International Journal of Population Data Science

IJPDS International Journal of Population Data Science

Swansea University Prifysgol **Abertawe**

Quantifying depression-related language on social media during the COVID-19 pandemic

Brent D. Davis[1,*], Dawn Estes McKnight[2], Daniela Teodorescu[2], Anabel Quan-Haase[3,4], Rumi Chunara[5,6], Alona Fyshe[2,7], and Daniel J. Lizotte[1,8]

Abstract

Introduction

The COVID-19 pandemic had clear impacts on mental health. Social media presents an opportunity for assessing mental health at the population level.

Objectives

1) Identify and describe language used on social media that is associated with discourse about depression. 2) Describe the associations between identified language and COVID-19 incidence over time across several geographies.

Methods

We create a word embedding based on the posts in Reddit's /r/Depression and use this word embedding to train representations of active authors. We contrast these authors against a control group and extract keywords that capture differences between the two groups. We filter these keywords for face validity and to match character limits of an information retrieval system, Elasticsearch. We retrieve all geo-tagged posts on Twitter from April 2019 to June 2021 from Seattle, Sydney, Mumbai, and Toronto. The tweets are scored with BM25 using the keywords. We call this score rDD. We compare changes in average score over time with case counts from the pandemic's beginning through June 2021.

Results

We observe a pattern in rDD across all cities analyzed: There is an increase in rDD near the start of the pandemic which levels off over time. However, in Mumbai we also see an increase aligned with a second wave of cases.

Conclusions

Our results are concordant with other studies which indicate that the impact of the pandemic on mental health was highest initially and was followed by recovery, largely unchanged by subsequent waves. However, in the Mumbai data we observed a substantial rise in rDD with a large second wave. Our results indicate possible un-captured heterogeneity across geographies, and point to a need for a better understanding of this differential impact on mental health.

Keywords

machine learning; information retrieval; social media; Twitter; COVID-19; depression; public health surveillance

*Corresponding Author:
Email Address: bdavis56@uwo.ca (Brent D. Davis)

Introduction

The COVID-19 pandemic has had a significant impact on population mental health. For example, a meta-analysis of community-based studies conducted during the early stages of the pandemic indicates there was an increase in the prevalence of depression in various countries from January 2020 to May 2020 [1]. The reasons for this increase are complex; COVID-19 changed day-to-day life significantly for many people, and the individual burden of public health restrictions is variable. For example, highly individual factors such as living situation, income, and pre-existing mental health conditions all interact with pandemic-unique stressors to constitute mental health burdens [2]. A study by Daly & Robinson found that psychological distress levels began to decrease following a peak that occurred near the start of the pandemic, and were generally in decline by July 2020 [3]. As public health efforts continue, being able to describe and understand the population mental health burden imposed by the pandemic is an important part of generating an effective response.

Social media can provide a useful source of freely available data for population health monitoring as an adjunct to survey-based methodology. Advantages include its timeliness, lack of recall bias (though other biases may be present), and the fact that it can be retrospectively assessed for its relationships to different events [4]. Furthermore, social media data permit rapid analysis across a variety of geographic areas that may be challenging for survey-based approaches. Such data have been used to inform a variety of public health domains, including mental health and mental illness, disease surveillance, rapid knowledge dissemination, quantifying community supports, and characterizing misinformation in health communication [5–9]. Using sentiment analysis (a natural language processing technique), social media data have also been used with large social media datasets to quantify population mental health during the pandemic [10].

However, sentiment analysis can give an incomplete picture of population mental health for three main reasons. First, sentiment and mental health are different constructs. Many sentiments expressed on social media have nothing to do with the author's mental health, and sentiment analysis techniques are not designed to distinguish between negative sentiment associated with poor mental health and negative sentiment not associated with mental health status [11]. Second, even if poor mental health is evidenced by negative sentiment, many sentiment analysis methods are based on fixed vocabularies and rule-based annotations that are subject to degrading applicability as language evolves over time [12]. This can be problematic, as language used to indicate poor mental health can vary widely from individual to individual and across cultures [13]. Due to nuances in online discourse about mental health, techniques that rely on closed-vocabulary approaches with fixed keyword sets may miss mental health indicators, particularly ones expressed using slang or memetics. Third, language on social media associated with poorer mental health *at the population level* may not express negative sentiment at all—for example, an increase in positive or supportive posts from "helpers" may be a response to a perceived increase in poor mental health at the population level. Sentiment analysis techniques are not currently designed to detect language that is associated with poor mental health at the population level.

Aims

Our overall aim is to present a framework for describing population mental health using social media data that uses a transparent, open-vocabulary approach based on discourse in online communities. To illustrate its use, we present a case study that describes keywords associated with online discussion about depression as well as its rise and fall in different geographic regions over the first year of the COVID-19 pandemic. The aims of our case study are as follows: (I) identify and describe language used on social media that is associated with discourse about depression, and (II) describe the associations between the identified language and COVID-19 incidence over time across several geographies. To assess the framework's suitability for supporting public health practice, we compare to published results not relying on social media such as that of Aknin et al. and Shields et al. [14, 15]. Our goal is for the framework to be understandable and for the resulting keywords to be easy to re-purpose so that other practitioners can perform related analyses on other topics and geographies of interest.

Methods

Our framework follows three basic steps. First, we identify a set of authors (social media users) online whose discourse we want to characterize, and another "control" set of authors who do not participate in that discourse. Second, we use a combination of representation learning and supervised learning to identify the keywords that are most strongly associated with the discourse of interest. Representation learning is the creation of numerical approximations of symbols, concepts, phenomena, or structures; supervised learning refers to algorithms which learn from labelled data [16]. Third, we curate and use the resulting keywords to quantify the amount of discourse of interest we see in other settings, potentially where directly identifying users who participate in the discourse would be impossible. This quantification can include investigating geographic differences, time trends, and associations with other phenomena such as prominent events in the world.

In the following, we detail the methods used to apply our framework to social media discourse about depression. We use Reddit to characterize discourse about depression, and we use Twitter to investigate changes in discourse over time across different geographic regions. We compare the changes in social media discourse we identify as being related to depression with COVID-19 incidence.

Datasets and data providers

Following our framework, the first step is to identify authors who participate in our discourse of interest and authors who do not. To do this, we used Reddit, one of the largest anonymous social media sites in the world [17]. Reddit is composed of sub-communities that discuss different topics; these are called "subreddits." Subreddit names are prefixed with "/r/"; examples include /r/Depression and /r/PublicHealth. Conversations within subreddits are composed of "posts," which initiate new specific topics of conversation, and "comments," which are made in reply to a post. We use participation (or lack thereof) in the /r/Depression subreddit

through posting and commenting to identify authors who participate in online discourse about depression, and authors who do not.

For the second step of analyzing the discourse and producing keywords, we retrieved data from the social media data repository PushShift, which archives and makes available all historical data from Reddit, including comments and submissions with information about when each was posted, which subreddit it was posted to, and the username associated with the author [18]. Using PushShift, we retrieved every post and comment made in /r/Depression between November and December 2019. From this, we extracted 81,118 authors with 3,718,640 posts that were active in the subreddit over that time period. For each of these authors, we extracted their complete posting history on Reddit for the time period. To build a contrasting set of authors who do not participate in discourse about depression, we collected data from non-related subreddits by retrieving the posts of authors who posted in /r/aww, /r/CasualConversation, /r/TotallyNotRobots and /r/AskReddit, excluding any authors who also participated in /r/Depression. These posts are aggregated by author and restricted to these four subreddits. This control set had a total of 143,737 authors with 4,829,822 posts.

For the third step of the framework, we used data from the Twitter microblogging site [19]. Geotagged data were retrieved from the full-archive endpoint of the official Twitter API, which our team was granted access to through Twitter's research program. We retrieved all non-reply, non-retweet posts from January 2020 to April 2021 from the following regions: a 25-mile radius about Mumbai, India; Sydney, Australia; and Seattle, USA; and a 10-mile radius about Toronto, Canada. These locations were chosen because they have many English-language tweets, they experienced different timings and intensities of COVID-19 waves, and there were different public health measures implemented at different times. The choice of using 25-mile radii was due to collection limits imposed by Twitter's API, while a smaller 10-mile radius was selected for Toronto due to the extreme density of tweets in the area. We retrieved confirmed COVID-19 case count information from the COVID-19 Data Repository, a resource from the CSSE at Johns Hopkins University [20, 21] for Seattle, Sydney, and Mumbai (respectively, King County data, New South Wales data, and Maharashtra data). Toronto data were retrieved from the Toronto Public Health Unit [22].

Data pre-processing

Reddit data were cleaned of all special characters, emoticons, and numbers. To cut down on the number of automated accounts (bots) included in the Reddit data set, authors who wrote more than 1500 posts over the two-month period were removed. We also discarded any authors with fewer than 50 words, as well as any with fewer than 10 unique posts; this limit is based around the ability of the algorithm to calculate approximations for these representations, with any less textual information causing problems computing the representation. Further, any accounts with names that contained the suffix "-bot" were removed. Twitter data were filtered for bots by removing any accounts which had made more than 23 tweets over a one-day period (to remove hourly bots) or over 100 geotagged posts overall (to limit individual posters' influence

on data trends). Further, the tweets were filtered to only contain English tweets that were geotagged to be within each area of interest.

Vocabulary extraction methodology

To identify keywords associated with /r/Depression discourse (rDD) using the data retrieved during our first step above, we follow the Archetype Based Modelling and Search (ABMS) approach of Davis et al. [23], which we briefly describe here. The approach consists of four main steps.

1. Learn a word representation using the discourse of interest (here, discourse on the topic of depression)

2. Learn author or document representations for the archetypal class (e.g., documents written by authors who participate in discourse about depression) and for a control class

3. Separate the two classes of authors using supervised learning

4. Extract words which are most strongly related to the archetypal class

First, we construct a word embedding, which creates a vector representation for every word in the vocabulary used by the individuals engaging in discourse related to depression. The word embedding was trained using GloVe [24]. The dimensionality was set to 300, the learning rate to 0.1, the number of iterations to 100, and the convergence tolerance to 0.01. The full vocabulary size of 201,375 words is used to train the word embedding. GloVe was chosen for this study, as it has been shown to capture semantic relationships in the distances between word representations [25].

Second, we construct author representations for all authors, both affiliated with /r/Depression and contrasting, using as a foundation the word representations of the words they use. We follow the usr2vec protocol [26] to do so. Author representations maintain the same dimensionality as the word representations, a fact which is important for subsequent processing. Out-of-vocabulary words (that is, any words that occur that are not in the original corpus from which rDD are extracted), are not used in the construction of author representations; if too many words used by an author were out of vocabulary, that author was discarded from the training set. The amount of computation needed to train this stage is approximately 39 CPU-years; however, individual representations of authors can be trained in parallel. Training was distributed over multiple nodes on Compute Canada.

Third, using the resulting author representations as feature vectors, a support vector machine (SVM) was trained to separate cases from controls using the ThunderSVM package for GPU acceleration [27]. To evaluate the effectiveness of the SVM at separating contrasting the authors active in /r/Depression from others, we made an 80/20 train/test split. The SVM achieved an 81.6% accuracy on the test set, demonstrating that the SVM was able to reliably distinguish between our two groups of authors. We then re-trained the SVM on all of the data before proceeding to the next step.

Fourth, we identified the vocabulary that was most strongly associated with being the /r/Depression authors. To do this,

we first pruned the original vocabulary of 201,375 words by removing those that occurred fewer than 750 times in the corpus. This removed many unusual and infrequent words as well as most of the words with strange characters or artifacts. Within this subset, we used the decision direction vector from the SVM, which is orthogonal to the separating hyperplane, to identify the words in the vocabulary most strongly associated with rDD. To do this, we computed the dot product similarity between each word representation and the decision direction vector and sorted the words from positively (maximum 1.0) to negatively (minimum -1.0) associated with rDD. The resulting top 1000 words were extracted based on their computed association and then refined down to a list of 400 words by manual review; this process and its rationale are described below. Words which were deemed to be too common or uninformative out of context were removed.

Information retrieval scoring & case count comparison

With the keywords from the modeling stage extracted, the next step was to index the documents of interest.

For Toronto, Seattle, Sydney, and Mumbai, we retrieved all Twitter posts from April 2019 to June 2021 with a geotagged location within a certain radius of the city center (as according to Twitter's API). These cities were selected for their relative isolation from one another, different COVID-19 case trends, and availability of geotagged tweets. The date range was chosen to provide context of a year of Twitter activity prior to the general presence of COVID-19.

These tweets were then indexed using Elasticsearch, a search engine that uses the BM25 scoring metric to match documents with keyword queries [28]. The BM25 score is used in information retrieval to assess how relevant a document is to a query; documents with more keywords in common with a query have a higher BM25 score, and documents with no keywords in common have a score of zero [29]. To retrieve documents, Elasticsearch uses the query to compute the BM25 score for each document in the corpus and then returns the highest-scoring documents. For our work, we retrieved the BM25 score for *all* documents (tweets) within a corpus (all tweets in a geographic region) and aggregated by week to evaluate how much the tweets for that week were aligned with the rDD score. Individual scores were assigned to every tweet in the set for each city. For a given city, each week's score was the average of the scores of all posts occurring that week.

To investigate potential relationships between the rDD scores and COVID-19 case counts, data were retrieved for Toronto, Seattle, Sydney, and Mumbai. Using the number of confirmed cases, the change per day was calculated and summed per week. We calculated the Pearson correlation between the average weekly rDD scores and the change in number of cases for each week.

Results

rDD keywords

A list of the top 1000 words and the manually curated list of 400 are provided in Appendices A and B. These are the words

that are most associated with the representations of users who participate in /r/Depression in our model as measured by dot product. Words throughout the range of the top 1000 had plausible connections to discussion about depression; for example, the first ten words appearing in the list were "care, exams, rocks, bottles, tear, controlled, methods, inch, violates, storm" and the last of the 1000 was "rehab". In the 450–550 range, there are words such as "grandparents", "deaths", "unstable", "flawed", "psychologist", "trash", and "hopeless". The diversity of words that occur in this range may reflect different aspects of the experience of depression. There are two issues with the list that led us to undertake some manual curation to reduce its length. First, for this analysis we elected to include only words that had at least some "face validity," meaning that they had a plausible connection to discourse about depression; this led us to eliminate, for example, single characters like "i" and "F." Second, because our Elasticsearch instance limited the maximum number of characters in a given query, we were restricted to using just over 400 words (which hit the character limit); in response to this, we filtered our list down to approximately match the maximum number of words we could use. We note that others following our framework could filter the list themselves to match their particular needs, and could manually add additional words based on their expertise if they felt any were missing.

Changes in rDD over time

Using Elasticsearch, we scored all tweets from our Toronto, Seattle, Sydney, and Mumbai Twitter datasets by running a query using the 400 manually curated keywords we identified as being associated with a higher rDD score. The rDD trends over time from April 2019 to June 2021 for Toronto, Seattle, Sydney, and Mumbai can be seen in Figure 1. In all cases, there is an increase of rDD language that occurs at the start of the pandemic in March 2020, and then gradually returns to levels at or below what is observed before the pandemic. A notable exception occurred in the case of Mumbai, where the rDD score increased at the same time as a very large second wave of COVID-19 across India.

Associations between rDD and COVID-19 case counts

In Figure 2, the scatterplots show weekly COVID-19 case counts (horizontal axis) and weekly rDD scores (vertical axis) over time (point shading) for Toronto, Seattle, Sydney, and Mumbai, respectively. When considering January 2020 to June 2021, significant ($p < 0.05$) correlations were found between the rDD scores in Toronto and Seattle, where rDD was very high during the announcement of the pandemic in March 2020, and cases did not increase until later (Table 1). However, a positive linear trend was observed among the points representing the April-through-June 2021 wave in Mumbai (darker coloured points.)

We present the COVID-19 case counts and weekly rDD scores together over time from April 2019 to June 2021 in Figure 3 for Toronto, Seattle, Sydney, and Mumbai, respectively. The scale of each axis for the cities is based on the highest amount of daily change for the time period considered, which varies considerably between cities. Mumbai's graph is

Figure 1: rDD scores over time for each city between April 2019 and June 2021
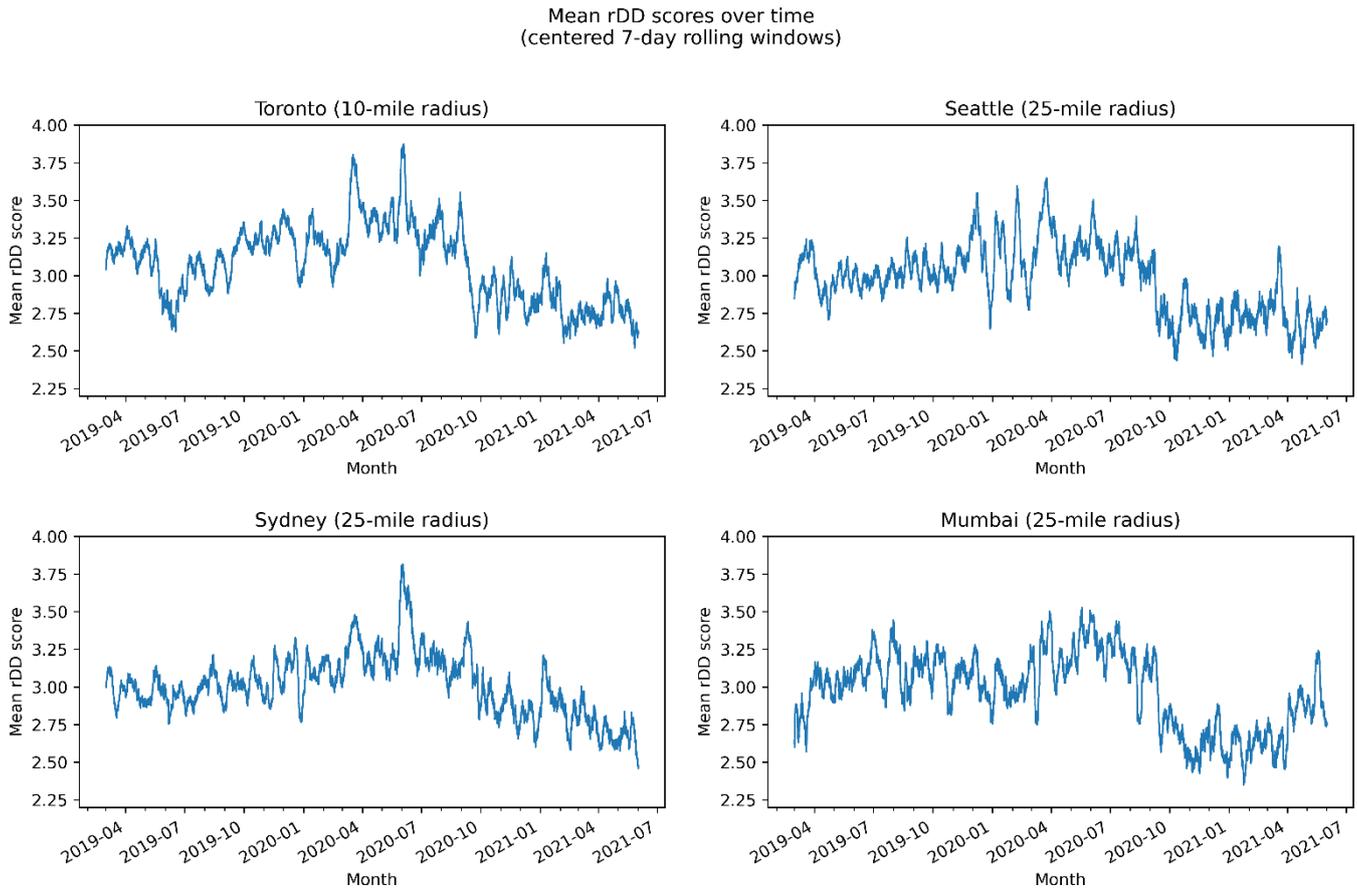
Mean rDD scores over time
(centered 7-day rolling windows)

Discussion

The emerging narrative in the research literature on mental health during the pandemic is that there was an increase in poor mental health in early 2020 when little was known about what pandemic life would be like, and that this tapered off after it became clear what the government and society's response would be, and what day-to-day life would look like in the coming months [14]. The largest population-wide peak of rDD score occurred during the early days of the pandemic around March 2020, which lends credibility to rDD score being a measure of population mental health. Furthermore, we find that, in Toronto, our results are consistent with the Shields report, which found that population-wide mental health began to return to normal around July 2020 [15]. In addition, our results are consistent with the conclusions of *The Lancet*'s COVID-19 Commission Mental Health Task Force, which summarizes the findings of several high-quality studies [14]. The report in *The Lancet* found that overall anxiety, depression, and distress increased in the early months of the pandemic, starting around March 2020. Meanwhile, suicide rates, life satisfaction, and loneliness remained largely stable throughout the first year of the pandemic.

The exception to this pattern was observed in Mumbai's second wave. This large wave, which occurred in mid-2021, was associated with significant worldwide media coverage of the severe strain on the healthcare system in India, which was exacerbated by shortages of key supplies like oxygen [30]. This period was also associated with a substantial increase in rDD keywords. This may be a situation that calls into question the hypothesis that mental health was negatively impacted primarily at the beginning of the pandemic. A targeted survey could be used to assess whether this is the case.

Unexpectedly, in all cities in our dataset (though less prominent in Seattle), we observed a dip in the rDD score around September 2021. Initially, we hypothesized that the dip could be the result of faulty data collection or some other problem with data integrity. However, we performed an in-depth investigation of the data collected around this time period, including manual examination of selected tweets, and did not find evidence to support this theory. Seasonal effects were considered, but they seemed unlikely to persist across the geographies involved in Toronto, Seattle, Sydney, and Mumbai. Some of the factors we speculate may have contributed to this period include loosening of pandemic restrictions across the globe, a coincidental occurrence of the start of the school year, holidays or celebrations, and potential changes in social media moderation and accounts in close proximity to the 2020 US elections.
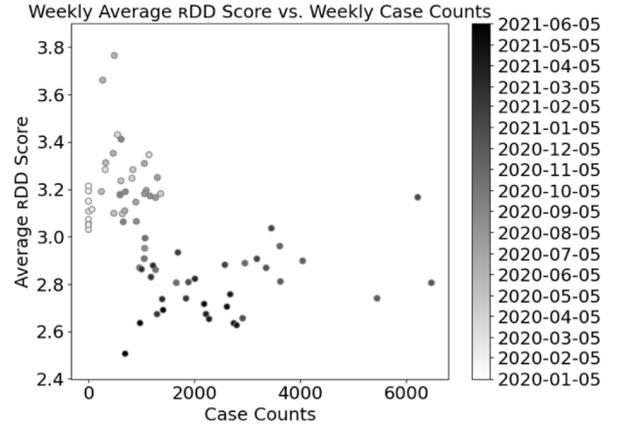
Related work and comparisons

One of the challenges in using social media data for research is finding "labelled data." For example, identifying
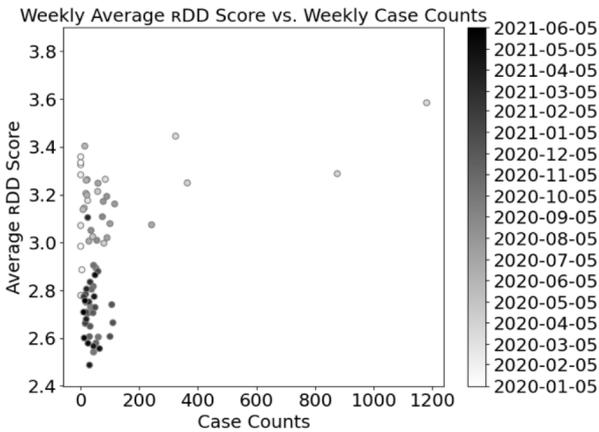
Figure 2: rDD scores vs. case counts per city. Points are shaded to show progression over time
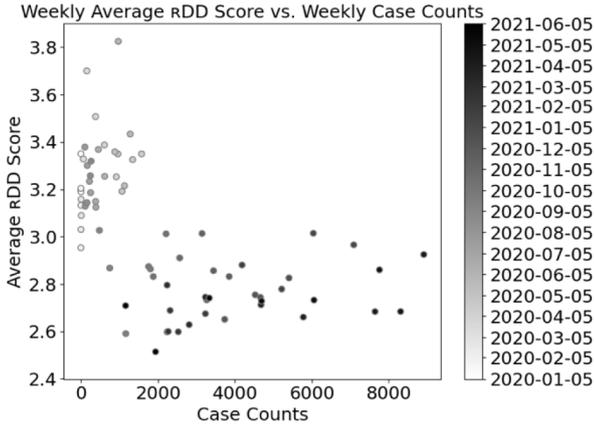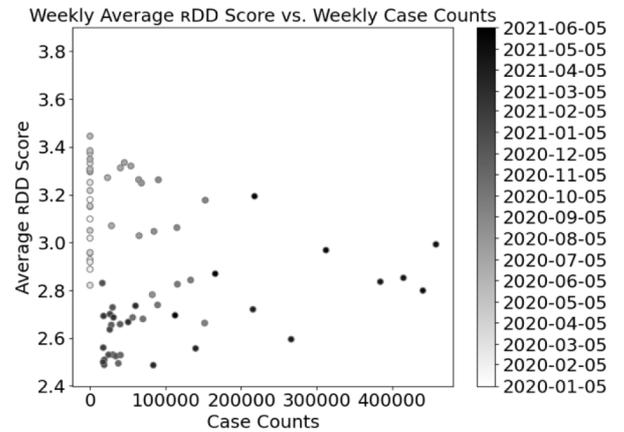

(a) Toronto rDD scores vs. case counts

(b) Seattle rDD scores vs. case counts

(c) Sydney rDD scores vs. case counts

(d) Mumbai rDD scores vs. case counts

Table 1: Pearson correlation over various time periods between weekly case counts and average weekly rDD scores. Bold values are significant, with p-values ≤ 0.05

City	January 2020–Sept 2020	Sept 2020–June 2021	Jan 2020–June 2021
Toronto	**0.378**	0.103	**−0.581**
Seattle	0.060	−0.002	**−0.458**
Sydney	**0.471**	−0.098	**0.382**
Mumbai	−0.086	**0.566**	−0.157

content that is associated with specific characteristics like depression is notoriously difficult [31]. Whereas in our work we directly sought out users who discuss depression through their participation in /r/Depression, related work has relied on manually labelled data or explicitly constructed rules that are described as identifying individual users who "have depression." There have been a number of other models intended to detect depression on social media; these range from using other methods intended for sentiment analysis, to using specialized deep learning architectures which attempt to estimate living environments, and which may not transfer well outside of their training data [32]. There have been other studies which use machine learning approaches to study language associated with depression, such as one using pre-trained transformer models in conjunction with a smaller set

of individual users who have been manually labelled as having depression based on pre-specified non-clinical criteria [33]. This study found that sentiment analysis did not reliably capture indicators of depression that their transformer-based models could find.

As noted previously, sentiment analysis techniques have been used to assess population-level sentiment, which is sometimes used as a surrogate for a measure of mental health. We compared our approach to the scoring produced by a common sentiment analysis tool, VADER [34]. Line plots for the April 2019 to June 2021 time period are given in Appendix C. We observe that overall, rDD and VADER scores track (inversely because lower VADER scores indicate more negative sentiment) over the time period we investigated. However, VADER scores do not change between May 2021 and June

Figure 3: Region case counts (centered 7-day window mean daily change) in orange and rDD scores (centered 7-day window mean) in blue for the Twitter posts

COVID-19 case change and rDD scores over time
(centered 7-day rolling windows)
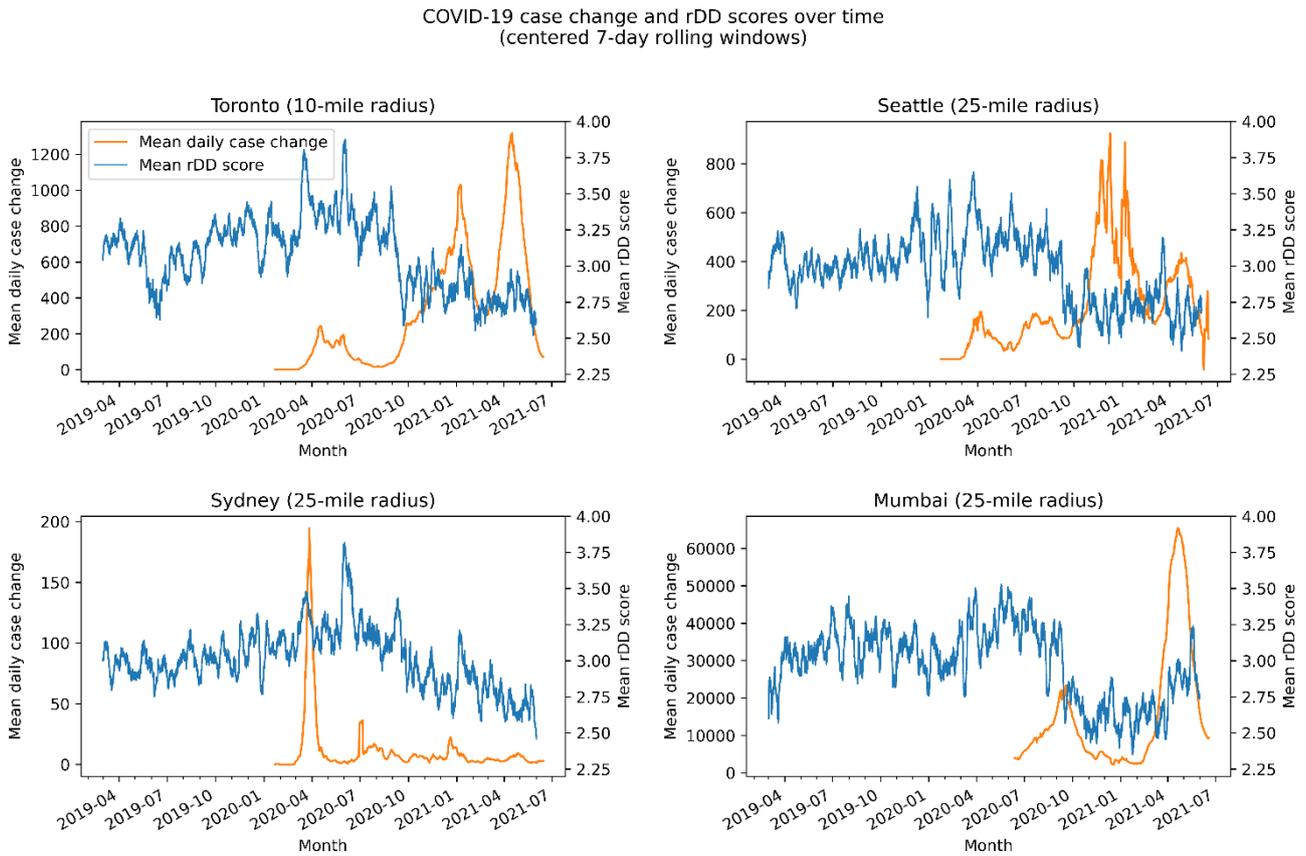
2021. We briefly contrast rDD and VADER by interpreting them as a measure of population depression related mental health.

The VADER score for Toronto from May to June 2021 indicates that the sentiment is the lowest, i.e., negative sentiment is the highest, that it has ever been during the entire pandemic; we see a decrease from the original score of 0.275 in February 2020 to a score of 0.150 by June 2021; this is a worsening of 45%. The rDD score changes from 3.0% to 2.7% in the period of February 2020 to June 2021, which is a 10% improvement since the start. This reduction in rDD score is more consistent with the results from Aknin et al., who found that after the initial months of the COVID-19 pandemic, many measures of mental health burden returned to normal amounts, and in some cases improved when individuals were able to engage in other activities such as gardening [14].

For Toronto, Shields et al., the rate of depression increased from 7% to 16% [15] by September 2020. From February 2020 to September 2020, there was a change in the rDD score of 3.0 to 3.3. Following this period, we do see a decrease in rDD score which persists through June 2021. In contrast, VADER has a change from 0.3 to 0.2, suggesting a 33% increase in negative sentiment, to an eventual 50% increase in negative sentiment by June 2021. Based on the studies identified, rDD seems to be matching the magnitude of change observed from surveys and clinical reports more closely than VADER is.

When contrasting rDD to VADER, we find there are some advantages to our approach. The rDD approach (i) uses language extracted from individuals active in communities associated with being depressed, and (ii) achieves comparable results even after filtering down to 400 keywords instead of the over 7000 used in the VADER vocabulary. The adaptability of the rDD approach allows the vocabulary to find words which capture associations that VADER potentially does not; the increase in rDD with the spike in cases in Mumbai could be through one or more of these words. Finally, the language used in rDD can be revised either automatically or manually on a regular basis to stay current.

Lastly, we contrast rDD with a population-based study of acute mental health service use by Saunders et al. in Ontario, Canada [35]. Administrative data were examined for the trends in hospitalizations and emergency department visits due to mental health diagnoses and substance use disorders, and emergency department visits for intentional self-injury from 1 January 2019 to 31 March 2021. Significant inflection points were found after the pandemic began in March 2020, with a decrease in overall emergency department visits and hospitalizations for mental health and substance use disorders. Visits for self-intentional injury also dropped by 33% in April 2020 compared to April 2019 and returned close to pre-pandemic levels by August 2020. From these trends, Saunders et al. conclude that increased stressors of the pandemic did not equate to increased service use for acute care in the 12 months following the pandemic; however, they note that this does not mean there was no mental health impact, reasoning that this could be due to changes in admission thresholds, system capacity, reduction in lethality of cases, or shifts to ambulatory care or outpatient care which was not captured

in the analysis. Furthermore, they note that people could have avoided hospitals and emergency rooms due to fear, and could have availed themselves of newly available virtual care options. The results by Saunders et al. contrast trends in rDD, as rDD scores peaked at the start of the pandemic rather than dropped. This may be because of the reasons noted by Saunders et al. which emphasize that experiences of poor mental health need not be associated with health services use. Furthermore, the relationship between poor mental health and service use is likely highly context-dependent; the reduction in service use noted by Saunders et al. did not match trends in other jurisdictions [36–38]. Thus, rDD and social media generally represent a different, complementary view of population mental health than measures of mental health care utilization.

Limitations

The ability of our approach to detect discourse about depression as it appears on social media is limited by the way it can be expressed through word frequency: presentations outside of word frequency cannot be captured by a keyword model, as only the terms in the query are used. There are many possible explanations as to which words and frequencies capture this. For example, one of the words in the model is "grandparents". In the context of authors describing depression, this could be involving medical problems with their grandparents. This transfers to the context of Twitter, where individuals concerned about the well-being of their grandparents or elderly in general who are more susceptible to COVID might vent. Words such as these, which can have complex significance, seem to be responsive to events that are causing widespread increases in discourse about depression. Extracting these words and using them as part of public health surveillance of social media helps with estimating trends at a population scale, even if the words on their own may not form a comprehensive monitor.

Assessing the representativeness of social media users is challenging. For general social media use, reports from Australia, Canada and the USA detail a trend across both Twitter and Reddit that social media users tend to be younger, and in Reddit's case a higher proportion of active users identify as men [39–41]. Furthermore, Twitter users who use geotagging, whose posts we exclusively used in our experiment, are demonstrably different from those who do not, with some evidence for a slightly older and more female population than on Twitter overall [42, 43]. We did not find demographics by age for India. The specific demographics of individuals that are active in /r/Depression are unknown, so we cannot directly assess any differences for that specific subreddit. The language extracted as a result will be representative of the entire social media user population, but this population has some differences to the general population of each country that should be considered. The severity of COVID-19 is highly correlated with advanced age, while social media usage is dominated by younger age groups. Additionally, depression symptoms are known to be more prevalent among women than the population average [44]. Hence, because of a potential underrepresentation of women and elderly posters in our dataset compared with the broader populations for

each country, our study may under-capture the depressive experiences of each city's population.

Our work is descriptive rather than analytic and does not address the causal mechanisms relating COVID-19 to mental health. There are events that are not directly related to the pandemic that can cause great stress and strain at the population level, and these can cause an increase in rDD. The governmental response in different locales, including lockdowns, public health restrictions, and pandemic-related policies can all cause strain that is not directly related to the case count, except, potentially, for the window of cases leading directly up to the restriction.

Another limitation is that individuals discussing their experience with depression or other conditions may move to other messaging mediums, such as direct, or private, messages. These are not included in the Twitter feed, and would not be counted in any word frequency approach. Ongoing long-term effects are harder to see in these population-level trends. It is reported that a 9% or so increase over the regular rate of depression was observed in Canada during 2020, equating to slightly more than a doubling of the rate in the population [15]. Individuals experiencing slowly increasing chronic effects would not be captured by this model unless it were happening at scale, so it is not suitable for identifying increased individual distress. These slowly increasing chronic effects for a subsection of the population are different considerations than trying to capture population-wide changes at scale in rDD. Capturing the continued worsening experience of these chronic effects over social media at scale may be difficult, as it could be dispersed across other independent trends and across the time range considered.

Conclusion

We have presented a framework and case study that examines language on social media that is associated with discourse about depression. Rather than relying on manual categorisation of content and users, our approach leverages publicly available social media data from a community that discusses depression. From this, we generate both a set of keywords associated with depression discourse and a score that indicates how prevalent this discourse is over time. Examining the score over the COVID-19 pandemic reveals trends that are concordant with established research on the rate of depression being highest at the beginning of the pandemic and decreasing over time; however, our analyses also indicate that in Mumbai there was a significant increase in rDD which may be associated with a second wave. This finding identifies a potentially important area for future investigation by population health and global health researchers to better understand whether experience of depression induced by the pandemic was different in this region.

There are many avenues for future work based on new applications and extensions of the framework we have presented. For this analysis, we chose to manually curate the list of keywords down to a more manageable size based on face validity. However, our approach could be used to discover new relationships between social media-specific language and discourse about depression; for example, the single character 'F' has a memetic meaning often related to someone else

posting 'Fs in the chat' to incite public sympathy for an event. Posts containing only the character 'F' will follow in reply to indicate sympathy. Future work could include an online ethnographic approach that uses knowledge of Internet subculture to identify new ways that users express themselves when discussing depression or other topics of public health interest. A related avenue would be to develop interactive visualizations to help users understand whether specific subsets of words are "driving" the rDD score, and whether those change over time.

While we focus on population-level analyses, similar approaches may be used to capture individual journeys of depression on social media; however, this requires particular attention to privacy considerations and is complicated by the possibility of intervention on a personal scale. As an intermediate approach between the individual and the population levels, individual trajectories of rDD could be clustered to identify heterogeneous groups of individuals whose experiences over time are similar.

Expanding our methodology to use social media data to better understand the way language is used and to better characterize different experiences of mental health over time will be an important piece of understanding the long-term pandemic-related impacts on mental health.

Acknowledgements

Statement on conflicts of interest

The authors have no conflicts of interest to declare.

Ethics statement

We did not seek ethics approval since the data are publicly available, and we only present information in aggregate so that no individual participants' identities are exposed.

References

1. J. Bueno-Notivol, P. Gracia-García, B. Olaya, I. Lasheras, R. López-Antón and J. Santabárbara, "Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies," *International journal of clinical and health psychology*, vol. 21, p. 100196, 2021. https://doi.org/10.1016/j.ijchp.2020.07.007

2. O. V. Ebrahimi, A. Hoffart and S. U. Johnson, "Physical distancing and mental health during the COVID-19 pandemic: Factors associated with psychological symptoms and adherence to pandemic mitigation strategies," *Clinical Psychological Science*, vol. 9, p. 489–506, 2021. https://doi.org/10.1177/2167702621994545

3. M. Daly and E. Robinson, "Psychological distress and adaptation to the COVID-19 crisis in the United States," *Journal of psychiatric research*, vol. 136, p. 603–609, 2021. https://doi.org/10.1016/j.jpsychires.2020.10.035

4. K. Relia, M. Akbari, D. Duncan and R. Chunara, "Socio-spatial self-organizing maps: using social media to assess relevant geographies for exposure to social processes," *Proceedings of the ACM on human-computer interaction*, vol. 2, p. 1–23, 2018. https://doi.org/10.1145/3274414

5. A. Wongkoblap, M. A. Vadillo and V. Curcin, "Researching mental health disorders in the era of social media: systematic review," *Journal of medical Internet research*, vol. 19, p. e228, 2017. https://doi.org/10.2196/jmir.7215

6. C. Cuello-Garcia, G. Pérez-Gaxiola and L. van Amelsvoort, "Social media can have an impact on how we manage and investigate the COVID-19 pandemic," *Journal of clinical epidemiology*, vol. 127, p. 198, 2020. https://doi.org/10.1016/j.jclinepi.2020.06.028

7. A. K. M. Chan, C. P. Nickson, J. W. Rudolph, A. Lee and G. M. Joynt, *Social media for rapid knowledge dissemination: early experience from the COVID-19 pandemic*, Wiley Online Library, 2020. https://doi.org/10.1111/anae.15057

8. N. H. Di Cara, J. Song, V. Maggio, C. Moreno-Stokoe, A. R. Tanner, B. Woolf, O. S. P. Davis and A. Davies, "Mapping population vulnerability and community support during COVID-19: a case study from Wales," *International journal of population data science*, vol. 5, 2020. https://doi.org/10.23889/ijpds.v5i4.1409

9. M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo and A. Scala, "The COVID-19 social media infodemic," *Scientific Reports*, vol. 10, p. 1–10, 2020. https://doi.org/10.1038/s41598-020-73510-5

10. D. Valdez, M. Ten Thij, K. Bathina, L. A. Rutter and J. Bollen, "Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of twitter data," *Journal of medical Internet research*, vol. 22, p. e21418, 2020. https://doi.org/10.2196/21418

11. A. Zunic, P. Corcoran, I. Spasic and others, "Sentiment analysis in health and well-being: systematic review," *JMIR medical informatics*, vol. 8, p. e16023, 2020. https://doi.org/10.2196/16023

12. G. D. Rosin, E. Adar and K. Radinsky, "Learning Word Relatedness over Time," July 2017. https://doi.org/10.18653/v1/D17-1121

13. K. Loveys, J. Torrez, A. Fine, G. Moriarty and G. Coppersmith, "Cross-cultural differences in language markers of depression online," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018. https://doi.org/10.18653/v1/W18-0608

14. L. Aknin, J. E. De Neve, E. Dunn, D. Fancourt, E. Goldberg, J. Helliwell, S. P. Jones, E. Karam, R. Layard, S. Lyubomirsky, A. Rzepa, S. Saxena, E. Thornton, T. VanderWeele and others, "A review and response to the early mental health and neurological consequences of the COVID-19 pandemic," *Perspectives on Psychological Science*, 2021. https://doi.org/10.31234/osf.io/zw93g

15. M. Shields, L. Tonmyr, A. Gonzalez, M. Weeks, S.-B. Park, A.-M. Robert, D.-L. Blair and H. L. MacMillan, "Symptoms of major depressive disorder during the COVID-19 pandemic: results from a representative sample of the Canadian population," *Health Promotion and Chronic Disease Prevention in Canada*, vol. 41, September 2021. https://doi.org/10.24095/hpcdp.41.11.04

16. Y. Bengio, A. Courville and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, p. 1798–1828, 2013. https://doi.org/10.1109/TPAMI.2013.50

17. *15 Most Popular Social Media Sites of the Last 10 Years.* URL: https://www.newsweek.com/15-most-popular-social-media-sites-facebook-tiktok-whatsapp-1627665

18. J. Baumgartner, S. Zannettou, B. Keegan, M. Squire and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, 2020. https://ojs.aaai.org/index.php/ICWSM/article/download/7347/7201/10577

19. L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar and R. M. Merchant, "Twitter as a tool for health research: a systematic review," *American journal of public health*, vol. 107, p. e1–e8, 2017. https://doi.org/10.2105/AJPH.2016.303512

20. CSSEGISandData, *Covid-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*, GitHub, 2021. URL: https://github.com/CSSEGISandData/COVID-19

21. E. Dong, H. Du and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, vol. 20, p. 533–534, May 2020. https://doi.org/10.1016/S1473-3099(20)30120-1

22. *Ontario COVID-19 Data Tool |Public Health Ontario.* https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool

23. B. Davis, K. Sedig and D. Lizotte, "Archetype-Based Modeling and Search of Social Media," *Big Data and Cognitive Computing*, vol. 3, no. 3, p. 44, July 2019. https://doi.org/10.3390/bdcc3030044

24. J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014. https://doi.org/10.3115/v1/D14-1162

25. T. H. Muneeb, S. Sahu and A. Anand, "Evaluating distributed word representations for capturing semantics of biomedical concepts," in *Proceedings of BioNLP 15*, 2015. https://doi.org/10.18653/v1/W15-3820

26. S. Amir, G. Coppersmith, P. Carvalho, M. J. Silva and B. C. Wallace, "Quantifying mental health from social media with neural user embeddings," in *Machine Learning for Healthcare Conference*, 2017. http://proceedings.mlr.press/v68/amir17a.html

27. Z. Wen, J. Shi, Q. Li, B. He and J. Chen, "ThunderSVM: A fast SVM library on GPUs and CPUs," *The Journal of Machine Learning Research*, vol. 19, p. 797–801, 2018. https://doi.org/10.5555/3291125.3291146

28. C. Gormley and Z. Tong, Elasticsearch: the definitive guide: a distributed real-time search and analytics engine, O'Reilly Media, Inc., 2015. ISBN: 9781449358549

29. S. Robertson and H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Now Publishers Inc, 2009. https://doi.org/10.1561/1500000019

30. *Oxygen shortage forces BMC to shift 168 Covid-19 patients from 6 hospitals |Mumbai news - Hindustan Times.* https://www.hindustantimes.com/cities/mumbai-news/oxygen-shortage-forces-bmc-to-shift-168-covid-19-patients-from-6-hospitals-101618766338181.html

31. E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the international conference on Web search and web data mining - WSDM '08*, New York, New York, USA, 2008. https://doi.org/_10.1145/1341531.1341557

32. G. K. Gupta and D. K. Sharma, "Depression detection on social media with the aid of machine learning platform: A comprehensive survey," *Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom 2021*, p. 658–662, March 2021. https://doi.org/10.1109/INDIACom51348.2021.00116

33. Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang and J. Luo, "Monitoring Depression Trends on Twitter During the COVID-19 Pandemic: Observational Study," *JMIR infodemiology*, vol. 1, p. e26769–e26769, July 2021. https://doi.org/10.2196/26769

34. C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399

35. N. R. Saunders, A. Toulany, B. Deb, R. Strauss, S. N. Vigod, A. Guttmann, M. Chiu, A. Huang, K. Fung, S. Chen and others, "Acute mental health service use following onset of the COVID-19 pandemic in Ontario, Canada: a trend analysis," *Canadian Medical Association Open Access Journal*, vol. 9, p. E988–E997, 2021. https://doi.org/10.9778/cmajo.20210100

36. S. A. Simpson, R. M. Loh, M. Cabrera, M. Cahn, A. Gross, A. Hadley and R. E. Lawrence, "The Impact of the COVID-19 Pandemic on Psychiatric Emergency Service Volume and Hospital Admissions," *Journal of the Academy of Consultation-liaison Psychiatry*, 2021. https://doi.org/10.1016/j.jaclp.2021.05.005

37. K. M. Holland, C. Jones, A. M. Vivolo-Kantor, N. Idaikkadar, M. Zwald, B. Hoots, E. Yard, A. D'Inverno, E. Swedo, M. S. Chen and others, "Trends in US emergency department visits for mental health, overdose, and violence outcomes before and during the COVID-19 pandemic," *JAMA psychiatry*, vol. 78, p. 372–379, 2021. https://doi.org/10.1001/jamapsychiatry.2020.4402

38. L. R. Joyce, S. K. Richardson, A. McCombie, G. J. Hamilton and M. W. Ardagh, "Mental health presentations to Christchurch Hospital Emergency Department during COVID-19 lockdown," *Emergency Medicine Australasia*, vol. 33, p. 324–330, 2021. https://doi.org/10.1111/1742-6723.13667

39. A. Gruzd and P. Mai, "The State of Social Media in Canada 2020," *SSRN Electronic Journal*, 2020. https://doi.org/10.5683/SP2/XIW8EW

40. S. Wojcik and A. Hughes, "Sizing Up Twitter Users," Pew Research Center, 24 April 2019. [Online]. Available: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. [Accessed 21 01 2022].

41. A. Ramshaw, "Social Media Statistics for Australia (Updated January 2022)," Genroe, 11 1 2022. [Online]. Available: https://www.genroe.com/blog/social-media-statistics-australia/13492#australian-social-media-usage-by-age. [Accessed 21 1 2022].

42. L. Sloan and J. Morgan, "Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter," *PLoS ONE*, vol. 10, no. 11, p. e0142209, 2015. https://doi.org/10.1371/journal.pone.0142209

43. R. R. Kadari, L. Panati, S. P. Nooli, H. Bheemreddy and P. Bozorgi, "Analysis of geotagging behavior: Do geotagged users represent the Twitter population?," *ISPRS International Journal of Geo-Information*, vol. 10, no. 6, p. 373, 2021. https://doi.org/10.1145/3341161.3342870

44. R. H. Salk, J. S. Hyde and L. Y. Abramson, "Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms," *Psychological Bulletin*, vol. 143, no. 8, pp. 783-822, 2017. https://doi.org/10.1037/bul0000102

Appendices

Appendix A   Unfiltered search list

(Please note that this list may contain words some readers may find offensive.)

some, cause, Good, affects, hunt, care, exams, tf, sits, websites, Hi, path, UH, every, daughters, rocks, broad, generations, Depends, F, tracks, albums, cents, bottles, sisters, songs, shine, edits, closely, Some, tear, savings, headed, channel, lord, channels, controlled, methods, inch, violates, Last, profanitycounter, nipples, investigation, storm, teammates, self, warrior, dwight3469, pregnancy, attempts, course, favorites, shit, rain, amazed, rule, shit, decrease, wounds, of, rough, suited, ghost, system, yourself, destiny, addiction, paths, shits, scars, courses, I, ratelimited, warren, tricks, arrest, envy, behaviors, sites, hill, transition, listening, wind, account, rainbow, backs, hearts, semester, suicidebombereyelash, quitting, sentences, cousins, One, First, phones, dip, favourite, losses, investing, bless, programs, sale, youtube, bald, tests, fortnite, work, scrolling, unaware, assure, falling, bond, recovering, billions, nI, fund, patches, giant, versions, you, improving, or, projecting, minecraft, i, hunting, discussing, steps, taper, alarm, roads, planets, consumption, offense, No, confuse, spray, records, guard, I, episodes, riot, YouTube, liberals, crack, Okay, knee, dependent, click, staying, urban, bright, wages, fields, spamming, weeks, combo, boost, station, To, days, weeks, months, green, draining, articles, All, addicts, man, plays, rounds, deaf, abuse, ago, spam, wide, duty, depend, journey, paste, trail, well, elaborate, blind, supply, tube, drain, bull, be, events, nuclear, unfortunate, habits, suits, nnI, I, treats, poll, interacting, system, continuing, pops, roommates, dying, rap, hop, decades, it, wiki, i, parties, sigh, streak, financially, emotionally, tour, subs, eyes, subreddits, it, chances, cap, miracle, surrounded, profile, focusing, masculinity, wild, lease, recovery, overcome, lil, yrs, systems, decks, road, poor, heads, tears, smoke, cycle, increase, shoulders, guide, affecting, cs, adventure, opponents, photos, ancientnucleus, bleed, blowing, batman, angel, becoming, Usually, shield, instances, pattern, viewing, distant, podcast, independent, long, days, politically, tripping, tall, red, largest, see, ads, adopt, cameras, survivors, Black, debt, fell, grace, officers, collect, twitter, businesses, Could, alt, loneliness, theories, national, zeebuoy, projects, is, mixing, drops, cherry, recording, satisfaction, huge, right, big, sessions, saving, voice, pays, off, earned, quality, breaks, Oh, fits, elsewhere, too, abusing, tricky, however, lead, traumatic, friendships, When, nights, reminding, tower, campaign, bat, silence, daddy, brainwashed, asses, sucks, strike, chapter, analysis, shows, apps, interviews, command, slip, agency, notice, slow, music, meals, hours, season, too, army, baking, grad, suffer, attack, scar, casually, FUCKING, protests, improve, towns, enough, addict, december, dxm, films, dear, ear, song, purple, royal, goals, clip, 30s, desert, walks, nnWhat, dive, ruin, falls, nnIt, factors, someday, unattractive, feel, doing, dare, same, shitting, typically, fall, unlimited, treatments, wipe, lover, reflect, break, whining, sorry, It, pink, studying, it, agent, guy, soon, you, games, pack, damned, spots, album, books, reminded, inches, Which, Same, tape, squeeze, list, relatives, moments, binge, cave, isolated, rock, fox, that, dating, this, cups, thing, sets, behaviour, myself, indicate, tanks, We, hunger, fulfilling, swim, am, heading, sir, prices, tolerance, fishing, neglect, illusion, pray, alert, classes, foster, jealous, trend, widely, increased, breakdown, canadian, inclined, grandparents, Just, isolate, scroll, adayinthelife, pathetic, pull, deaths, incidents, connections, sitting, shed, neat, corporations, cake, God, ultra, shit, grain, statements, consuming, growth, 4k, man, lap, disagree, imaginary, plants, thanks, grass, cleared, playing, solar, twitch, More, You, farm, spends, flesh, art, collecting, proves, clips, saves, birthday, someone, creep, pc, pulling, wonders, blow, version, nails, aware, psychologist, flawed, well, unstable, relationships, humanity, gas, friend, area, cells, incapable, stars, destroying, mozza7, spirit, about, trash, dirty, vanilla, oldest, vary, hopeless, releasing, Please, fur, harmless, apples, dry, insert, sucking, buttons, That, devices, periods, friends, rooting, empire, accounts, memes, behavior, homes, raising, cheer, idiots, It, lungs, inevitably, takes, reviewed, buddies, lights, hours, buildings, Will, People, seed, dates, years, music, A, passive, Which, endless, pile, sun, philosophy, republicans, tinder, grams, monster, typing, massive, politicians, legacy, lessons, conditions, following, now, hunter, exceptions, breakup, elections, collection, wing, david, itself, months, themselves, blew, rape, care, digging, driving, bubble, trap, Thanks, cry, unknown, clarification, simultaneously, advocate, off, research, yourself, glasses, bags, mountain, influence, fair, continue, lifestyle, years, rant, dried, agree, Sorry, downvotes, guests, audio, lol, loads, weigh, reposting, question, opening, decade, hall, privacy, a, surviving, too, Reddit, well, listen, walking, drinking, are, communities, Omg, stubborn, receive, bosses, precious, dreaming, work, sat, options, TV, nut, handling, That, buddy, This, sensitive, gt, exploring, John, next, round, strictly, surely, dark, loans, blocking, vehicles, salt, dc, last, victim, trucks, hm, safely, social, to, automatically, acting, fb, edge, dude, introvert, spending, depending, whom, one, Now, feed, breath, lips, understand, technology, want, sit, gods, itll, evil, deep, armed, trips, favorite, funds, fuel, neighbors, harsh, infinite, posting, upside, 2020, Ill, legends, karma, software, graphics, coverage, cares, crossing, delay, faq, pup, risk, abandon, snap, worthy, freak, thick, cow, negative, celebrating, Tell, tiny, interpretation, bleeding, corporate, As, creatures, passion, rehab, n, organization, pants, passes, dropped, relying, turns, speakers, runs, guitar, Wow, opposed, times, chunk, holy, not, to, shit, imagination, glitch, mommy, resources, forgetting, While, payment, classmates, environmental, lane, blocks, sleeping, Think, highlight, observation, controllers, units, attacks, threatening, services, hollywood, change, naive, platform, bedroom, status, away, haha, ears, Yes, puberty, spring, strings, trees, stays, ruins, are, things, precisely, apart, frequently, ones, raise, videos, Right, habit, fallout, mike, footage, directed, deciding, things, major, mechanisms, worthless, small, oral, socialism, xp, By, turning, deeply, planet, earn, jedi, nations, gear, crisis, wealth, review, hilarious, boi, too, deleting, sticking, watching, can, depression, reviews, him, contain, Nice, sexually, prone, vocal, network, year, luke, Ah, button, forest, instructions, soundtrack, wears, negatively, pm, Im, builds, del, choices, take, neighborhood, monsters, greatly, health, These, behavioral, cuts, its, Great, patch, Was, bitches, cats, work, With, shorts, separation, bro, led, factory, heavily, letting, flags, mind, drop, signal, housing, its, magnon, gti, well, games, chew, free, privilege, disability,

strongly, echo, inner, escape, rejection, lack, influenced, followers, professionals, brains, partners, economic, replacing, curve, dozens, Facebook, burns, damaging, one, man, shallow, candidates, Let, mana, them, serving, qualify, because, One, were, leaning, onion, century, will, streets, realm, orange, China, life, yr, lets, brief, yoga, resentment, happy, vaping, cutting, invisible, replacement, refund, catch, root, clients, crystal, Haha, chin, dads, route, Most, hood, O, thoughts, damage, developing, know, voting, know, friend, universe, proud, central, As, fade, tires, introduce, remote, itll, watches, minds, writer, fire, Well, spouse, vision, play, off, frequent, employees, shopping, ahead, years, apparent, shoot, For, memory, leaves, of

Appendix B Manually refined query list

(Please note that this list may contain words some readers may find offensive.)

care, exams, rocks, bottles, tear, controlled, methods, inch, violates, storm, teammates, self, warrior, pregnancy, attempts, rain, amazed, rule, shit, decrease, wounds, rough, ghost, system, yourself, destiny, addiction, paths, shits, scars, courses, warren, tricks, arrest, envy, behaviors, sites, hill, transition, listening, wind, account, backs, hearts, semester, quitting, cousins, dip, favourite, losses, investing, bless, programs, youtube, bald, tests, fortnite, work, scrolling, unaware, assure, falling, bond, recovering, fund, improving, projecting, minecraft, hunting, steps, taper, alarm, consumption, offense, confuse, records, guard, episodes, riot, YouTube, liberals, crack, knee, dependent, 151, wages, spamming, draining, addicts, deaf, abuse, duty, depend, journey, supply, drain, bull, nuclear, unfortunate, habits, roommates, dying, parties, sigh, streak, financially, emotionally, chances, miracle, surrounded, profile, focusing, masculinity, recovery, overcome, road, poor, tears, smoke, cycle, guide, affecting, cs, opponents, bleed, blowing, angel, becoming, Usually, shield, instances, pattern, viewing, distant, independent, politically, tripping, survivors, Black, debt, fell, grace, officers, loneliness, drops, recording, satisfaction, saving, off, earned, quality, breaks, Oh, abusing, tricky, traumatic, friendships, nights, reminding, tower, bat, silence, daddy, brainwashed, asses, sucks, strike, chapter, analysis, shows, apps, interviews, command, slip, agency, notice, slow, meals, hours, season, too, army, baking, grad, suffer, attack, scar, casually, FUCKING, protests, improve, enough, addict, 30s, desert, dive, ruin, falls, factors, someday, unattractive, feel, doing, dare, same, shitting, fall, unlimited, treatments, wipe, lover, reflect, break, whining, sorry, studying, damned, relatives, binge, cave, isolated, rock, fox, dating, behaviour, myself, tanks, hunger, fulfilling, swim, tolerance, neglect, illusion, pray, alert, classes, foster, jealous, increased, breakdown, inclined, grandparents, isolate, scroll, pathetic, pull, deaths, incidents, connections, sitting, shed, neat, corporations, cake, God, ultra, shit, consuming, growth, 4k, disagree, imaginary, plants, twitch, spends, flesh, art, creep, pc, pulling, blow, version, nails, aware, psychologist, flawed, well, unstable, relationships, humanity, gas, friend, cells, incapable, stars, destroying, spirit, about, trash, dirty, oldest, vary, hopeless, releasing, harmless, sucking, periods, friends, rooting, memes, behavior, homes, idiots, lungs, inevitably, passive, endless, philosophy, republicans, tinder, grams, monster, massive, politicians, legacy, lessons, conditions, following, exceptions, breakup, elections, collection, wing, months, blew, rape, care, digging, driving, bubble, trap, Thanks, cry, unknown, clarification, simultaneously, advocate, off, influence, fair, continue, lifestyle, rant, Sorry, downvotes, lol, loads, weigh, surviving, drinking, stubborn, receive, bosses, precious, dreaming, work, sat, options, TV, nut, dark, loans, blocking, salt, last, victim, safely, social, automatically, acting, edge, dude, introvert, spending, depending, sit, gods, evil, deep, armed, trips, favorite, funds, fuel, neighbors, harsh, infinite, karma, software, graphics, coverage, cares, crossing, delay, faq, risk, abandon, snap, worthy, freak, thick, cow, negative, celebrating, tiny, interpretation, bleeding, corporate, creatures, passion, rehab
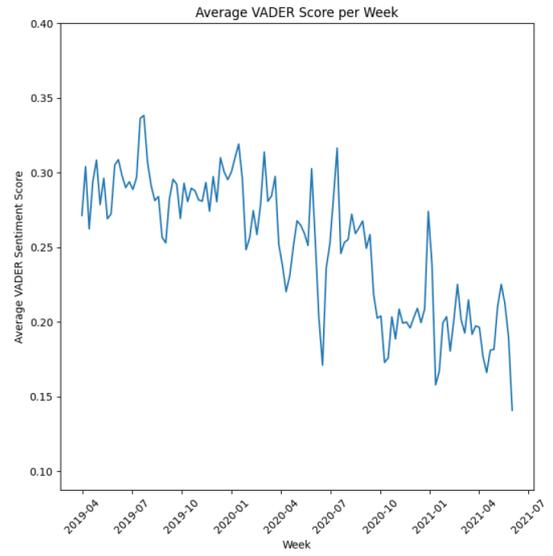
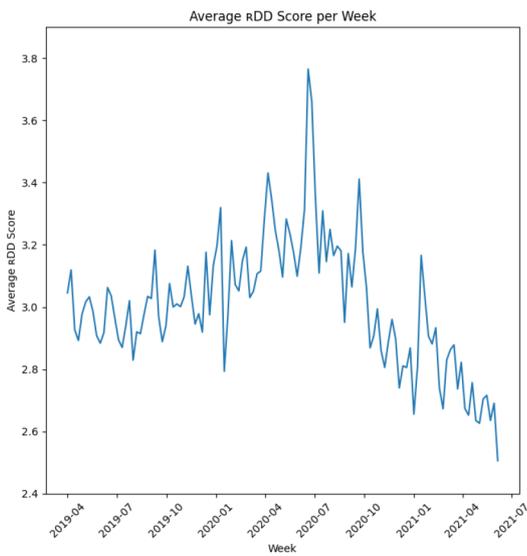# Appendix C   VADER score graphs

Figure 4: rDD and VADER Scores for Toronto and Seattle
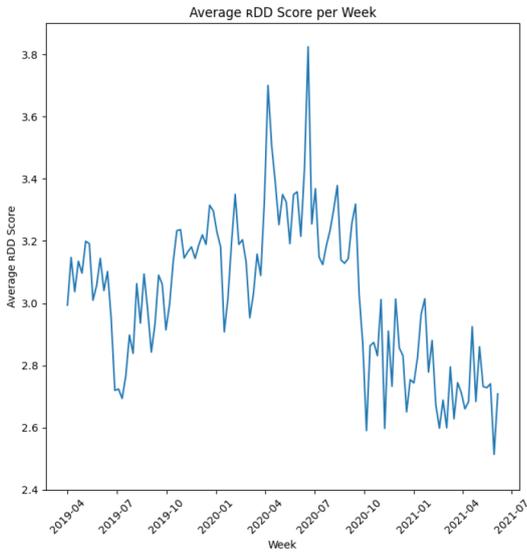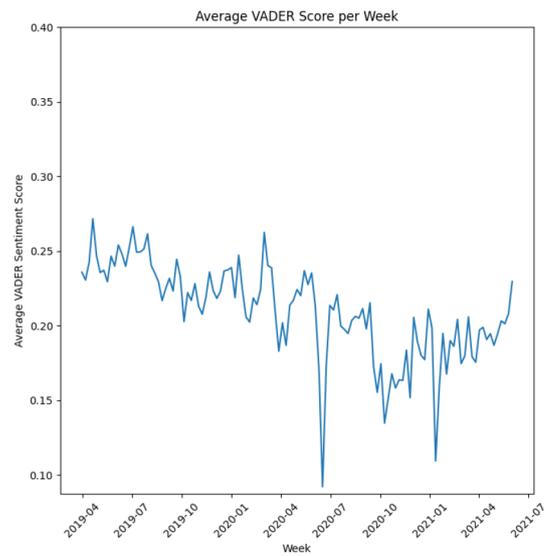

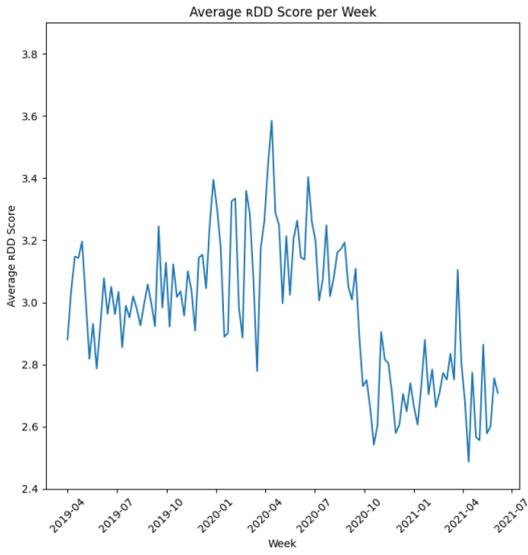(a) Toronto rDD score

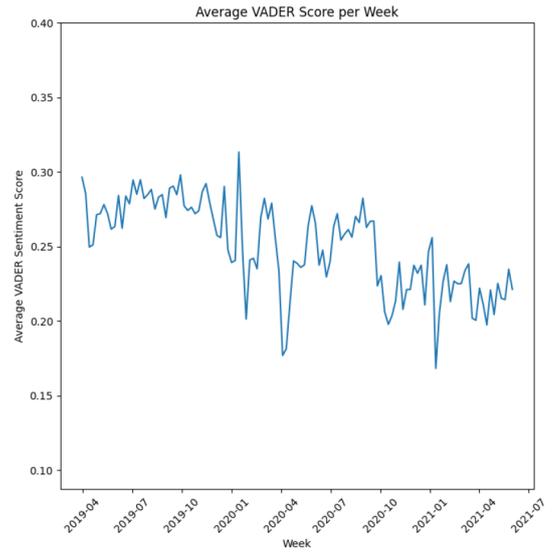(b) Toronto VADER score

(c) Seattle rDD score
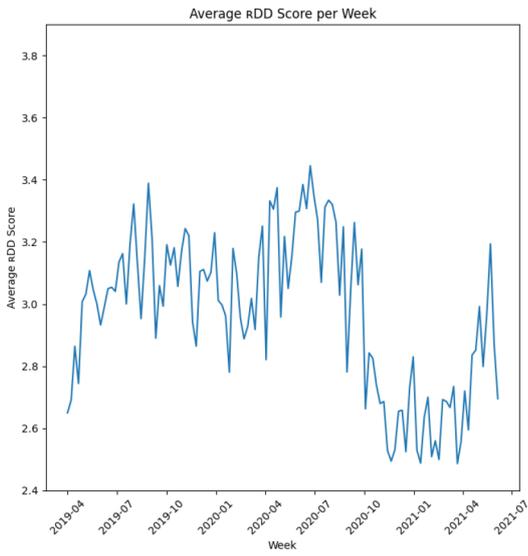

(d) Seattle VADER score

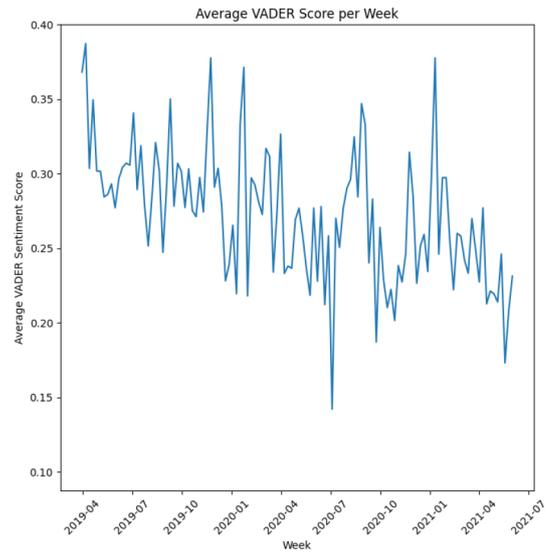Figure 4: rDD and VADER Scores for Sydney and Mumbai

(e) Sydney rDD score

(f) Sydney VADER score

(g) Mumbai rDD score

(h) Mumbai VADER score