four years, only a few dates result in ranks indicating less attention than the median. This substantial increase is evidence that the conversation around mental health is happening far more frequently.

We also examine the positivity of this conversation, calculating the "ambient happiness" score of messages mentioning the phrase "mental health" for each day, which is also shown in Fig. 4.1. Ambient happiness scores for each day are computed by averaging the scores of each word that appear in a message with "mental health" for a given day, using the labMT dictionary [16]. While the rank of this 2-gram has increased over the past decade, the ambient happiness of these messages have decreased.

Examining the daily behavior of these timeseries, several dates emerge where either the rank or ambient happiness deviate largely from their baseline behavior. In Fig. 4.1, key events associated with large spikes or drops in the timeseries are
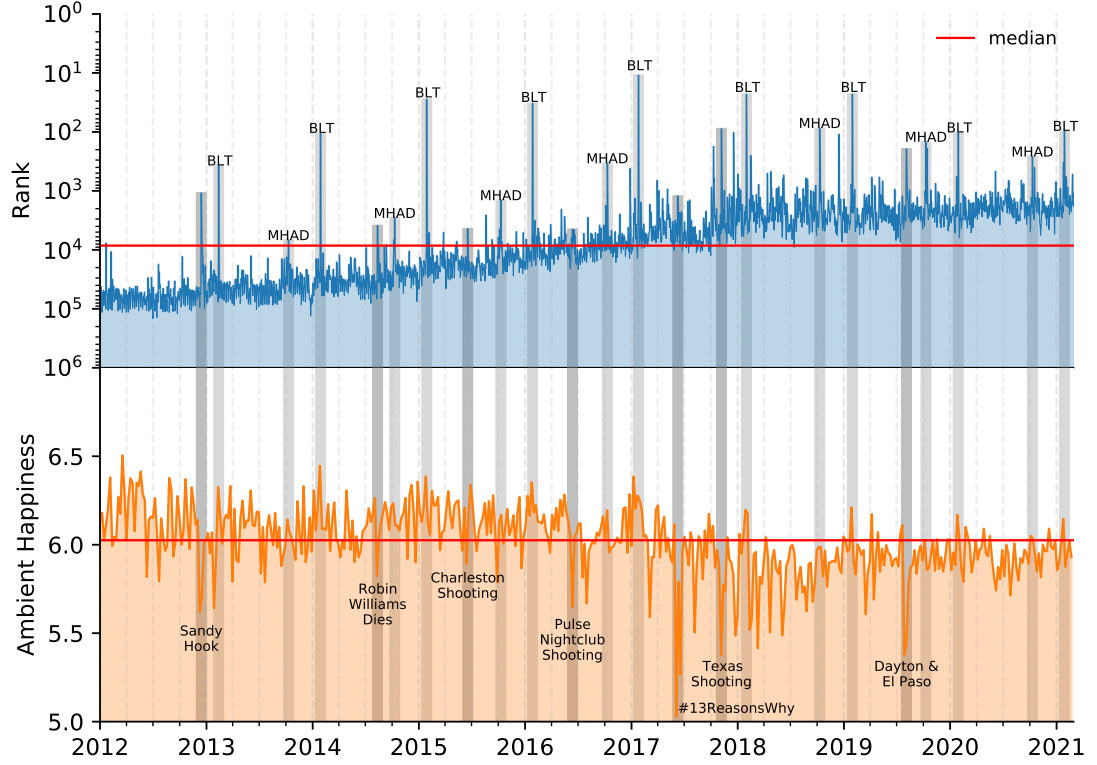
Figure 4.1: **Timeline of mental health discourse on Twitter.** *The top panel shows the rank timeseries of the 2-gram "mental health" over the past decade on a logarithmic axis. Rank is determined by ordering 2-grams in descending order of counts for each day, and plotted on an inverted axis. The median rank value of the timeseries is highlighted by a horizontal red line. Between 2012 and 2018, the phrase increased in rank by nearly two orders of magnitude, reflecting a dramatic increase in the discussion of mental health on Twitter. The bottom panel shows the "ambient happiness" of all messages containing the 2-gram "mental health" for each day over the same time period. For clarity, this data is shown as a weekly rolling average, and again the median is highlighted by a red horizontal line. Ambient happiness remained roughly constant during the period of increasing volume, but has dropped since 2017. Across both panels, key dates are highlighted in grey and annotated with the associated event. These are dates that led to large spikes or drops in either timeseries. Annually occurring events, such as Bell Let's Talk (BLT) or Mental Health Awareness Day (MHAD), are shown with light grey, and unexpected events are highlighted with a darker grey. Ambient happiness dips tend to correspond to mass shootings, with the lowest period coinciding with discussion of the Netflix series "13 Reasons Why."*

14

highlighted across both panels. Awareness events such as Bell Let's Talk (BLT) and Mental Health Awareness Day (MHAD) contribute to the large, annual spikes in rank beginning in 2013.

Bell Let's Talk, falling on the last Wednesday of January each year, was started by the Canadian company Bell Telephones, and aims to bring awareness to the general public about mental health issues by donating five cents for each tweet using their hashtag '#BellLetsTalk'. The 2-gram 'mental health' reached its highest rank ever on Bell Let's Talk day in 2017, peaking at the 18th most popular phrase compared to all other 2-grams on Twitter that day.

Other spikes in rank, and concurrent drops in ambient happiness, occurred on dates with national tragedies such as mass shooting events or celebrity deaths. The largest drop in ambient happiness occurred in 2017, immediately following the death of a teen that was connected to the Netflix series "13 Reasons Why" [26]. Looking at the events that sparked more conversation around the topic of mental health, and their associated levels of ambient happiness, awareness campaigns tend to lead to a rise in ambient happiness, while the unexpected events, of which all would be considered tragedies, lead to drops in ambient happiness.

Looking further into the language used on these specific dates, we show the top 2-grams found in messages containing "mental health" in Fig. 4.2. Similar figures showing the top 1-grams (Fig. S1) and 3-grams (Fig. S2) can be found in Appendix A.1. These co-occurring $n$-grams are shown with their usage rate, rather than rank, so that we can visually see how phrases are being used compared to the others in the same list.

For example, a popular article shared on December 14, 2012 contained the phrase
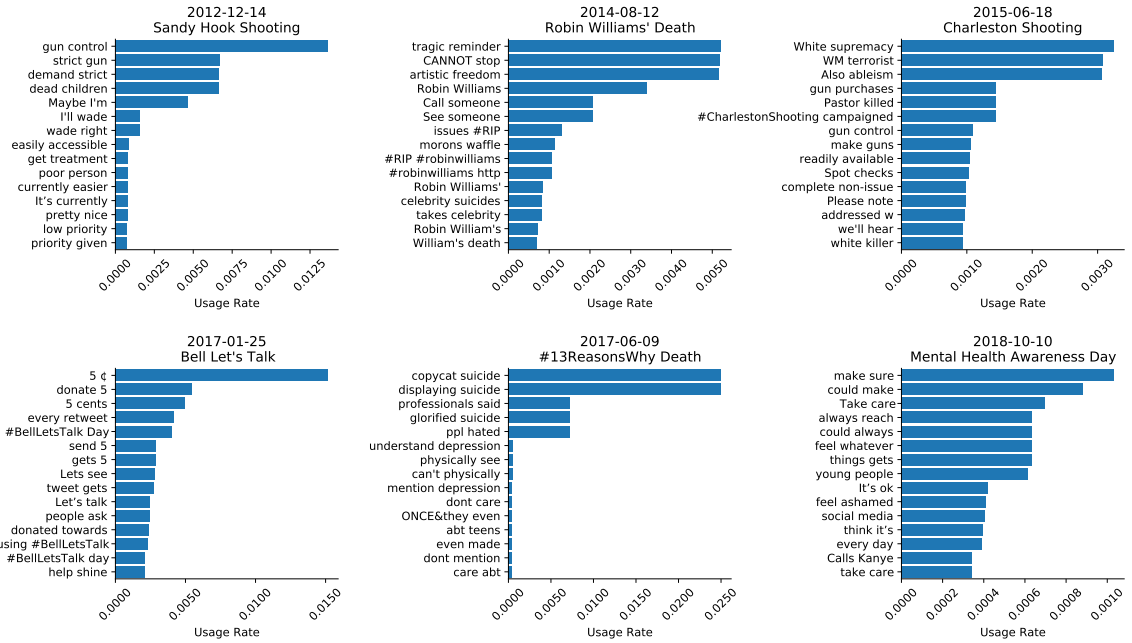
*Figure 4.2:* **Top n-grams used in discussions of mental health.** *Here we show the top fifteen 2-grams that appear in the "mental health" tweet collection for a few outlier dates noted in Fig. 4.1. Each subplot lists the date and its associated event, along with a bar graph of the usage rate. It is worth noting that the bars in each subplots cannot be compared to those of the other subplots, as the range of the x-axes are varied for clarity.*

"It's currently easier for a poor person to get a gun than it is for them to get treatment for mental health issues." which was subsequently quoted by thousands of accounts on Twitter [29]. The resulting phrases seen in this figure provide more insight into what the broader conversation around mental health looks like following these events.

## 4.2 Happiness Word Shifts

To understand the rise and fall of the ambient happiness scores over the timeseries in Fig. 4.1, we can look at the words that most heavily contribute to these shifts [21]. Fig. 4.3 highlights words associated with the same key events shown in Fig. 4.2, using messages from a week before the event as a reference. Words highlighted with a blue bar are ones that have been coded as negative, and words with a yellow bar have been coded as positive.

The darker shades of these two colors represent words that have increased in usage compared to the reference, while lighter shades represent words that have decreased in usage. The left side of these panels shows words that are bringing the average score down, whether with an increase in negative words or a decrease in positive words, and the right side shows words that are raising the score. The average ambient happiness scores for the day of the event and a week before the event are also highlighted at the top of each panel. The 1-grams are also ordered by rank from top to bottom, as shown by the vertical axis.

Looking at Fig. 4.3, we see that mass shooting events have an increase in negative words such as "gun", "guns", and "shocked", and a diminishing use of negative words such as "depression", "disease", and "crisis". The day of the Sandy Hook shooting
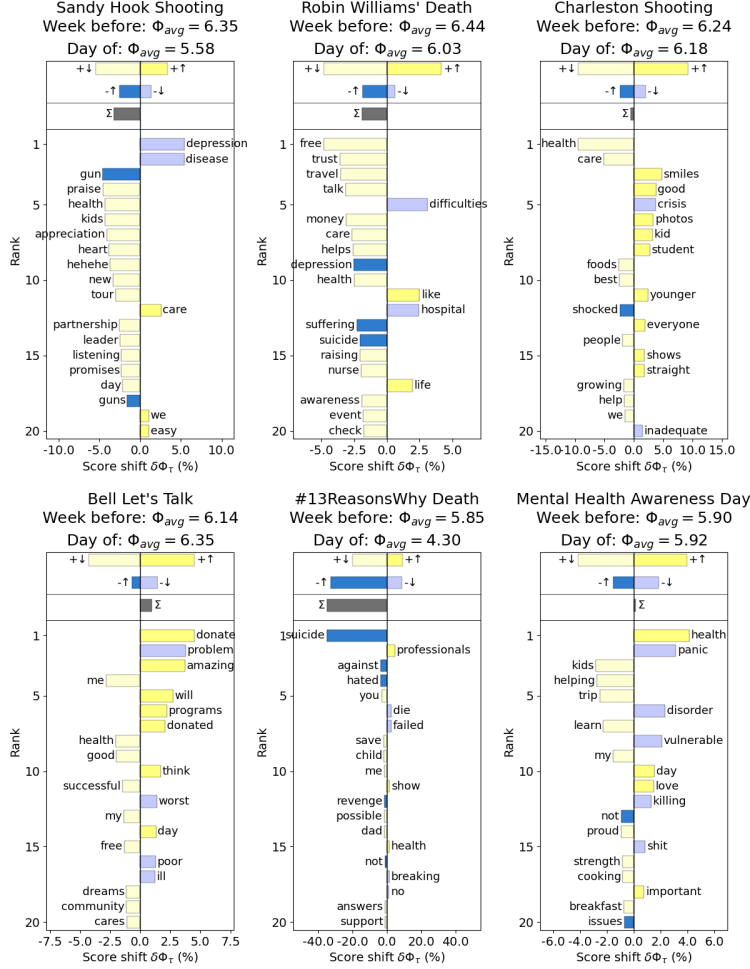
Figure 4.3: **Happiness word shift graphs.** In each of the six panels, we show the twenty 1-grams that contribute most to the shift in ambient happiness on key dates shown in Fig. 4.1, relatively to the prior week. The words shown in blue are ones that have been labeled as relatively negative, and the ones shown in yellow have been labeled as relatively positive [21]. For example, on the day of the Sandy Hook shooting, the relatively negative word "gun" appeared more often in mental health tweets than during the prior week, while the relatively negative words "depression" and "disease" appeared less often. On Bell Let's Talk day, the relatively positive words "donate" and "amazing" appear more often, and the relatively negative words "problem" and "worst" appear less often. The darker shade of these colors tells us where there is an increase in these words, while the lighter shade represents a decrease in usage. The happiness score shift is shown on the horizontal axis, representing how positive or negative the language on these days becomes, and the happiness rank of the 1-gram in this subset is shown on the vertical axis. Average ambient happiness scores for the day of the event, as well as a week before the event, are also noted at the top of each subplot.

18

saw less positive words such as "praise", "appreciation", and "listening", which would usually be seen in the daily mental health content on Twitter.

While the Charleston shooting saw a decrease in words such as "health" and "care", it also saw an increase in positively coded words such as "smiles", "kid", and "student", which likely refer to the shooter in this event. This example highlights the drawbacks of dictionary-based ambient happiness analysis without context of the words being used, as independently positive words can be used to describe a tragic event and vice versa. The middle panels in both rows highlight word shifts following death by suicide tragedies, and include an increase in the words "depression", "suffering", and "suicide", which explain the drops in ambient happiness seen on these days.

The awareness events Bell Let's Talk and Mental Health Awareness Day, which represent the only increases in ambient happiness of the dates shown in Fig. 4.3, both show an increase in quite a few positive words: "donate", "amazing", "programs", "health", "love", and "important". These days also notably see a decrease in strongly negative words, such as "problem", "disorder", "vulnerable", and "killing". These results highlight the shift in language on awareness days, away from phrases with negative connotations and focusing on language relating to community support and aid.

# CHAPTER 5

# NARRATIVE AND SOCIAL AMPLIFICATIONS

## 5.1   Rank-Turbulence Divergence

The increasing appearance of the phrase "mental health" could be due to several factors. We analyze the corpus associated with the topic of "mental health" using the $n$-grams and their relative frequency and rank values for each day, and compare the word usage in this subset to a random sample of messages on Twitter.

To compare differences in language usage, we use rank-turbulence divergence [18]. With this method, we can examine the shift in language between the two samples of tweets. We aggregate $n$-gram counts for phrases found in tweets containing "mental health" over the span of each year, getting annual counts for each of these phrases.

We do the same aggregation for a smaller random subset of Twitter data, aggregating yearly data for a one percent sample of the Decahose API. Fig. 5.1 highlights the results of rank divergence comparing the two subsets of messages across the year 2020.

Each square histogram bin reflects the relative ranks for 3-word phrases in each respective subset. Bins to the right side contain 3-grams with relatively higher rank in the right subset than the left. The bins down the middle of the plot contain words with a similar rank in both subsets. The bands of bins on the bottom edges of these plots represent words that are exclusive to their respective side's dataset.

The color of each bin correlates with the density of words contained in it, and the words appearing on the plot are randomly selected representatives from the bins on the outer edges. The table on the right shows the words that contribute most to the divergence of the two datasets, with small triangles indicating when a word is exclusive to one system.
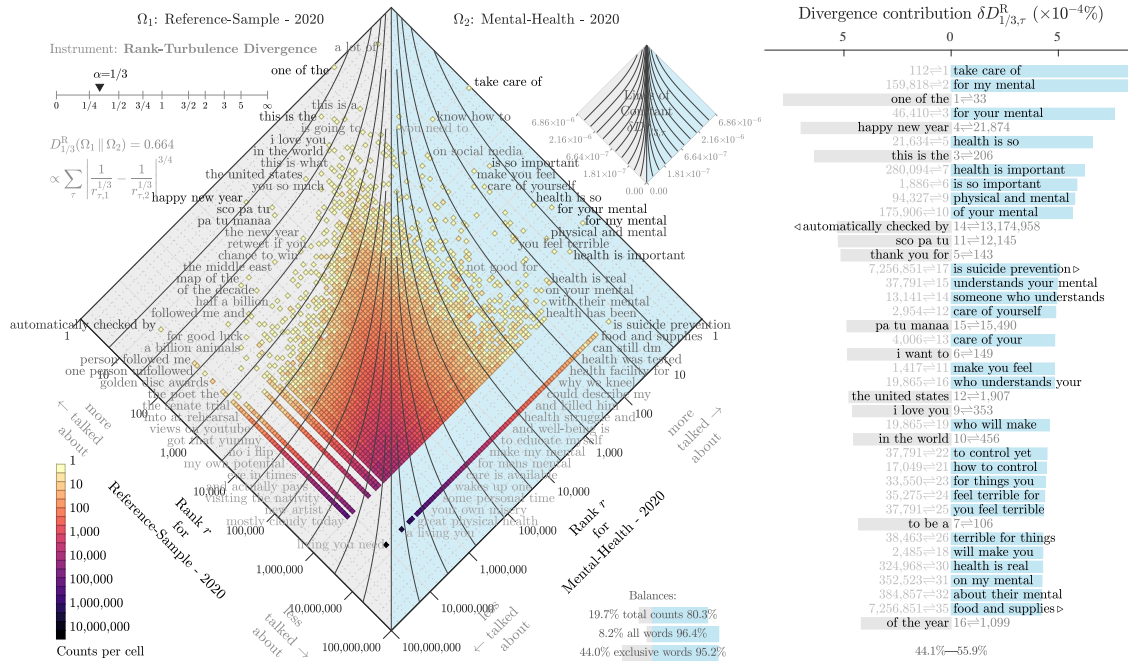
Figure 5.1: **Allotaxonograph using rank-turbulence divergence of 1-grams from tweets in 2020 containing the anchor phrase "mental health", compared to a random sample of tweets in 2020.** *In the central 2D rank-rank histogram panel, phrases appearing on the right have higher rank in the mental health subset than in random tweets, while phrases on the left appeared more frequently in the random sample. The table to the right shows the words that contribute most to the divergence. For example, the phrase "take care of" was the 112th most common 3-gram in random tweets posted during 2020, but it was the most common 3-gram in tweets containing "mental health". Note that when ranking 3-grams from mental health tweets, "* mental health" and "mental health *" phrases were removed for clarity. The balance of the words in these two subsets is also noted in the bottom right corner of the histogram, showing the percentage of total counts, all words, and exclusive words in each set. See Dodds et al. [18] for a detailed description of our allotaxonometric instrument.*

When comparing $n$-grams from these subsets in Fig. 5.1, we see that the mental health dataset, shown on the right side of the figure, includes language related to taking care of your physical and mental health, suicide prevention, men's mental health, social media, and personal time. These topics seem to have become more prominent in the year 2020 with people being at home and isolated during the COVID-19 pandemic, and with more awareness being brought to the relationship between social media and mental health. While we would expect to see pandemic-related phrases show up in 2020, these topics were equally mentioned across both samples, so they do not appear on either side of this histogram.

Studies this year have shown that at the onset of the pandemic, Google searches for terms related to mental health increased initially, followed by a "flattening out" after stay-at-home orders were announced [24]. It has also been recorded that in the time between March and July 2020, average phone screen time doubled to 5 hours per day and rates of depression increased by 90 percent [22]. While these figures cannot tell us everything about how language differs between subsets of conversation, they do provide a sense of the mental health topics individuals discussed in 2020.

## 5.2 CONTAGIOGRAMS

To better understand the dynamics of phrases related to mental health, we explore ways in which these messages are spreading across Twitter. Tweets can be either posted as original content in a new message, or a user can retweet a message that another user has posted.

Organic messages show that users are writing their own content related to a topic,

while retweeted messages show that this topic is being shared and spread to other groups of users; both are important means of contributing to conversation. Both organic messages (OT) and retweeted messages (RT) appear in our dataset and are included in the previous analyses, so it is important to also examine the proportion of messages that fall into these two categories.

Fig. 5.2 shows "contagiogram" plots, as implemented by Alshaabi *et al.* [2], which highlight the relationship between retweeted and organic content for a given $n$-gram on Twitter. The top panel of these plots shows the monthly relative usage of the specified $n$-gram, highlighting usage of organic messages in blue and shared retweets in orange. A shaded area in this top panel represents time periods when the number of retweeted messages surpasses that of organic messages, highlighting social amplification.

The middle panel shows retweet usage of an $n$-gram, relative to the rate of all retweeting behavior across English Twitter, using a heatmap for each day of the week across the timeseries. In this heatmap, darker red shades represent a higher relative rate of retweets for the given $n$-gram compared to a random English $n$-gram on Twitter, and grey shades represent a higher rate of original messages. The bottom panel provides the rank timeseries of the $n$-gram, with a month-scale smoothing of the daily values shown in black. In Fig. 5.2, we look at these contagiogram plots for a collection of key $n$-grams related to the discussion of mental health on Twitter.

Looking at English Twitter overall, the balance of messages was primarily organic until around 2017, when the practice of retweeting messages tipped the balance [3]. Around this same time, retweeted messages reach higher numbers than organic messages for most mental health related $n$-grams, as seen in the top panels of these
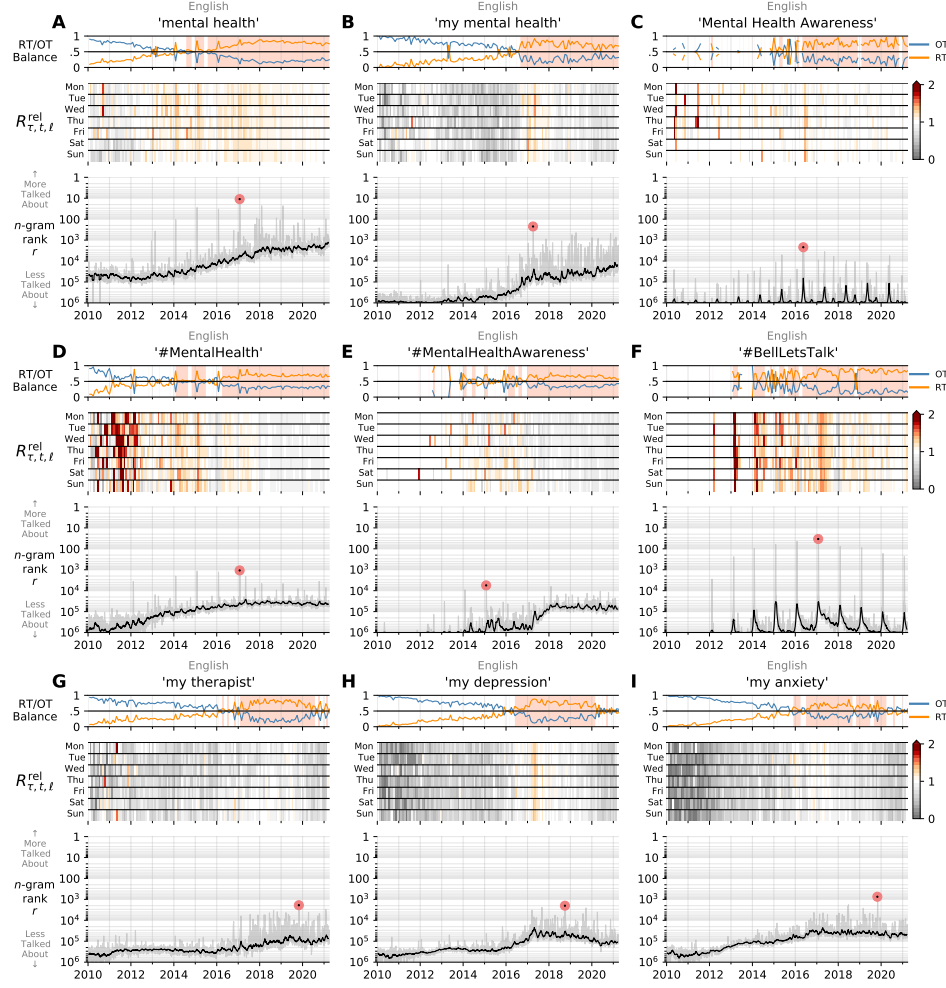
Figure 5.2: **Contagiograms for mental health related** $n$**-grams.** *Phrases and hashtags related to the topic of mental health have grown in volume throughout the time period studied, as reflected by their popularity relative to all tweets. In each subplot, the top panel displays the monthly relative usage of each n-gram, indicating whether they appear organically in new tweets (OT, blue), or in shared retweets (RT, orange). The shaded area highlights time frames when the number of retweeted messages is higher than that of organic messages, suggesting social amplification [2]. The middle panel of each subplot shows the retweet usage of each n-gram relative to the background rate of retweets among all English tweets, with a heatmap for each day of the week. For these heatmaps, the color map is shown to the right, with darker red representing a higher relative rate of retweeting among these messages compared to general messages, and grey representing a higher rate of original messages. The bottom panel shows the basic n-gram rank timeseries, with a month-scale smoothing of the daily values shown in black, and background shading in grey between the minimum and maximum rank of each week. Note that phrase counts only reflect tweets that have been identified as messages written in English as discussed by Alshaabi et al. [3].*

25

subplots.

Examining the heat map panels of these subplots, we observe a larger social amplification effect in hashtags related to mental health, highlighted by the darker red shades across the heatmaps. In recent years, however, these hashtags shift to more organic messages, with the heatmaps becoming more grey after around 2018. The hashtag "#BellLetsTalk" sees the most retweeted behavior of these hashtags, as well as an annual spike on the day of the event, followed by a substantial tail of conversation following this date. On Mental Health Awareness Day (October 10th) of 2018, organic tweets referencing #BellLetsTalk spiked, leading to the inversion of RT/OT in late 2018 that we see in Fig. 5.2F. We also see more original content containing self-disclosure phrases, such as "my therapist" or "my depression", as seen in the third row of $n$-grams which appear to have largely grey shades across the heatmaps. These relationships suggest that users are sharing hashtags in order to spread awareness, and feel comfortable retweeting hashtags posted by others. The public disclosure of private personal anecdotes, which helps to normalize conversation about personal struggles with mental health, is treated differently.

Overall, our results suggest that a larger number of individuals feel comfortable making mental health disclosures publically, but they are amplified relatively less often than other types of mental health messages. Across the subplots, we see a substantial increase in the rank of all phrases/hashtags over time, with annual awareness days resulting in spikes corresponding to their given date each year. These findings offer evidence that understanding mental health conversations have increased substantially over time, reducing the stigma surrounding mental illness.

# Chapter 6

# Concluding Remarks

In this project, we explored the conversation around mental health and its appearance on the social media platform Twitter. Using a collection of phrases, we examined how often the topic of mental health is discussed in tweets, finding that the 2-gram "mental health" has increased in rank by nearly two orders of magnitude since 2012. We calculate the associated ambient happiness for the same time series, finding that happiness is largely effected by key dates and has generally decreased over the past decade.

Compiling a new dataset of $n$-grams found in the subset of tweets mentioning "mental health", we analyzed text associated with this specific term, finding the top $n$-grams related to the topic and their usage rates. We examine the the language in this conversation across years, finding topics that emerged over the past year since the pandemic began.

Comparing usage rates of retweeted content and original content, we find that common "awareness" messages are being amplified on the social media platform, while personal self-disclosing statements are being seen more in organic, originally authored content. These results provide valuable insight into how the discussion of mental health has changed over time, and suggests that more awareness and acceptance has been brought to the topic compared to past years.

We acknowledge that using Twitter as a data source for this research has many limitations, as its user base is not a broadly representative sample of the human population. A study by the Pew Research Center [33] shows that as of June 2019, a only 22 percent of all US adults reported using Twitter, smaller for example than the 69 percent who use Facebook. The age breakdown of users is also skewed, with 38 percent of 18-29 year-olds using Twitter while only 17 percent of 50-64 year-olds use

the site.

While demographics of race are fairly uniform (21 percent of white adults, 24 percent of black adults, and 25 percent of Hispanic adults), the platform is more often used by individuals with a college degree (32 percent) living in an urban area (26 percent) [33]. We also recognize that a portion of Twitter accounts are run by businesses, institutions, and other organized groups, rather than simply individual people. These corporate accounts, such as "@Bell_LetsTalk", would have more of a pattern and agenda to their posted tweets, and there is not currently a way to filter out these messages. Due to these complexities of the Twitter user base, care must be taken when interpreting findings based on tweets.

These limitations could be addressed in future studies by expanding the data sources, e.g., by looking to other available online sites such as Reddit, Instagram, or Facebook, whose user bases differ in some regards. Turning away from social media, one could examine clinical records for cases of diagnosed mental illness, analyzing the language and positivity of physician notes. Rather than looking at simply the messages of this social media platform, this work could be expanded to address the conversation on a network scale, determining how interactions between users impact the discourse.

The work presented here is also limited to the anchor phrase "mental health", and thus could be leaving out conversation related to the topic. To further enrich these findings, future work could expand the existing mental health dataset to include tweets with additional anchor $n$-grams, although a method for determining these anchors would be necessary.

We believe the results presented here provide useful texture regarding the growing

conversation around mental health on Twitter, and evidence that more people are contributing to this conversation on the public social media platform than ever before. Public health campaigns aiming to reduce stigma surrounding mental health can leverage success stories to improve their messaging. As this conversation continues to grow, and perhaps becomes more normalized, it will be useful to examine the language or events that could be contributing to these shifts.

# Bibliography

[1] Everybody hurts, 2021. Available online at https://www.crisistextline.org/everybody-hurts/.

[2] T. Alshaabi, J. L. Adams, M. V. Arnold, J. R. Minot, D. R. Dewhurst, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. *Science advances*, 2021. (In press).

[3] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Science*, 10(15), 2021. URL https://doi.org/10.1140/epjds/s13688-021-00271-0.

[4] T. Althoff, K. Clark, and J. Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476, 2016.

[5] N. Andalibi and J. Buss. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2020.

[6] K. C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, pages 1–9, 2021.

[7] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, and M. De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 79–88, 2019.

[8] M. Conway and D. O'Connor. Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, 9:77–82, 2016.

[9] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, 2014.

[10] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. 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*, pages 1–10, 2015.

[11] P. Corrigan and A. B. Bink. *On the stigma of mental illness.* American Psychological Association, 2005.

[12] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[13] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff. Characterizing and predicting postpartum depression from shared Facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 626–638, 2014.

[14] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2098–2110, 2016.

[15] R. Detels and C. C. Tan. *The scope and concerns of public health.* Oxford University Press, Oxford, UK, 02 2015. ISBN 9780191788567. doi: 10.1093/med/9780199661756.003.0001.

[16] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS One*, 6(12):e26752, 2011.

[17] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth. Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop, 2019. Available online at http://arxiv.org/abs/1910.00149.

[18] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems, 2020. Available online at http://arxiv.org/abs/2002.09770.

[19] P. M. Editors et al. The paradox of mental health: Over-treatment and under-recognition. *PLoS Med*, 10(5):e1001456, 2013.

[20] R. G. Fortgang, S. B. Wang, A. J. Millner, A. Reid-Russell, A. L. Beukenhorst, E. M. Kleiman, K. H. Bentley, K. L. Zuromski, M. Al-Suwaidi, S. A. Bird, et al. Increase in suicidal thinking during COVID-19. *Clinical Psychological Science*, 2021.

[21] R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4, 2021.

[22] O. Giuntella, K. Hyde, S. Saccardo, and S. Sadoff. Lifestyle and mental health disruptions during COVID-19. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[23] S. Gohil, S. Vuik, and A. Darzi. Sentiment analysis of health care tweets: Review of the methods used. *JMIR Public Health and Surveillance*, 4(2):e43, 2018.

[24] N. C. Jacobson, D. Lekkas, G. Price, M. V. Heinz, M. Song, A. J. O'Malley, and P. J. Barr. Flattening the mental health curve: COVID-19 stay-at-home orders are associated with alterations in mental health search behavior in the United States. *JMIR Mental Health*, 7(6):e19347, 2020.

[25] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[26] K. Kindelan and S. Ghebremedhin. California families claim '13 Reasons Why' triggered teens' suicides. *ABC News*, 2017. URL https://abcnews.go.com/US/california-families-claim-13-reasons-triggered-teens-suicides/story?id=48323640.

[27] A. Li, D. Jiao, and T. Zhu. Detecting depression stigma on social media: A linguistic analysis. *Journal of Affective Disorders*, 232:358–362, 2018.

[28] G. McClure. Suicide in children and adolescents in England and Wales 1970–1998. *The British Journal of Psychiatry*, 178(5):469–474, 2001.

[29] S. Mukherjee. It's easier for Americans to access guns than mental health services, 2012. URL https://archive.thinkprogress.org/its-easier-for-americans-to-access-guns-than-mental-health-services-13064

[30] T. Nabity-Grover, C. M. Cheung, and J. B. Thatcher. Inside out and outside in: How the COVID-19 pandemic affects self-disclosure on social media. *International Journal of Information Management*, 55:102188, 2020.

[31] J. A. Naslund, A. Bondre, J. Torous, and K. A. Aschbrenner. Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science*, 5(3):245–257, 2020.

[32] W. H. Organization. The world health report: Mental disorders affect one in four people. 2001.

[33] A. Perrin and M. Anderson. Share of US adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center*, 10, 2019.

[34] M. Price, A. C. Legrand, Z. M. Brier, K. van Stolk-Cooke, K. Peck, P. Dodds, Z. W. Adams, and C. M. Danforth. Doomscrolling during COVID-19: The negative association between daily social and traditional media consumption and mental health symptoms during the COVID-19 pandemic, 2021. Available online at https://psyarxiv.com/s2nfg/.

[35] N. J. Reavley and P. D. Pilkington. Use of Twitter to monitor attitudes toward depression and schizophrenia: An exploratory study. *PeerJ*, 2014.

[36] A. G. Reece and C. M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1):1–12, 2017.

[37] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, 7(1):1–11, 2017.

[38] H. Ritchie and M. Roser. Mental health, 2018. Available online at https://ourworldindata.org/mental-health.

[39] P. Robinson, D. Turk, S. Jilka, and M. Cella. Measuring attitudes towards mental health using social media: Investigating stigma and trivialisation. *Social Psychiatry and Psychiatric Epidemiology*, 54(1):51–58, 2019.

[40] K. Roemmich and N. Andalibi. Data subjects' conceptualizations of and attitudes toward automatic emotion recognition-enabled wellbeing interventions on social media. *Proceedings of The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (PACMHCI'21)*, 2021.

[41] D. Rose, G. Thornicroft, V. Pinfold, and A. Kassam. 250 labels used to stigmatise people with mental illness. *BMC Health Services Research*, 7(1):1–7, 2007.

[42] R. Zhang, N. N. Bazarova, and M. Reddy. Distress disclosure across social media platforms during the COVID-19 pandemic: Untangling the effects of platforms, affordances, and audiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

# Appendix A

# Supplementary Information
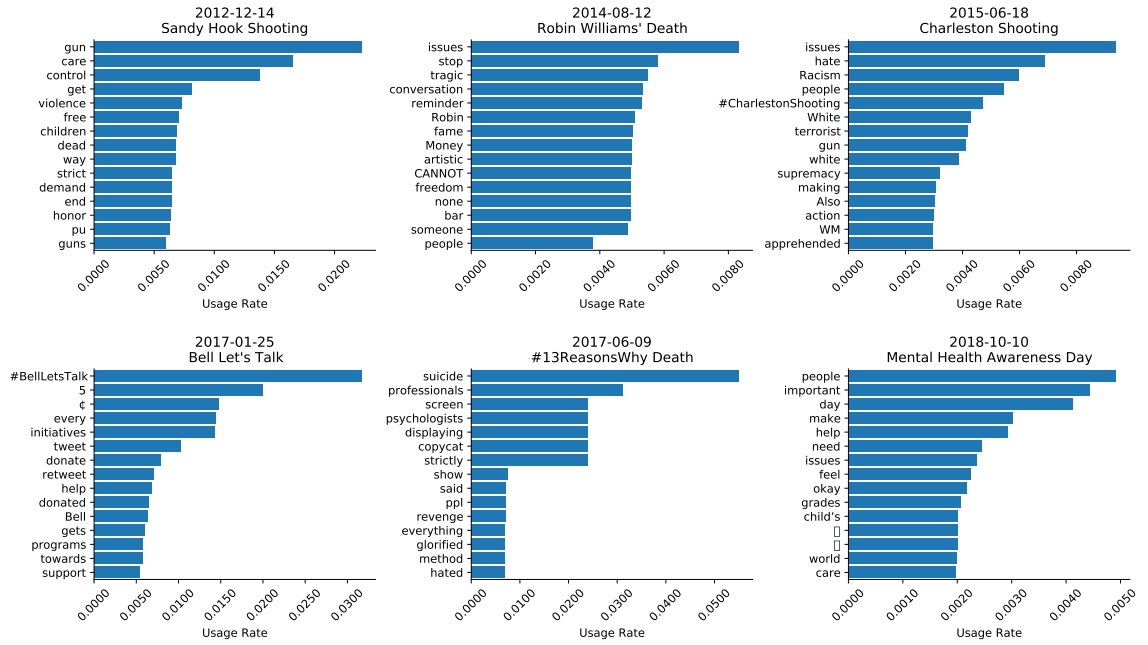
# A.1 Top $n$-grams in Mental Health Collection



*Figure S1:* **Top $n$-grams used in discussions of mental health.** *Here we show the top fifteen 1-grams that appear in the "mental health" tweet collection for a few outlier dates noted in Fig. 4.1. Each subplot lists the date and its associated event, along with a bar graph of the usage rate.*
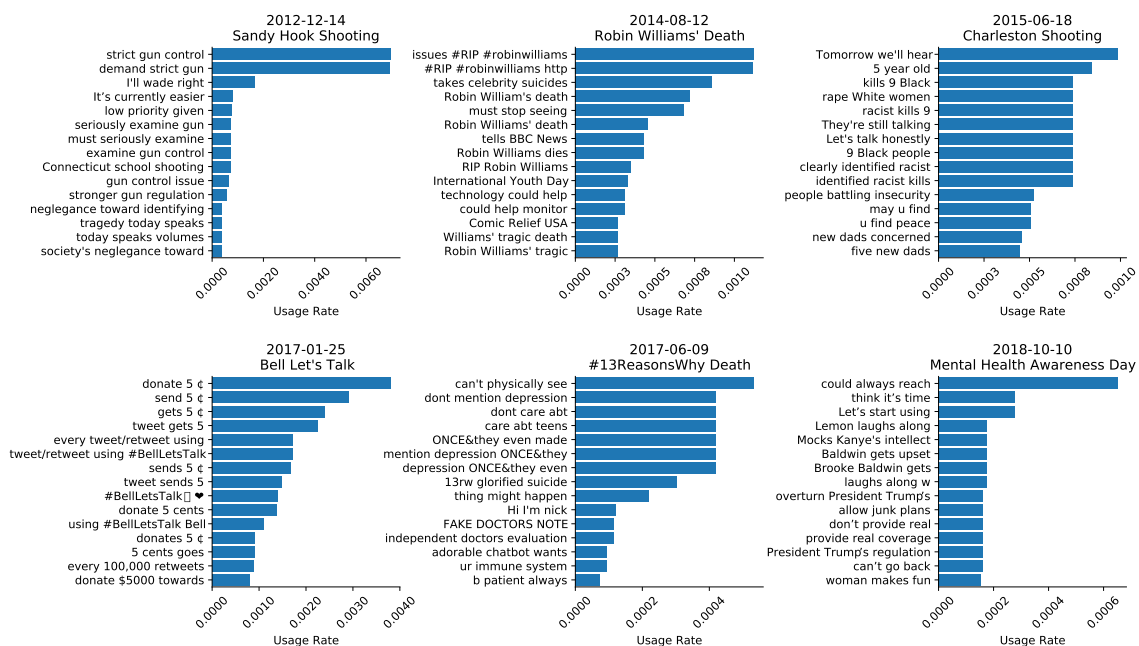
*Figure S2:* **Top** *n***-grams used in discussions of mental health.** *Here we show the top fifteen 3-grams that appear in the "mental health" tweet collection for a few outlier dates noted in Fig. 4.1. Each subplot lists the date and its associated event, along with a bar graph of the usage rate.*