Measuring Mental Health Stigma on Twitter

Anne Marie Stupinski
University of Vermont

Follow this and additional works at: https://scholarworks.uvm.edu/hcoltheses

Recommended Citation
Stupinski, Anne Marie, "Measuring Mental Health Stigma on Twitter" (2020). UVM Honors College Senior Theses. 369.
https://scholarworks.uvm.edu/hcoltheses/369

This Honors College Thesis is brought to you for free and open access by the Undergraduate Theses at ScholarWorks @ UVM. It has been accepted for inclusion in UVM Honors College Senior Theses by an authorized administrator of ScholarWorks @ UVM. For more information, please contact donna.omalley@uvm.edu.
Measuring Mental Health Stigma on Twitter

Anne Marie Stupinski
Christopher M Danforth, Peter Sheridan Dodds, Matthew Price

May 1, 2020
## Contents

1 Abstract 2

2 Background and Motivation 3

3 Methods 6

3.1 Data 6

3.2 n-grams 7

3.2.1 General 7

3.2.2 Mental Health-related 7

3.2.3 Self-disclosures 8

4 Results and Discussion 9

4.1 Negative Stereotyping on Twitter 9

4.2 Growth of Collective Attention to Mental Health 11

4.3 Disclosures of Personal Mental Illness 16

4.4 Dynamics of Social Contagion 20

5 Concluding Remarks 22

6 Acknowledgements 24

A Supplementary Information 28

A.1 Rank Divergence 28

A.2 Self-disclosures 32
1 Abstract

Major depression is a serious health issue afflicting hundreds of millions of people each year, with many going untreated due to the intense stigma surrounding mental illness. In this project, we explore perceptions of mental health on social media, attempting to quantify the level of stigma present on Twitter and track how it has changed in the past decade. To explore trends in the appearance of various words and phrases, we collect roughly 10% of all tweets starting in 2008, process English tweets into 1-, 2-, and 3-grams, and determine their usage frequency and rank. Using these values, we can examine how often the topic of ‘mental health’ is discussed on Twitter, and we find that the phrase has increased in rank by an order of magnitude since 2013. We attempt to disentangle the components of this rise in prevalence, determining how much of the rise is explained by decreased stigma and how much is explained by a convergence in linguistics. We look at messages containing ‘mental health’ posted in 2012 and 2018, as these years are before and after the drastic increase in rank of this phrase, and examine the divergence of the language in both subsets. In further efforts to measure stigma, we compile a list of negative labels commonly used in stigmatizing language and track the rank and frequency throughout the past decade, and we find that many of these labels have decreased in rank in recent years. We also identify statements of self-disclosures of Twitter users and examine how many appear over the years in a subset of tweets specifically about depression. These results all provide valuable insight into how the discussion around mental health has shifted over time.
2 Background and Motivation

It is estimated that nearly 450 million people suffer from mental illness, with 300 million of those people suffering from depression. These numbers put mental illnesses among the leading causes of ill-health and disability worldwide. Rates of mental disorders and deaths by suicide are increasing, especially among young people. However, services for identifying and treating mental illnesses are insufficient, and under-diagnosis is a persistent problem. Many people who would benefit from mental health services decide not to seek or participate in care, as they are either unaware of such services, are unable to afford them, or they wish to avoid the label of mental illness and the judgement surrounding it. In fact, two-thirds of people with a known mental disorder never seek help from a health professional.

The stigma existing around mental health issues is a huge barrier between those who are affected by mental illness and the help they need. Stigma, as a general term, is when someone is seen in a negative way because of a particular attribute, such as skin color, cultural background, a disability, or in this case, a mental illness. Stigma can be broken down into three different components: lack of knowledge about the subject, negative attitudes towards the given attribute, and excluding or avoiding behaviors around people with this attribute. Stigma can have many negative affects on those who are targeted by it. People struggling with mental health issues often experience feelings of shame, hopelessness, and isolation as a result of the stigma society holds towards them. Friends and family members often have a lack of understanding of the struggles they face, and this can lead people to be reluctant to reach out for help or get treatment. Stigma is also a major factor of psychological distress and may affect relationships and educational goals, resulting in fewer opportunities for employment or social interaction. In extreme cases, stigma held against people with mental illness can lead to bullying, physical violence or harassment.

Due to the fact that a major component of stigma is a lack of knowledge, several global campaigns have been started which aim to bring awareness and education to the general public about mental
health issues. World Mental Health Day, celebrated annually on October 10th, was started in 1992 with the initiation of the World Federation for Mental Health. Its goal is global mental health education, awareness, and advocacy in an effort to mitigate stigma faced by those affected by mental illness. Bell Let’s Talk, falling on the last Wednesday of January each year, was started by the Canadian company Bell Telephones in 2011 with similar goals in mind. On this day, Bell Telephones pledges to donate five cents for each tweet, retweet, Snapchat, and text containing their hashtag to mental health organizations, motivating thousands of online discussions around issues of mental health and methods of getting help and support. Despite the large amount of attention these events generate, there has not yet been any analysis on whether awareness campaigns truly accomplish their goal of reducing stigma. As part of the broader effort to measure stigma on social media, this paper will look into the effects these events have on online users in the days and years following.

Many previous researchers have used various social media platforms in order to explore and understand dynamics of mental illness online. A study by De Choudhury [6] uses Twitter activity from users who have been diagnosed with depression along with clinically validated measures in order to predict users who may be at risk of the mental illness. Reece et al. [7] improved upon this study, using tweets posted prior to a user’s diagnosis date to better capture the onset of depression. De Choudhury has also worked on predicting postpartum depression in new mothers, using Facebook activity, linguistic expression in status updates, and demographic survey data [8]. Using consenting Instagram users’ photos, Reece et al. [9] found that there are distinct predictive markers of depression in users’ profiles. Work by Coppersmith et al. [10] classifies online users who suffer from Post Traumatic Stress Disorder by using self-disclosing messages on Twitter. Another study using self-disclosures [11] trained 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. Another study by De Choudhury [12] uses 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 mental health content online.

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. [13] 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. These students were asked: “What sorts of words or phrases might you use to describe someone who experiences mental health problems?” The resulting words were manually grouped by their connotation and five main themes emerged. The majority of the phrases fit into the theme “popular derogatory terms” and included words such as “freak”, “retard”, and “braindead”. Words in this category will be used later in our study in order to track the appearance of stigmatizing labels on Twitter.

Reavley and Pilkington [14] takes 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. 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. This study makes good use of Twitter to measure attitudes toward mental illness, but can be extended to look at tweets over a longer period of time.

While negative stereotypes and attitudes of the general public are decent measures of stigma, the experiences of people with mental illnesses should be most important. When people report fewer hurdles to getting opportunities and less fear of seeking help, there will be true evidence of a lack of stigma [15]. Being willing to publicly discuss personal experiences with mental illness is another good sign of an individual feeling supported and accepted by their community. Several studies, while not directly related to mental health, attempt to explain why individuals choose to post
sensitive personal disclosures on social media accounts. A recent study by Andalibi and Forte [16] explores disclosures of pregnancy loss on Facebook, interviewing women who had experienced this tragedy and shared their story on their personal Facebook page. They find that public disclosures are more likely when a user has previously seen other related disclosures, and are motivated by six main factors: self, audience, network, society, passage of time, and platform. While these factors were observed in this specific context, they could very well apply to the disclosure of other sensitive topics such as mental illness.

Previous research in measuring stigma is fairly limited, and even more so in relation to stigma on social media. The existing literature is mostly qualitative and exploratory, with human researchers manually parsing through messages and assigning them to categories. Due to this hand-coding process, only a small time period of tweets have been analyzed. The goal of this project is to improve upon this area of work with advanced computing power and a data-driven approach, which allows us to examine tweets over a full decade rather than a week. Using messages from Twitter, we measure stigma in three ways: quantifying the use of stereotyping labels, examining the growth of public attention to mental health, and tracking the frequency of self-disclosures.

3 Methods

3.1 Data

Twitter is a valuable source of information of the views people hold on various topics, as tweets are public and the platform is commonly used by both adults and young people. We acknowledge that Twitter may not capture all aspects of mental illness, however, as many people prefer to cope with their pain on their own and likely will not disclose that on a public platform such as this. Nevertheless, it is a great source to sketch a rough portrait of the stigma around mental health among people on this platform. Through Twitter’s Decahose API, we have collected a 10% random sample of all public tweets between September 2008 and the present date. These tweets come in
the format of JSON files, where each tweet object has a user, message, timestamp, and sometimes
geo-metadata, along with many other potential attributes. Zipped files for each day of tweets are
stored on the Vermont Advanced Computing Core (VACC).

3.2 n-grams

3.2.1 General
To explore trends in the appearance of words, we take the English Twitter corpus (as outlined
in [17]) and process tweets 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. For each day, we count the amount of times these n-grams appear in
tweets and determine their usage frequency relative to the appearance of other phrases on Twitter.
Then, we rank n-grams by descending order of count; n-grams with a low rank value assigned to
them are phrases that appear on Twitter very often, while those with a high rank value do not
appear as often. 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 [18]. 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-related
To explore the more specific language used when discussing mental health on Twitter, we compile a
new dataset of n-grams from tweets related to this topic. We first parse all tweets from the decahose
and filter out messages that do not contain the 2-gram “mental health”. With the remaining
messages that do contain this phrase, we create n-grams in the same fashion as previously described,
determining their usage frequency and ranking them by descending order of counts. Summary
statistics of this new dataset compared to the general 1-grams dataset are shown in Table 1. We
also compute the aggregated frequency and rank of these n-grams over each year in the decahose,
rather than just for each individual day. We do this using the existing count values for each day, summing them over each year and ranking them by these counts. With this data, we can analyze text associated with this term and see how it has shifted from year to year.

<table>
<thead>
<tr>
<th></th>
<th>2012-02-08</th>
<th>2014-01-28</th>
<th>2018-01-31</th>
</tr>
</thead>
<tbody>
<tr>
<td>MH</td>
<td>General</td>
<td>MH</td>
<td>General</td>
</tr>
<tr>
<td>Unique 1-grams</td>
<td>3,039</td>
<td>23,332</td>
<td>49,788</td>
</tr>
<tr>
<td>Total 1-grams</td>
<td>30,336</td>
<td>24,493,496</td>
<td>547,043,674</td>
</tr>
<tr>
<td>Total 1-grams no RTs</td>
<td>9,380</td>
<td>227,861,890</td>
<td>266,028</td>
</tr>
</tbody>
</table>

Table 1: Summary statistics of the mental-health n-gram dataset compared to the general Twitter n-gram dataset. Dates shown are Bell Let’s Talk Day from several years, which is the annual peak in conversation regarding mental health. Unique 1-grams make up the set of distinct words found in tweets on these dates. The count of total 1-grams is the sum of the counts of each unique 1-grams, and total 1-grams with no retweets is the sum of the counts of 1-grams in tweets not including any messages that were retweeted. The counts of 1-grams in the mental-health related database is substantially smaller than the count of all 1-grams on the same dates.

### 3.2.3 Self-disclosures

One method of measuring stigma is to identify statements disclosing personal experience with mental illnesses. Using the general n-gram dataset previously described, we can track the appearance of specific phrases over time. A list of common phrases of self-disclosure were determined by starting with a couple of seed n-grams (“my depression”, “my anxiety”) and collecting messages from Mental Health Awareness Day that contained these phrases. A subset of these messages were manually examined to find similar phrases of self-disclosure, such as “my mental health”, “I have depression”, and “my therapist”. With this list of phrases, we collect tweets from Bell Let’s Talk Day each year and count the appearance of these n-grams. We can then track how their usage frequency has changed throughout the decade.
4 Results and Discussion

4.1 Negative Stereotyping on Twitter

Using a list of negative labels commonly seen in stigmatizing language, we track how often such phrases appear on Twitter. This list was compiled in the previously mentioned study by Rose et al. [13], in which 14-year-olds were asked “what sorts of words or phrases might you use to describe someone who experiences mental health problems?” While roughly 50 different words were recorded in their category “popular derogatory terms”, we used a subset of the most common and most directly related phrases in our analysis. Several words on their list, such as “loser” or “crazy” for example, are not directly referencing mental health issues but are not far off from doing so. These words, when appearing in tweets, cannot be assumed to be related to the state of someone’s mental health and therefore are not included in our results.

Other words, such as “retard” or “braindead”, on the other hand, do typically engage in directly stigmatizing language. Using our general $n$-gram dataset, we find the count and rank of these phrases compared to other words on Twitter for each day. With this data, we can determine if the use of these negative labels has increased or decreased over the past decade. The timeseries of six of the most commonly occurring phrases on Twitter are shown in Figure 1. The daily resolution (blue) and weekly rolling average (black) are plotted to get a sense of where the large spike days are as well as the broader behavior over time. Of these six main phrases, “freak” and “retard” are evidently on a steady decline. The phrases “braindead” and “demented” both initially dropped in rank and then increased again around 2015, while the phrases “disturbed” and “loony” remain relatively constant.

It is worth noting that while these phrases are more directly related to mental illnesses than others on Rose’s list, the context of their use on Twitter cannot be assumed. The $n$-gram “disturbed” for example is apparently often used in reference to a band of that name, and perhaps the spike days in the timeseries line up with album releases or other related news. Despite the drawbacks,
these timeseries still provide a general picture of the use of these negative labels in society.

![Graphs showing rank of stigmatizing labels on Twitter from 2009 to 2020.](image)

**Figure 1:** Rank of stigmatizing labels on Twitter from 2009 to 2020. ‘Rank’ is determined by ordering all 1-grams in descending order of counts for each day, and then plotted on an inverted logarithmic axis. Daily resolution is shown in blue and a weekly rolling average is shown in black. The 1-grams “freak” and “retard” appear to have substantially decreased in use. Other words such as “demented”, “braindead”, and “loony”, however, appear to dip and then increase again around 2015.
4.2 Growth of Collective Attention to Mental Health

Public awareness and education of an issue is an important step in reducing negative attitudes, as a major component of stigma is a lack of knowledge. In order to understand the general public’s level of awareness of mental health issues, we examine the frequency at which people on Twitter have discussions about the topic of mental health. Using our Twitter n-gram data, we construct a rank timeseries of the 2-gram “mental health” on a logarithmic axis, which can be seen in Figure 2. We find that this 2-gram has increased in rank by an entire order of magnitude since 2013. This substantial increase is evidence that the conversation around mental health is happening more frequently than ever before.

![Figure 2: 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 then is plotted on an inverted axis. The logarithmic plot is evidence that since 2013, the phrase has increased in rank by an order of magnitude. This represents the increased discussion of mental health on Twitter. Large spike days are annotated with the associated event contributing to the increase in rank.](image)

Examining the daily behavior of this timeseries, we see that several dates emerge where the rank largely deviates from the baseline. The timeseries in Figure 2 is annotated with the events associated with these large jumps in rank. After researching the events occurring on these dates, we find that the awareness events Bell Let’s Talk and Mental Health Awareness Day (MHAD) are contributing to the large, annual spikes beginning in 2013. The 2-gram “mental health” reaches its highest rank on record on Bell Let’s Talk day in 2017, peaking with a rank of 18 compared to all other 2-grams on Twitter that day. Other spikes in rank occurred on dates when celebrities such
as Robin Williams died by suicide, or on dates when mass shootings occur.

![Figure 3: Rank timeseries of the 2-gram ‘mental illness’ over the past decade on a logarithmic axis. ‘Rank’ is determined by ordering 2-grams in descending order of counts for each day, and then is plotted on an inverted axis. Similarly to “mental health”, this logarithmic plot shows that since 2013, the phrase “mental illness” has increased in rank over time. Large spike days are annotated with the associated event contributing to the increase in rank.](image)

Figure 3: Rank timeseries of the 2-gram ‘mental illness’ over the past decade on a logarithmic axis. ‘Rank’ is determined by ordering 2-grams in descending order of counts for each day, and then is plotted on an inverted axis. Similarly to “mental health”, this logarithmic plot shows that since 2013, the phrase “mental illness” has increased in rank over time. Large spike days are annotated with the associated event contributing to the increase in rank.

![Figure 4: Timeseries comparing the rank of the 2-grams “mental health” to that of the 2-gram “mental illness”. The phrases show similar behavior, with the general trend as well as dates of spike days almost perfectly aligning.](image)

Figure 4: Timeseries comparing the rank of the 2-grams “mental health” to that of the 2-gram “mental illness”. The phrases show similar behavior, with the general trend as well as dates of spike days almost perfectly aligning.

We can also construct the same annotated timeseries of rank using the 2-gram “mental illness”, which is shown in Figure 3. This timeseries shows incredibly similar behavior to that of “mental health”, and a side-by-side comparison of the two is plotted in Figure 4. The trends of the two phrases very closely follow each other, with “mental health” in tending to be just higher than “mental illness” in rank. The dates that they drastically jump up in rank almost always match, with the exception of the Charleston, Dayton, and El Paso shootings where “mental illness” is more prevalent and higher in rank. While the two phrases are so clearly intertwined in conversations, it
is notable that they take the lead in rank on different types of events. The 2-gram “mental health” is higher in rank on awareness days, where the topic is generally discussed in a more positive light. The 2-gram “mental illness”, on the other hand, only overtakes “mental health” on events of mass shootings, where the discussion tends to revolve around the shooter’s negative mental state.

While the increase in rank of these phrases is sizable, it of course could be due to several other factors. We attempt to disentangle the components of this rise in prevalence, separating it into contributions of decreased stigma and a convergence in linguistics. In order to analyze any potential shift in linguistics over time, we compile text associated with the term “mental health”. We create a new dataset of n-grams found in the subset of tweets mentioning this phrase. After collecting these n-grams and calculating their relative frequencies and ranks for each day, we can compare the word usage in different dates.

Using a method of calculating rank turbulence divergence [19], we examine the shift in language between several specific dates and years. First, we compare 1-grams contained in messages from the awareness events in 2012 and 2018. The results from comparing tweets on Bell Let’s Talk Day from these years are shown in Figure 5. We then aggregate mental health n-gram counts over the span of each year, getting annual counts for each of these phrases. We look at messages containing “mental health” posted on 2012 and 2018 and examine the divergence of the language in these subsets, which is shown in Figure 6. These figures show data from subsets of tweets where retweets were included. Figures S1–S4 in Appendix A.1 show similar plots for subsets where retweets were not included, as well as for the years 2014 versus 2018 and for Mental Health Awareness Day.

These rank divergence figures highlight the shift in language between two subsets of text. Each square on the plot contains words that fall on that position after the rank divergence calculation, where boxes appearing on the right side of the figure contain n-grams that increased in rank compared to their appearance in the dataset on the left side, and vice versa. The boxes down the middle of the plot contain words that remained relatively stable in rank between the two datasets. The bands of squares on the bottom edges of these plots represent words that are exclusive to their
Figure 5: Rank divergence of 1-grams from tweets on Bell Let’s Talk Day containing the anchor phrase “mental health” on the years 2012 and 2018. Words appearing on the right increased in rank in 2018, while words on the left decreased in rank in 2018 and appeared more frequently in 2012. The table to the right shows the words that are most contributing to the divergence seen.

In the comparison of different years of Bell Let’s Talk Day, we see the appearance of their own hashtag, as well as the hashtag #EndTheStigma, on the 2018 side. Celebrity Twitter accounts also show up, such as Ellen Degeneres (@TheEllenShow). When comparing aggregated n-grams from the years 2012 and 2018, we find that topics such as #BellLetsTalk and Donald Trump appear more often in 2018. The fact that #BellLetsTalk shows up in the general yearly dataset is evidence respective side’s dataset. The color of each square correlates with the density of words contained in it, and the words appearing on the plot are randomly selected from the squares on the outer edges. The table on the right shows the words that are most contributing to the divergence of the two datasets.
Figure 6: Rank divergence of 1-grams from tweets containing “mental health” between the years 2012 and 2018. Words appearing on the right increased in rank in 2018, while words on the left decreased in rank in 2018 and appeared more frequently in 2012. The table to the right shows the words that are most contributing to the divergence seen.

that this event has substantially increased in popularity and is now driving a large portion of the conversation related to mental health on Twitter. Donald Trump’s appearance is likely linked to his increased presence on Twitter in recent years, as well as his tendency to comment on the mental state of mass shooters. On the other hand, topics such as gun control appear more often in 2012. This is almost certainly due to the fact that the Sandy Hook school shooting occurred in 2012, which in turn sparked a still existing debate over gun control in the US. While these figures cannot tell us everything about how language has changed throughout the years, they do provide a sense of the topics that emerge from each dataset and what users of Twitter were talking about during each of these days or years.
4.3 Disclosures of Personal Mental Illness

One sign of decreasing stigma is an increase in people’s willingness to publicly discuss personal experiences with mental illness. Using our self-disclosure n-grams, which were described in Section 3.2.3, we track the frequency of personal accounts of mental health on Twitter. Common phrases of self-disclosure were determined by looking at tweets from Mental Health Awareness Day, and include phrases such as “my depression”, “my anxiety”, and “my mental health”. Looking at tweets on Bell Let’s Talk Day each year, we count the appearance of these phrases both in tweets that contain the hashtag #BellLetsTalk and in tweets from general Twitter on this day. We keep track of the raw count of these phrases as well as this measure normalized by the total number of messages in this subset, as to account for fluctuating rates of Twitter usage throughout the years.

![Normalized count of self-disclosure phrases - logarithmic scale](image)

**Figure 7:** Normalized counts of phrases of self-disclosure appearing in tweets with the hashtag #BellLetsTalk on this day each year. Normalized values are calculated using the total number of tweets identified with this hashtag on that day. Counts are plotted on a logarithmic scale.
Figure 8: Normalized counts of phrases of self-disclosure appearing in any tweet on Bell Let’s Talk Day each year. Normalized values are calculated using the total number of tweets identified on that day. Counts are plotted on a logarithmic scale.

Figure 7 shows the counts of self-disclosure phrases appearing in tweets with the hashtag #BellLetsTalk on Bell Let’s Talk Day for each year, normalized by the total number of tweets containing the hashtag on that day. Figure 8 shows the counts of self-disclosure phrases appearing in any tweet on Bell Let’s Talk Day for each year, normalized by the total number of tweets on that day. Figures S5–S10 in Appendix A.2 show plots of the raw counts of these phrases and plots where the y-axis is on a linear scale.

Figures 7 and 8 show similar behavior, providing further evidence that #BellLetsTalk really does take over Twitter on these dates. Several n-grams, especially “my therapist”, spike in 2017, which was the day that the 2-gram “mental health” reached its peak rank of 18 compared to all of Twitter. Other phrases, such as “my experience”, “my depression”, and “my mental health” rise steadily throughout the years.
Figure 9: Rank of 2-gram self-disclosure phrases on Twitter from 2009 to 2020. ‘Rank’ is determined by ordering all 2-grams in descending order of counts for each day, and then plotted on an inverted logarithmic axis. Daily resolution is shown in blue and a weekly rolling average is shown in black.

While evidence of these phrases appearing more often on these days is exciting, it doesn’t say much about whether users feel a lack of stigma on every other day they’re using Twitter. However,
Figure 10: Rank of 3-gram mental health related phrases on Twitter from 2009 to 2020 - “my mental health” being a self-disclosing phrase and “mental health issues” being used as a baseline 3-gram for comparison. ‘Rank’ is determined by ordering all 2-grams in descending order of counts for each day, and then plotted on an inverted logarithmic axis. Daily resolution is shown in blue and a weekly rolling average is shown in black.

many of these phrases cannot be assumed to be related to mental health outside the context of Bell Lets Talk Day. The phrase “my experience”, for example, could be in reference to almost anything on the other 364 days of the year. In order to get patterns of self-disclosure throughout the past decade at a daily resolution, we choose the phrases from this list that are directly related (“my depression”, “my anxiety”, “my mental health”, “my therapist”) and look for them in our general n-grams database.
When looking at rank timeseries for these \( n \)-grams, it is worth mentioning that 2-grams are only ranked based on the counts of other 2-grams, and 3-grams are only ranked based on the counts of other 3-grams. Therefore, it is misleading to compare the ranks of a 2-gram and a 3-gram, as they are not in the same dataset. To avoid the tendency to compare them this way, we show the plots of the 2-grams together in Figure 9. We also provide another 3-gram, “mental health issues”, to be able to compare our 3-gram self-disclosure phrase to some baseline. These 3-grams are shown in Figure 10. Looking at the 2-grams, it is clear that the phrases “my depression” and “my anxiety” have increased over time. The timeseries of the phrase “my therapist” also becomes more turbulent in recent years, with spike events happening much more frequently. Both of the 3-grams clearly increase in rank, with “my mental health” increasing quite drastically around 2017. These timeseries could contribute to the rise in the rank of the general 2-gram “mental health”. From this process, we have found initial evidence that online users are discussing their personal struggles with mental illness more now than 10 years ago.

### 4.4 Dynamics of Social Contagion

In order to better understand the dynamics at play behind each of the phrases examined in this paper, we explore the ways in which these messages are spreading across Twitter. Tweets can be either posted as new, original content, or a user can retweet a message that another user has posted. Both these organic messages as well as retweeted messages 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. 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.

Looking into the negative labels from Section 4.1, the \( n \)-grams from Section 4.2, and the self-disclosures from Section 4.3, we find there are some interesting dynamics between the number of retweeted and organic messages containing these phrases. Figure 11 shows the timeseries of the
Figure 11: Contagiograms. Timeseries for organic messages (blue), retweeted messages (orange), and total messages (black). Phrases shown are n-grams of relevance to previous sections of this document, and represent either stigmatizing labels or self-disclosures of mental illness. The areas shaded in red highlight occurrences where the number of retweeted messages is higher than that of organic messages. Counts of these phrases only account for tweets that have been identified as messages written in English as discussed by Ref. [17].
rank of each of these $n$-grams on Twitter, along with the corresponding timeseries of organic and retweeted counts for each phrase. For most of Twitter, messages are mostly organic until around 2016 when the practice of retweeting begins to take over [17]. This is about the same time that retweeted messages reach higher numbers than organic messages containing these phrases as well, with the exception of a few. The stigmatizing phrases “retard” remains almost entirely in messages that are original content posted by users. Appearances of the phrase “braindead” in retweeted content and original content converges to the same level over time, but counts of retweeted messages almost never surpass counts of organic messages. This relationship could perhaps be due to other users not wanting to share messages containing these negative labels.

5 Concluding Remarks

In this project, we explored the stigma around mental health and its appearance on the social media platform Twitter. Using our $n$-grams collection, 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 an order of magnitude since 2013. Compiling a new dataset of $n$-grams found in the subset of tweets mentioning “mental health”, we analyzed text associated with this specific term. We examined the divergence of this specific language in different years and found that popular topics related to mental health have changed over the years.

In efforts to measure stigma, we compiled a list of negative labels commonly used in stigmatizing language and tracked their rank and frequency over the past decade. Doing this, we found that many of these labels have decreased in rank in recent years. We also identified statements of self-disclosure by users on Twitter, finding evidence that people with mental illnesses are discussing their experiences more now than they were 10 years ago. These results provide valuable insight into how the discussion around mental health has shifted over time, strengthening evidence that stigma around mental illnesses is on the decline.
We acknowledge that using Twitter for this research has many limitations, as its user base is not a broad enough sample of the human population. A study by the Pew Research Center [20] shows that as of June 2019, a only 22 percent of all US adults reported using Twitter, which is quite low when compared to the 69 percent who use Facebook. The age breakdown of users is also very 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 consistent (21 percent of white adults, 24 percent of black adults, and 25 percent of Hispanic adults), the platform is mostly used by individuals with a college degree (32 percent) living in an urban area (26 percent) [20]. Due to this limited user base, Twitter cannot be regarded as a full representation of the human experience.

Several other limitations of this work exist as well. When examining timeseries of n-grams, it is difficult to know the context that these words are being used in. Therefore, it cannot be assumed that these phrases are directly related to mental health. When looking at the stereotyping labels, we cannot assume that these phrases actively engage in stigmatizing language. In future works, methods to find words and phrases adjacent to a set of seed words should be explored in order to account for this missing context.

In regards to examining the prevalence of stigmatizing labels on Twitter, we also acknowledge that while Rose’s study [13] provides a convenient list of keywords, these phrases may not reflect the type of language that is used to stigmatize mental illness on social media. As the study was done using a class of 14-year-olds, it is also likely that this language does not hold for other age groups or people who live in different areas. However, as the user base of Twitter is strongly skewed towards younger ages, we believe these phrases to be at least somewhat representative of the language that could be used in negatively stereotyping ways. In order to curate a more valid list of labels to track, future studies could use surveys, perhaps on Mechanical Turk, to ask adults suffering from mental health problems what types of negative language they hear and are most sensitive to. These surveys could also ask what language they would prefer the general public to use when referring to people with mental illnesses, so that the rise of these proper phrases with more positive connotations could
be tracked as well.

Due to these drawbacks of Twitter demographics, usage rates, and word context, it would have been valuable to perform similar language analyses on text from individuals diagnosed with a mental illness. With more time, we could have recruited people diagnosed with depression or anxiety and examined their language on Twitter, comparing it to that of the general public and looking at how it has changed over time. Future work in this area could be incredibly useful in recognizing traits of a stigmatizing societal environment. By being able to recognize these patterns, we can bring more awareness to these issues and hopefully lessen the burden placed on those who are negatively affected by stigma.

6 Acknowledgements

I am extremely grateful to my advisor, Chris Danforth, for his support and motivation throughout this project, as well as for inspiring my love of this field at the beginning of my undergraduate career. I would like to thank my advisor and committee member Peter Dodds for his wisdom, insightful feedback on this work, and assistance creating beautiful visualizations. I am also grateful to my committee chair Matt Price for taking the time to be involved in this project, as well as for the inspiring my involvement in psychological science research. Thank you to every member of the Computational Story Lab for the feedback and support, and especially to Thayer Alshaabi for technical assistance throughout this project and for holding me to the standards of a graduate student. To Vanessa Myhaver and Lincoln Pierce, thank you for the company, support, and encouragement throughout the past four years. And to my family, thank you for always believing in me and supporting me in my academic endeavors. This work was also funded in part by the Massachusetts Mutual Life Insurance Company and the University of Vermont Department of Mathematics and Statistics, and would not have been possible without the computing resources provided by the Vermont Advanced Computing Core.
References


[18] Peter Sheridan Dodds, Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Andrew J. Reagan, and Christopher 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.


A Supplementary Information

A.1 Rank Divergence

Figure S1: Rank divergence of tweets from Mental Health Awareness Day in 2012 versus in 2018.
Figure S2: Rank divergence of all tweets from 2012 versus 2018. No retweets are included in these datasets.
Figure S3: Rank divergence of all tweets from 2014 versus 2018. Retweets are included in these datasets.
Figure S4: Rank divergence of all tweets from 2014 versus 2018. No retweets are included in these datasets.
A.2 Self-disclosures

Figure S5: Raw counts of phrases of self-disclosure appearing in tweets with the hashtag #BellLetsTalk on this day each year. Counts are plotted with a linear y-axis.
Figure S6: Raw counts of phrases of self-disclosure appearing in tweets with the hashtag #BellLetsTalk on this day each year. Counts are plotted with a logarithmic y-axis.
**Figure S7:** Normalized counts of phrases of self-disclosure appearing in tweets with the hashtag #BellLetsTalk on this day each year. Counts are plotted with a linear y-axis.

**Figure S8:** Raw counts of phrases of self-disclosure appearing in any tweet on Bell Let’s Talk Day each year. Counts are plotted with a linear y-axis.
Figure S9: Raw counts of phrases of self-disclosure appearing in any tweet on Bell Let’s Talk Day each year. Counts are plotted with a logarithmic y-axis.

Figure S10: Normalized counts of phrases of self-disclosure appearing in any tweet on Bell Let’s Talk Day each year. Counts are plotted with a linear y-axis.