Social Media Mining to Understand Public Mental Health

Andrew Toulis and Lukasz Golab

University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
{aptoulis,lgolab}@uwaterloo.ca

Abstract. In this paper, we apply text mining and topic modelling to understand public mental health. We focus on identifying common mental health topics across two anonymous social media platforms: Reddit and a mobile journalling/mood-tracking app. Furthermore, we analyze journals from the app to uncover relationships between topics, journal visibility (private vs. visible to other users of the app), and user-labelled sentiment. Our main findings are that 1) anxiety and depression are shared on both platforms; 2) users of the journalling app keep routine topics such as eating private, and these topics rarely appear on Reddit; and 3) sleep was a critical theme on the journalling app and had an unexpectedly negative sentiment.

Keywords: text mining, social media mining, public mental health

1 Introduction

Many applications of social media involve text mining, such as understanding user interests, customer reviews, and sentiment around news events. We discuss an application of social media text mining in the context of understanding public mental health. This is an increasingly important application domain: prevalence of mental health conditions is increasing, and so is the amount of data we have to understand these conditions [2]. While text data has been analyzed in great depth for marketing purposes, there remains a large opportunity in using text data to understand public mental health.

Several researchers have identified social media data, such as Twitter posts, as a valuable source for mental health signals [3, 6]. However, there remain critical gaps in our ability to understand mental health. People may not be willing to share mental health content publicly, especially on the largest social media platforms, which are associated with a personal identity.

In this paper, we perform a mental health analysis on stigmatized topics by taking advantage of a unique dataset from a social media app for posting journal entires and sharing and tracking moods (referred to as the journalling app). We discover issues that are not widely discussed on other social media such as sleep. In addition, the journalling app requires users to track their moods in each journal. Hence, we are equipped with user-labeled sentiment, which otherwise is difficult to estimate.

By comparing the journalling app dataset to other social media (Reddit), we identify unique discussions that the mood-tracking community attracts. Furthermore, the journalling dataset is split into two segments: users may share their journals publicly or keep them private. Hence, we are able to understand which mental health issues are shared more publicly than others.

The questions we seek to answer include:

1. Does the journalling dataset cover a different set of topics than those discussed on other social media such as Reddit?
2. Are some topics shared more publicly than others? Are some topics kept private?
3. How is sentiment related to topics? Which topics elicit sad or happy feelings?

To answer these questions, we use a text mining methodology to derive topics from journals. Furthermore, we take advantage of already labelled moods to perform sentiment analysis. To summarize, we make the following contributions:

1. We apply text mining to a unique dataset of journals and associated moods that has not been studied before. We discover a set of mental health topics that are not frequent on other social media.
2. We quantify which topics are more public than others to identify gaps in available social media data for analyzing public mental health.
3. We compare how user-labeled moods vary across topics to identify important aspects of mental health that require attention.

The remainder of this paper is organized as follows. We discuss context and related work in Section 2; we describe our datasets in Section 3; we explain our methodology in Section 4 followed by our results in Section 5; and we conclude in Section 6.

2 Context and Related Work

This paper is related to two bodies of work: social media text mining and studies of public mental health. In the context of text mining, there are standard analysis techniques that enable topic modelling and sentiment analysis of social media posts. In the mental health domain, these techniques have had several successes including detecting users expressing suicidal thoughts on social media [11]. We also use standard topic modelling techniques but we utilize them to perform novel topic comparisons between datasets.

Mental health studies traditionally collect information via health care professionals, which is a costly process and only allows for analysis of a small subset of the public. A significant opportunity for understanding mental health through social media data has been identified by Harman et al. [6]. They focused on specific mental health conditions, and despite low incident rates they found a wealth of data on social media. They concluded that individual and population-level mental health analysis can be made significantly cheaper and more efficient than current methods.

Traditional therapy and studies of mental health conditions heavily utilize linguistic signals. In Diederich et al. [5], text processing is used to detect mental health conditions such as schizophrenia by analyzing conversations between patients and their psychiatrists using clustering algorithms and sentiment classifiers. The drawback of most studies utilizing doctor-patient data is privacy concerns and smaller datasets. These studies tend to be more ad-hoc due to the size of data. Another large opportunity for mental health data analysis is in electronic medical records. For example, natural language processing (NLP) was used to improve classification accuracy of depression in mood states of patients based on medical records [13].

The creation of a social media corpus for mental health data could significantly improve mental health research [2]. There are two main methodologies for analyzing mental health signals in social media using linguistic signals. The first relies on hand-crafted lexicons containing connotations and strengths of words. For example, the Linguistic Inquiry Word Count (LIWC) lexicon has been used to help clinicians understand mental states given a patient's writings [6]. The disadvantage of this method is that lexicons like LIWC cover a very small portion of possible language used in informal contexts such as social media.

The second common method is to train a language classifier model. This technique is limited when ground-truth labels are not available. Existing work has attempted to approximate labels, and a conservative labeling approach is to filter for users who self-identify with a condition. In particular, previous work searched for statements such as "I was diagnosed with X" [2]. However, there are caveats that the authors identify with this approach. In particular, only a small sample of people would publicly self-identify with a mental health condition. Despite this, through a language model they were able to compare language uses across specific mental health conditions [2].

Alternative pipelines for acquiring labels to model social media text include crowd-sourcing and developing custom apps [2]. Crowd-sourcing involves surveying users. In a previous study, surveys were used to study mental health trends in undergraduate students [4]. In our study, we also identify school-related issues (among other things) as a frequent topic discussed by journallers. While successes with surveys have been made, having users agree to honestly share their personal information is difficult and it can be costly to solicit other data such as social media from surveyed users.

On the other hand, apps that interact with social media such as Facebook can be used to collect personality information and grant access to public status updates. However, signals that are important for mental health analysis are not typically shared on Facebook.

While existing work has focused on traditional social media, talking about difficult issues is not common on these platforms. On the other hand, the dataset we are studying is specifically designed for mood tracking. The journalling app's goal is to de-stigmatize the expression of mental health. It is fully anonymous, and hence includes topics that are typically considered taboo on personally identifiable social media platforms. Moreover, the dataset is a combination of both public and private journals, allowing for more private topics to be mentioned frequently.

Furthermore, instead of focusing on specific mental health conditions, we choose to take a broader look into the state public mental health. We demonstrate that simple, interpretable signals can be derived from our dataset. Furthermore, we use sentiment labeled by users to avoid relying on custom lexicons.

In particular, one of our most important findings is a large issue with sleep. In a study that correlated sleep problems with mental health problems, it was found that patients are much more able to identify when they have an issue with their sleep and more willing to reveal it to their doctors than a potential mental health concern [9]. Furthermore, patient self-perception of sleep issues was strongly associated with health issues, which demonstrates that people are able to accurately identify when a real problem is present. While wide-scale studies of sleep data using social media have not been performed, there is an increasing prevalence of sleep-tracking mobile apps and tools for analyzing the quality of an individual's sleep [7].

While traditional social media has helped people connect with friends and family, anonymous social media services are becoming increasingly used by people for sharing personal stories and looking for advice [14]. These communities are growing as the general public becomes comfortable sharing more information online, and benefit people who are unable or do not want to see a doctor in person to talk about mental health [14]. We believe that these types of datasets will become increasingly important to analyze for researchers. As such, we explore another anonymous social media platform, Reddit, and compare mental health topics discussed on Reddit to those written about on the journalling app.

3 Data

We analyze two datasets: 1) user communities on Reddit and 2) journals from a mental health journalling mobile app. We omit the name of the app for privacy, and we refer to it as the "journalling app".

Reddit is a social media platform that was originally used for sharing and rating content such as news, documentaries and music. Users post in and subscribe to self-organized communities known as subreddits; subscribing to a subreddit allows a user to view all posts from that subreddit. An advantage of analyzing Reddit data is that the subreddits are labelled according to their topics. Utilizing curated lists from volunteer Reddit users, we crawled all subreddits related to mental health, as well as all subreddits linked by these communities.

The second dataset consists of anonymized journal posts from a mobile app designed to help people track their moods and share them anonymously if they desire. For each journal post, the app requires the user to label the journal post with at least one mood selected from a pre-populated list including "happy", "sad", etc. We obtained all journals, and the associated moods, written between January 2016 and January 2017. This amounts to over 1.2 million journals written by approximately 75,000 users.

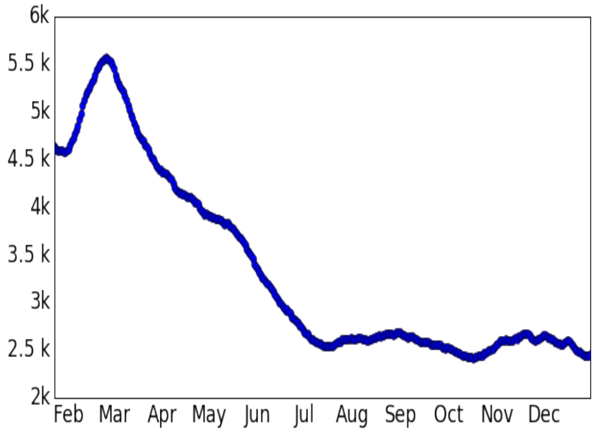

Fig. 1. Number of journals posted over time.

Figure 1 plots the number of journals posted over time. Most of the journals were written in the first half of 2016, although we inspected topic distributions per month and did not find seasonal effects. Towards the beginning of 2016, many new users registered on the app and eventually stopped using it. Like weight-loss and productivity apps, we believe this influx is tied to users looking to improve their habits as a New Year's resolution.

Each journal can be set to be private or public (visible to all other users of the app). Roughly one third of all journals are public. Figure 2 plots the number of users on the y-axis versus the percentage of journals they posted publicly. Most users are either mostly private or mostly public.

Most journals are relatively short, just like Twitter posts that are at most 140 characters. The average length of a journal with text in it is 128 characters; there are roughly 100,000 journal that have no text, only a mood label. We observed that private users tend to write journals that are slightly, but statistically significantly, longer than those written by public users by approximately 10 characters. Figure 3 shows the distribution of journal lengths, where the spikes correspond to 0 length (mood only), 200 characters (the default limit set by the app) and 300 characters (set as the maximum for visualization purposes).

Users of the app can optionally enter their location, age and gender. While most users did not enter this information, we found that those who revealed their location are mostly from North America, those who revealed their gender are predominantly female, and those who revealed their age have an average age of 25.

4 Methodology

The goal of this analysis is to understand public mental health by mining social media. We want to identify common topics discussed publicly (Reddit plus pub-

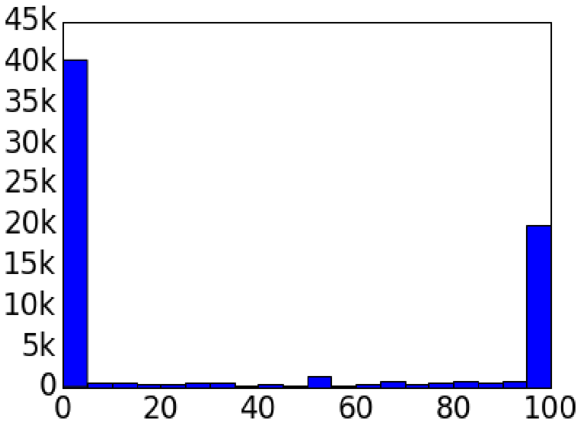

Fig. 2. Number of users plotted against the percentage of journals posted publicly.

lic journals from the journalling app) and privately (private journals). For the Reddit dataset, we simply count the number of subscribers in each subreddit related to mental health to discover popular topics and issues. Recall that each subreddit is labelled with its topic, so topic modelling is not necessary. On the other hand, for the journalling app, each journal post is labelled with a mood but not with a topic. Below, we describe our methodology for assigning topics to journals.

4.1 Topic Modelling

First, we removed journals with no text and those with fewer than 20 characters[1], leaving 1.1 million journals for topic modelling.

Next, we pre-processed the text using the Stanford Tweet Tokenizer, which is a "Twitter-aware" tokenizer designed to handle short, informal text [1]. We used the option that truncates characters repeating 3 or more times, converting phrases such as "I'm sooooo happyy" to "I'm soo happyy". On average, the number of tokens per journal was 27.7.

Since we are interested in topics, we removed stopwords and tokens with fewer than two letters, and we only retained nouns which appear in the WordNet corpus [10]. After this filtering, the average number of nouns per journal was 7.

Examples of frequently appearing nouns, in alphabetical order, include "anxiety", "class", "dinner", "family", "god", "job", "lunch", "miss", "school", "sick", "sleep", and "work". We then iteratively clustered the journals into topics (details below) and removed nouns that do not refer to topics such as numbers,

[1] We manually inspected a sample of short public journals and found that those under 20 characters long typically re-stated the mood of the user and did not refer to any specific topic.

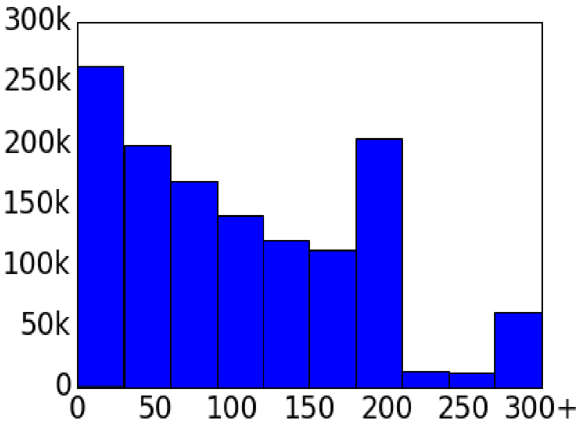

Fig. 3. Distribution of length of journals, with zero length journals included.

ALGORITHM 1: Text mining procedure to label journals with topics.

```
Data: Term frequency vector for each journal computed using TF-IDF
Result: A labeled set of journals
1. Perform clustering, iterating over different numbers of clusters
2. Manually label topics and evaluate them, selecting the best performing number
   of clusters
3. Determine a minimum threshold for the relevance of each topic in order to prune
   weak topic associations
4. Select the top two most relevant topic labels per journal, if any
```

timings (e.g., “today”, “yesterday”), general feelings (e.g., “feel”, “like”), proper nouns, and nouns that have ambiguous meanings (e.g, “overall”, “true”). Lastly, we only retained nouns that appeared more than ten times in the dataset. This process resulted in a vocabulary of 8386 words for topic modelling. Each journal is represented as a 8386-dimensional term frequency vector, with each component denoting the term-frequency/ inverse-document-frequency (TF-IDF) of the corresponding term.

Algorithm 1 summarizes our topic modelling methodology. Given a TF-IDF term frequency vector for each journal, we run non-negative matrix factorization (NMF) [8], implemented in Python’s scikit-learn package [12]. The objective of NMF is to find two matrices whose product approximates the original matrix. In our case, one matrix is the weighted set of topics in each journal, and the other is the weighted set of words that belong to each topic. Hence, each journal is represented as a combination of topics which are themselves composed of a weighted combination of words.

We chose NMF because its non-negativity constraint aids with interpretability. In the context of analyzing word frequencies, negative presence of a word

Table 1. Final list of journal topics, with the top 6 words per topic shown. We manually assigned the topic names based on the top words.

Topic						
Work	work	focus	money	meeting	friday	shift
Love	love	heart	man	fall	world	matter
School	school	high	break	summer	test	boring
Sleep	sleep	wake	sleeping	headache	nap	waking
Sickness	sick	stomach	cold	headache	throat	gross
Missing Someone	miss	old	baby	text	heart	times
Family	family	christmas	spending	health	food	husband
Career & Finances	job	interview	money	call	move	offer
Dinner	dinner	ate	movie	evening	shopping	walk
Physical Pain	pain	headache	period	empty	body	stomach
Homework	homework	finish	test	due	break	room
Anxiety/Depression	anxiety	depression	attack	high	stress	panic
School (Activities)	class	test	yoga	dance	english	teacher
Meals	lunch	ate	food	eating	break	breakfast

would not be interpretable. This is because we only track word occurrences and not semantics or syntax. Unlike other matrix factorization methods, NMF reconstructs each document from a sum of positive parts, which enables us to easily manually label the discovered topics.

Iterating from 4 to 40 topics, we derived 37 different topic matrices (steps 1 and 2 of Algorithm 1). Each matrix consists of one topic per row. Each topic has a positive weight for each word in the vocabulary. Stronger weights indicate higher relevance to the topic. The final topic matrix we used has 14 topics and is shown in Table 1. We show the first six words in this table for simplicity, where we sorted the words associated with each topic from highest relevance to lowest.

When judging the topic matrices, we considered the top twenty most important words per topic. Using this information, we manually labeled each row in the matrix with a corresponding topic. Furthermore, we manually evaluated each matrix based on the distinctness between topics, consistency within topics, and interpretability. During this process, we compiled a custom list of removed words that we mentioned earlier in this section. The groups of words we removed appeared as stand-alone topics that did not offer information about what the journal was about. For example, proper nouns appeared as a stand-alone topic. Other words, which we deemed too general or ambiguous, appeared across several topics and hence did not provide discriminative information.

We note that by default NMF does not enforce words to be assigned a nonzero weight to only a single topic. Using our pruning procedure, we ensured words that appeared across too many topics were removed. We did permit words with multiple meanings (for example, “high”) and words that apply in different settings (for example yoga “class” versus academic “class”). We note that the most important words (based on weights) for each topic generally did not overlap,

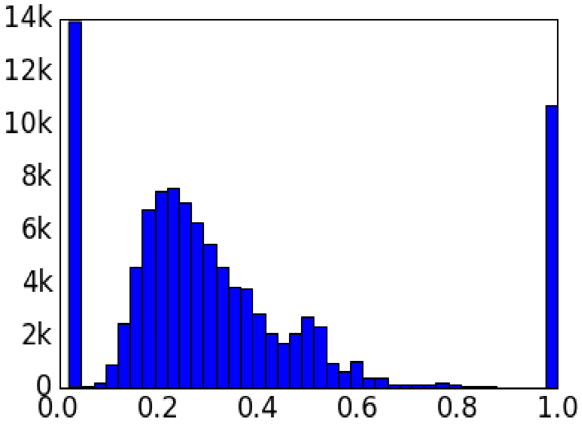

Fig. 4. Distribution of the importance of the topic ”Work” to journals, with an inflection point at 10%. Zero importance has been omitted for clarity.

with “ate” being the exception. Our validation procedure, outlined in section 4.2, ensured that the two topics ’Dinner’ and “Meals” were indeed distinct despite both assigning high weights to “ate”.

We tested different levels of regularization to enforce sparseness in our models (see [8] for a discussion), but did not find significant differences. However, one important modification we made to regularize each topic was to make their first words only as strong as their second ones (by default, first words are stronger than second words, which are stronger than third words, and so on). This is since the most relevant word for each topic tended to be too strong of a signal, regardless of how we changed the number of topics, pre-processing procedure, or regularization in the objective function. For example, the word “love” in a journal about sports would be so strong that the journal would be labeled as relating to romantic love. Lowering the importance of first words was sufficient to eliminate the false positives we identified.

Given the final topic matrix (summarized in Table 1), the next step is to use it to assign labels to journals (steps 3 and 4 of Algorithm 1). We plotted the distribution of how important each topic was to all journals in the dataset, with importance ranging from zero to one. Each distribution had a similar shape with a clear inflection point between 0.05 to 0.15 importance. Figure 4 shows an example importance distribution for the topic ”Work”, where the inflection point occurs at 0.1 importance.

We used these inflection points to set minimum thresholds of importance for each topic. We ignored any topic assignments below the thresholds. Then, we obtained the top two topics per journal, if any. We chose a maximum of two topics per journal due to the generally short length of journals.

Table 2. Accuracy of our topic model per topic.

Topic	Accuracy (%)
Work	88
Love	63
School	98
Sleep	72
Sickness	83
Missing Someone	88
Family	83
Career & Finances	88
Dinner	93
Physical Pain	75
Homework	98
Anxiety / Depression	98
School (Activities)	85
Meals	83

Using this topic modelling procedure, we assigned at least one topic to 430,000 (35 percent of all available) journals. The number of journals with one topic was 334,000, while 96,000 had two topics (the maximum).

4.2 Validation

To evaluate the effectiveness of our topic modelling methodology, we selected random subsets of 60 public journals for each topic and 100 public journals with no assigned topic. We manually labeled the sampled journals, looking for one of the 14 available topics, no topic or "other" topic. Including "other" allowed us to validate whether our list of manually labeled topic names were accurate and complete. We then compared our labels with those assigned by the model. For journals which were assigned two topics by the model, we considered the model correct if either one of the assigned topics was equal to the topic we chose manually. Table 2 shows the topic accuracies of our model. Overall, our model works well, with an average accuracy well above 80 percent.

Journals without topics were much shorter in length. The average journal length of a journal with no topic was 114 characters, while one topic was 142 and two topics was 185. Manual inspection confirmed that these journals indeed did not contain any topic more than 70 percent of the time. Instead, they mostly contained sentiment that was already available from mood labels.

We conclude this section by remarking that we analyzed activity around significant events such as the 2016 American Election. We did not find statistically significant anomalies in topics mentioned since the topics we derived are mostly related to day-to-day activities.

Table 3. Mental health related communities on Reddit with the most subscribers.

Subreddit	Subscribers
Depression	174k
Anxiety	110k
ADHD	70k
Suicide Watch	53k
Stop Smoking	46k
Mental Health	26k
Aspergers	22k
Dating	21k
Career Guidance	21k
BPD	17k
Bipolar Reddit	16k
OCD	12k
Sleep	12k
Eating Disorders	9k
Insomnia	9k
Alcoholism	8k
High School	4k
Family	2.5k

5 Results

This section presents the results of our analysis. As a reminder, the input consists of: 1) the number of Reddit users subscribed to various mental-health-related subreddits and 2) journals from the journalling app, each labeled with a timestamp, mood (entered by the user), visibility (public vs. private; set by the user), and up to two topics (assigned by our topic modelling algorithm).

5.1 Frequent Mental Health Topics across Reddit, Public Journalling and Private Journalling

We begin by comparing commonly subscribed topics on Reddit to common topics discussed on the journalling app both privately and publicly. Table 3 shows the most subscribed mental health related subreddits. Table 4 lists various statistics for the 14 topics we identified in the journalling app, including:

- Happiness percentage, corresponding to the percentage of journals whose associated mood was “happy”.
- Publicness percentage, corresponding to the percentage of journals whose visibility was set to public.
- Number of journals per topic (1000s).
- Number of users who posted at least one journal on the given topic (1000s).
- Average journal length per topic.

Table 4. Topic statistics for the journalling app.

Topic	Happiness (%)	Publicness (%)	Journals (1000s)	Users (1000s)	Average Length
Dinner	83	21	20	7	143
Meals	80	26	8	4	100
School (Activities)	68	30	16	6	136
Work	64	31	86	21	147
Sickness	36	32	23	9	118
School	58	32	39	12	140
Homework	63	32	11	4	124
Physical Pain	38	33	16	7	124
Family	66	34	23	10	155
Missing Someone	43	35	18	8	142
Sleep	43	36	35	14	135
Career & Finances	57	37	19	8	142
Love	70	38	58	17	154
Anxiety / Depression	38	42	17	7	147

While scanning for health-related communities on Reddit, we immediately noticed that physical health (exercise, weight loss) is a much larger theme compared to the journalling dataset. This is likely since there are other apps for tracking exercise. On the other hand, communities focused on mental health were relatively small given Reddit's large user base. Self-identified depression was the largest subreddit focused on a mental health condition, which in the journalling dataset was also a common topic. Additionally, Reddit includes smaller communities, such as ”High School”, ”Sleep” and ”Family” that correspond to important topics found in the journalling dataset. Notably, people with ADHD formed a very large community on Reddit, which was not a major theme in the journalling dataset and which could be a unique dataset for researchers interested in ADHD.

On Reddit, sleep-related communities are very small while in the journalling dataset it is a major theme. Sleep is a daily need that is critical to mood, which is what the journalling app is designed to track. Sleep is the third most common topic, and, as discussed later, it has a relatively negative sentiment. Based on manual inspection of a random subset of public journals, mentions of sleep are not mainly related to insomnia. Instead, we found that most mentions of sleep include commentaries on the quality of sleep, looking forward to go to sleep due to exhaustion, and (non-chronic) lack of sleep.

To further understand how users are logging their sleep, we analyzed the timing of journals that mentioned sleep. Figures 5 and 6 show the times of day that sleep and non-sleep related journals, respectively, were written across all journals posted in 2016. Sleep was uniquely mentioned in the mornings, whereas all other topics followed a very similar distribution (”Dinner” was mentioned later in the day than other topics and was removed for clarity). In agreement

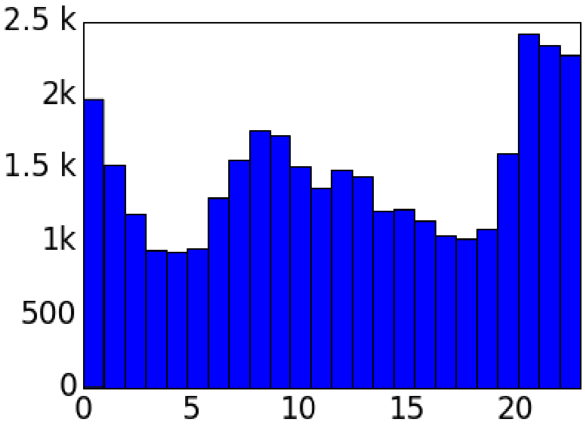

Fig. 5. Time of day when sleep was mentioned, summed across 2016.

with our manual inspection, sleep is mentioned before common hours of sleep and in the morning after waking up.

In addition, Reddit's community does not appropriately address specific issues that are affecting people in the journalling dataset, including family and school-related stress. Also, while a large subreddit exists for career advice, it does not specifically target job-related stress and workplace conflicts that are mentioned in the journalling dataset.

5.2 Analysis of Journal Moods and Visibility

Overall, one third of all journals are public. Based on Table 4, we find that social media has a gap in its ability to fulfill our social needs when expressing day to day activities. In particular, "Dinner" and "Meals" are topics that are shared (set to public) less than 30 percent of the time. Based on manual inspection of a random sample of public journals, those labelled with the topic "Dinner" tend to be about dates and family gatherings. On the other hand, "Meals" are generally short journals that are used to track how much was eaten and whether it was healthy or not. By creating a private medium, the journalling app helps people reflect upon these moments.

On the other hand, more public topics which were shared 35 percent or more of the time were "Missing Someone", "Sleep", "Career & Finances", "Love" and "Anxiety / Depression". Anxiety and depression are talked about the most publicly, which shows that users are aware of and comfortable sharing their mental state on the journalling app. In comparison, these topics are not usually found on traditional social media since there is a stigma around them.

Table 4 also contains the average mood of each topic, as labeled by users. While most topics are generally quite happy, there are some that are unexpectedly sad. Most surprisingly, "Sleep" is just as negative as "Missing Someone",

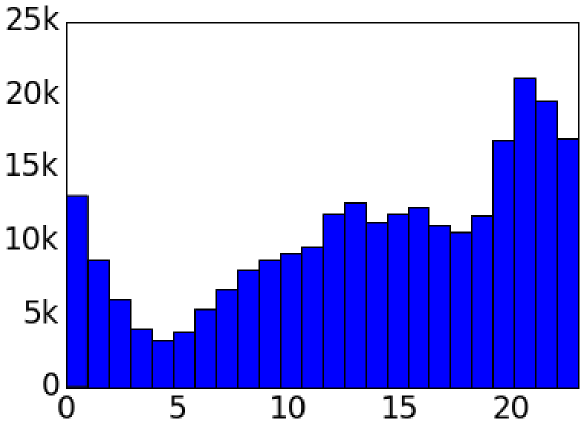

Fig. 6. Time of day when topics besides sleep and dinner were mentioned, summed across 2016.

with only 43 percent of journals happy, compared to the average happiness across the dataset of 60 percent. "Dinner" and "Meals" were especially happy, which also happened to be the most private topics.

6 Conclusions

In this paper, we used text mining to analyze a unique dataset of public and private journals in order to understand public mental health. We uncovered core themes affecting users. Based on user-labeled moods, we analyzed sentiment, revealing that the most private topics had the most positive mood. Despite being a very low mood topic, anxiety and depression were frequently publicly shared, showing the stigma around these issues can be mitigated in an anonymous environment.

By comparing public and private journals, we determined which topics are shared more than others, identifying new themes not available in currently analyzed social media. Routine topics such as eating meals are kept private by users. Across the dataset, most journals and topics were mostly private, suggesting that traditional social media cannot fulfill the need to express emotions during these moments.

We also compared the journalling app to Reddit, another service for anonymous sharing. We found that mental health topics such as family, school and work-related issues were missing from Reddit, perhaps because people are uncomfortable discussing these issues in a public forum, even anonymously. We believe there is an unfilled need for this user base. Future social media services may wish to offer a place to talk about these problems and make people comfortable enough to express emotions about them publicly.

An interesting finding is that sleep was a critical theme in the journalling dataset. This topic was frequently mentioned before and after sleeping. Sleep had an unexpected negative sentiment that is comparable with the topic of missing someone. Sleep is a daily activity that has a large impact on mood and is impacted by external factors such as stress. Hence, sleep monitoring data is critical for understanding public mental health.

In future work, we plan to collect more data to analyze issues related to sleep in more detail. For example, Twitter data offers a chance to study sleep patterns in users who post daily.

References

1. S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.
2. G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1 – 10, 2015.
3. G. Coppersmith, C. Harman, and M. Dredze. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 579 – 582, 2014.
4. M. Deziel, D. Olawo, L. Truchon, and L. Golab. Analyzing the mental health of engineering students using classification and regression. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*, pages 228 – 231, 2013.
5. J. Diederich, A. Al-Ajmi, and P. Yellowlees. Ex-ray: data mining and mental health. *Applied Soft Computing*, 7(3):923–928, 2007.
6. G. Harman, M. Coppersmith, and C. Dredze. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014.
7. H. S. Hossain, N. Roy, and M. A. A. H. Khan. Sleep well: a sound sleep monitoring framework for community scaling. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 44–53, 2015.
8. P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.
9. M. Kuppermann, D. P. Lubeck, P. D. Mazonson, D. L. Patrick, A. L. Stewart, D. P. Buesching, and S. K. Filer. Sleep problems and their correlates in a working population. *Journal of General Internal Medicine*, 10(1):25–32, 1995.
10. E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70, 2002.
11. D. D. Luxton, J. D. June, and J. T. Kinn. Technology-based suicide prevention: current applications and future directions. *Telemedicine and e-Health*, 17(1):50–54, 2011.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

13. R. Perlis, D. Iosifescu, V. Castro, S. Murphy, V. Gainer, J. Minnier, T. Cai, S. Goryachev, Q. Zeng, P. Gallagher, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological medicine*, 42(1):41–50, 2012.
14. M. White and S. M. Dorman. Receiving social support online: implications for health education. *Health education research*, 16(6):693–707, 2001.