Temporal and Spatial Patterns of Tweets

Do patterns vary for sleep disorder tweets?

Luke Larter, Trevor Seeger

Data Science 624-Winter 2020

April 19, 2020

Key Findings of the Project

We can list the key findings of this project as (3-5 key findings):

Key Findings

After careful analysis using both exploratory data visualization and supervised and unsupervised information extraction, a number of conclusions have been reached:

1. **There are apparent temporal trends in people's online interest in sleep disorders.** Sleep disorder tweets tend to dominate early in the morning, and online queries regarding sleep disorders seem to peak in the winter months.
2. **Certain words and word combinations appear at different frequencies in sleep disorder and non-sleep disorder tweets, and when sleep disorders are discussed on different social media platforms.** Words relating to sleep problems are, unsurprisingly, more common in sleep disorder tweets. Also, disorder sufferers seem to use Reddit to seek and give advice, moreso than Twitter.
3. **The language used in sleep disorder tweets is not the same as that in other tweets**. While there is an overlap in the words chosen, distinct words appear in the sleep disorder tweets that are not prevalent in other tweets.

1 Introduction

1.1 Project Motivation

- What is the problem you want to solve?

  We are seeking to find general patterns in tweets relating to sleep disorders in both space and time. This will give us insight into whether these tweets tend to cluster in certain provinces, rural/urban settings, at different times of day, and at certain times of the year. We will also assess the words that are commonly associated with sleep disorder tweets to assess what effect these have on the mental welfare of Canadians.

- What strategic goal is this connected to?

  This project is connected to the goal of establishing whether there are certain targets that would benefit most from interventions such as sleep disorder pamphlets or workshops relating to sleep hygiene. Sleep is essential for health, so the hope is that these interventions will increase the well-being of Canadians. By analyzing the words used in different classes of sleep, we will gain insight into how Canadians feel about their sleep disorders, and this may allow us to classify sleep disorder-related tweets so as to track sleep disorders and target interventions, going forward.

1.2 Problem Definition

We want to be able to identify whether tweets relating to sleep disorders cluster in meaningful ways across the provinces, between urban and rural areas, and at certain times of day and year. We also want to see which specific words and word combinations are indicative of sleep disorders.
The specific Questions posed by our clients are:

1. **What time of day do the tweets for sleep disorders most commonly appear?**
   1.1) Are sleep disorder-related tweets more common at certain times of day?
   1.2) Do Google searches regarding sleep disorders show patterns throughout the year?
2. **What kind of words are used most between these times? (How are sleep disorders described, externalized, or coped with?)**
   2.1) Are certain words associated more strongly with sleep disorder and non-sleep disorder tweets?
   2.2) Are sleep disorders expressed the same way across different social media platforms such as Reddit and Twitter?
3. **What cities show the most tweets listed under sleep disorder, and why?**
   3.1) What density distribution appears across Canada for each categorization of tweet?
   3.2) How are Google searches relating to sleep disorders distributed across Canada's provinces?
4. **What are the common themes that emerge when users tweet about sleep disorders (Insomnia, parasomnias, circadian rhythm disorders, etc)?**
   4.1) Can we classify if a tweet reflects a sleep disorder based on the language used/diction

- What **input** data do you have for your algorithms/solutions?
- For the geographic information we will need to use geographic location tags of tweets, and we will need to map these to provinces and urban hubs vs. rural areas. Temporal trends will be investigated by using the timestamps of the tweets. Finally, we will perform natural language processing on the text of the tweets to extract commonly used words and combinations of words. Additional data from Google Trends and Reddit will be Incorporated, allowing us to confirm and compare our results.

  - # of rows (roughly): ˜1500
  - temporal patterns: end user created at column as well as the behaviour classification of the tweet. And, for investigating geographic patterns, user.location will be used to pinpoint the location of the tweeter. For conducting lexigraphical analysis, NLP will be performed on the actual body of the tweet; the text column.
  - the time-span of your dataset: Tweets from 2019
  - Are there any challenges regarding this dataset? If yes, what is your plan for mitigating them?: Text analysis and using language to classify tweets will likely be the most challenging aspect as tweets are short and language processing can be tricky. By utilizing supervised and unsupervised NLP learned in class, we will hopefully be able to overcome these challenges.
- In your opinion, what are the most relevant factors (variables) in your dataset that can help to address your project's objectives?
- What is your plan for labelling the dataset?
- Each student will label roughly 800 tweets so that the whole dataset will be labelled

---

2 Methodology

2.1 Data Analysis

Methods for Addressing Question 1.1:
What time of day do the tweets for sleep disorders most commonly appear?
The first step in preparing the data for an analysis of temporal distribution of tweets relating to sleep disorders was to transform the time-stamps of the tweets into a more analysis-friendly format. For this, we used the 'lubridate' package in R. Then we plotted the number of tweets occurring in our chosen categories (sleep disorder/self-report, sleep disorder/not self-report, not sleep disorder/self report, and not sleep disorder/not self report) during each hour of the day. We plotted all categories of tweets, not just those pertaining to sleep disorders, so that any deviation from normal tweeting patterns due to sleep disorders might be recognized. Tweets had been human-labelled previously, allowing classification into these categories. We were interested in how raw numbers of tweets from different categories changed by the hour of the day, and we present this using a line graph. We were also interested in how tweeting patterns relating to sleep disorders change throughout the day as a proportion of overall tweeting activity, and we used a radar chart to represent this change in proportion by the hour.

Methods for Addressing Question 1.2:
Do Google searches regarding sleep disorders show patterns throughout the year?
To understand how google searches regarding sleep disorders were distributed throughout the course of the year, we used Google Trends. As Insomnia is the sleep disorder that most people will be familiar with, we queried Google Trends as to how searches for the terms 'insomnia' and 'sleep disorders' were distributed over 2019. Canada is a bilingual nation, so we performed combined searches for these terms as well as their French translations. We then plotted the data as a line graph, and also as a scatter plot with a line of best fit, to remove visual noise when assessing overall trends. Google trends gives normalized results of searches for your chosen keyword vs. the overall number of searches. Thus, the y axis for these plots is not the absolute number of tweets, but is a relative measure of their popularity.

Methods for Addressing Question 2:
What kind of words are used most between these times? (How are sleep disorders described, externalized,

or coped with?)
To address this question, we split the data into 6 hour groups for morning (6am to 12pm), afternoon (12pm to 6pm), evening (6pm to 12am), and late night (12am to 6pm), then plotted the most common words words in each of these times for both the sleep disorder tweets and non-sleep disorder tweets.

Methods for Addressing Question 2.1:
Are certain words associated more strongly with sleep disorder and non-sleep disorder tweets?
First, we explore the data by performing natural language processing on tweets belonging to our different categories. We then show via bar graphs which relevant words occur at which frequencies, and whether there is overlap in the words used for different categories. However, as can be seen in the subsequent section, this model performed poorly. As such, we opted for a statistical exploration of the problem. We conducted chi squared tests to see which of the most frequent words used in sleep disorders appeared at significantly different frequencies in sleep disorder and non-sleep disorder tweets.

Methods for Addressing Question 2.2:
Are sleep disorders expressed the same way across other social media platforms such as Reddit.
We performed web scraping on Reddit to compare whether users of these different social media platforms discuss sleep disorders in different ways. We extracted the 100 highest-rated posts the two most popular subreddits dealing with Sleep disorders: r/Insomnia and r/SleepDisorders. We then performed the same natural language processing methods that were used to extract keywords from our Twitter dataset on the titles and main bodies of these Reddit posts. We subsequently refined our stop words to ensure we only had words relevant to sleep disorders as our top words. We then plotted the frequency of all words with a frequency over 25, to compare these to the top words associated with sleep disorders from our Twitter dataset. Additionally, we also visualized the occurrence of frequent pairings of words via plotting frequent work pairings (>5 occurrences) as bigrams. This threshold is very liberal as our Reddit dataset is small.

Methods for Addressing Question 3:
What cities show the most tweets listed under sleep disorder, and why? (Rank by location?)
The locations were extracted from the twitter dataset, converted to factors, and any location with more than 15 tweets was plotted in a bar chart.

Methods for Addressing Question 3.1:
What density distribution appears across Canada for each categorization of tweet?
In order to show the relative density of all places across Canada, we imported the dataset into Tableau, then overlaid the counts for the labels onto a map of Canada. The locations that Tableau did not recognize were filtered out. A live-filter was enabled for the labels such that individual labels and combinations of labels can be investigated.

Methods for Addressing Question 3.2:
How are Google searches relating to sleep disorders distributed across Canada's provinces?
Using google trends, we queried the relative frequency of searches relating to sleep disorders across Canada's provinces. These included any searches using the terms 'insomnia', 'sleep disorder' or their French translations. We then plotted the relative rankings of provinces in a bar chart. Google Trends does not give you the absolute differences in numbers of tweets per province, but normalizes the data to give relative rankings among provinces. Also, as the metric Google uses to rank provinces is the frequency of searches containing the terms of interest against the overall volume of Google searches in that province, this effectively controls for differences in population size among provinces. Google Trends doesn't allow an appraisal at the level of cities, so we were not able to delve deeper than the provincial level.

Methods for Addressing Question 4:
What are the common themes that emerge when users tweet about sleep disorders (Insomnia, parasomnias, circadian rhythm disorders, etc)?
We constructed bigrams to visualize the how commonly different words occur together in sleep disorder and non-sleep disorder tweets. For these bigrams, overly common words were removed, and word pairings had to exceed a frequency of 15 occurrences to be plotted.

Methods for Addressing Question 4.1:
Can we classify if a tweet reflects a sleep disorder based on the language used/diction?

We then decided to try to answer our question with a more quantitative approach. We opted to build a machine learning model to assess whether we could use the words used in a tweet to predict whether it was a tweet pertaining to a sleep disorder. We opted for an logistic regression approach via an XGBoost model, as we had a binary target variable. We the used NLP in R (including packages tm, snowballC, and igraph) to determine the top 28 most common words used in tweets pertaining to sleep disorders. We went through several iterations of stop word removal before we were satisfied that our top words were informative. We then modified our data frame so that each of these words represented a separate feature for out set of tweets (which contained both sleep disorder and non-sleep disorder tweets, with self- and second hand-report collapsed into a single label), with a 1 indicating that word was present, and a 0 indicating absence. After processing and removing NaNs, we were left with 1066 tweets. We then split the data 80/20 for training/testing respectively, and assessed how well our model's prediction accuracy on our test data. We used a gridsearch to analyze the optimal combination of hyperparameters for our model; these can be seen below.

```
#Fit final XGBoost model with best hyperparameters

final_XGB=xgb.XGBClassifier(objective='binary:logistic', learning_rate=0.1, max_depth=5,
                            n_estimators=100, reg_alpha=0, reg_lambda=0.3, scale_pos_weight=0.492)

final_XGB.fit(X_train, y_train)
y_pred = final_XGB.predict(X_test)
y_prob = final_XGB.predict_proba(X_test)
```

Figure 1: *Final XGBoost Model*.

---

3 Performance Measurements

The area under the curve (AUC) is used in the logistic regression to measure performance. It is a measure of the area underneath the receiver operating curve, plotting the true positive rate and false positive rate. This areas describes the ability of the model to determine between the classes we are investigating: in this case, sleep disorder tweets or non-sleep disorder tweets. They provide a value between 0 and 1. The line drawn with a slope of 1 indicates a 50/50 probability of correctly classifying the tweet.

---

4 Results

4.1 Information Extraction

The labelled tweets were downloaded from google drive and joined to the twitter dataset using the tweet ids. The columns related to locations, timestamps, tweet contents and hashtags were selected, giving a dataset of 3245 tweets, over 18 variables. We then checked the tweet ids to ensure there were no duplicated tweets.

4.2 Exploratory Visualization

Results for Question 1:
What Time of Day do Tweets Discussing Sleep Disorders Usually Appear?

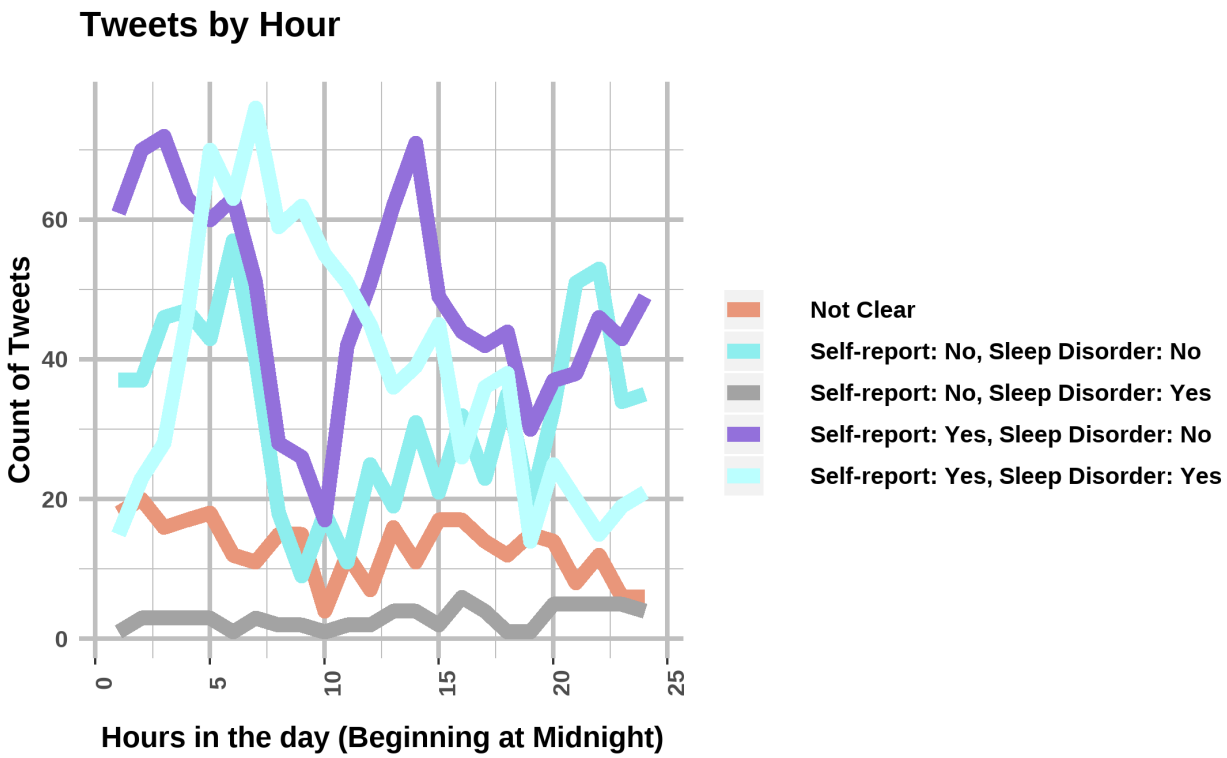


Figure 2: *Plots of the frequency distributions of different types of tweets throughout the day.*

Discussions of findings from Figure 2 to 4:

As can be seen in Figures 2, 3, and 4, tweets containing self-reports of sleep disorders can happen at all times of day, but the bulk of their occurrences happens from around 4:00am to 11:00am. This seems to suggest that people are most conscious of their sleep disorders, or a lack of sleep, when waking up in the morning, and in the earlier parts of their day. In contrast to this, we see that sleep disorders being reported second-hand seem to spike between 3:00 and 4:00pm, and then show a sustained rise between 7:00 and 11:00pm. This increase between 7:00 and 11:00pm could indicate that people are increasingly aware of the poor sleep habits of people around them later in the evening. When looking at the proportion of sleep disorder tweets as a proportion of tweets from all categories (figure 3), we see a similar trend to the one described in the line graph; sleep disorder tweets begin to be more common at around 4am, and come to dominate the tweets from 7am to 11am.

Interestingly, first-hand and second-hand reports of activities unrelated to sleep disorders seem to follow a somewhat similar pattern in terms of their frequency throughout the day (Best shown in Figure 2). They show a roughly bimodal distribution, both peaking at around 3:00 to 4:00am, and then showing another peak later in the day (at around 1:00pm for self-reported tweets, and at around 11:00pm for second-hand reported tweets). That sleep disorder tweets, both self-reported and second-hand reported, deviate from

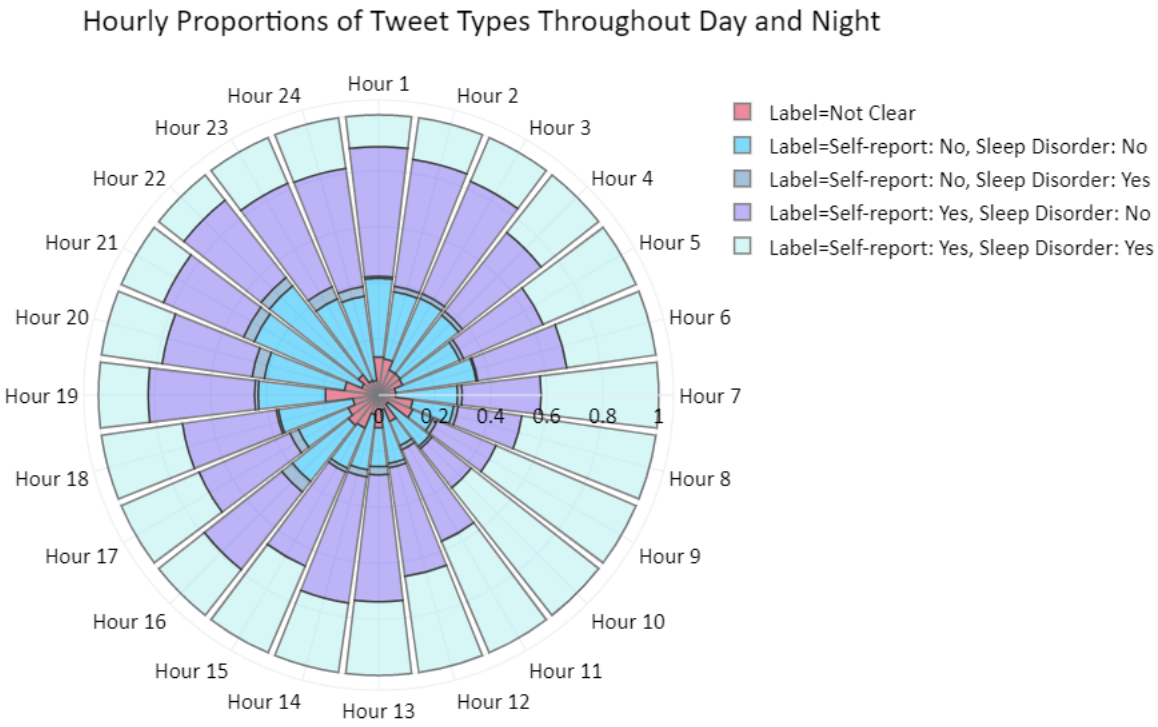


Figure 3: *Changes in proportion of tweets belonging to different categories throughout the day*.

this bimodal pattern seen in tweets unrelated to sleep disorders, suggests that the experience of people's own, or other peoples', sleep disorders are causing changes in their tweeting behaviour.

Results for Addressing Question 1.2:
What kind of Google searches have been performed over time regarding sleep disorders over time?

Moving a wider temporal lens, we can see the relative frequencies of Google trends searches in all of 2019 for insomnia and sleep disorders (figures 5 and 6). Spring and Summer months have the lowest prevalence of searches relating to sleep disorders, while the beginning and end of the year, which are winter months, show higher numbers of searches for sleep disorder-related terms.

Results for Addressing Question 2:
What kind of tweets/hashtags appear the most between these times? (How are sleep disorders described, externalized, or coped with?)

In these tweets, which do not pertain to sleep disorders, show a reasonable distribution of words throughout the day, and there appears to be several mentions of sleep related words no matter what time of day. For example, the mornings in figure 7a contain 'dream', 'bed', 'woke', but "bed" also occurs in the afternoon and late night tweets, and dream is also in the late night tweets. This gives us a baseline for how often people talk about sleep-related activities in general for this dataset. It appears that, even when not discussing sleep disorders, mentions of sleep-related words is fairly common.

In the sleep disorder tweets in figure 8a to 8d, there is a limited variety of words in the evenings but, the

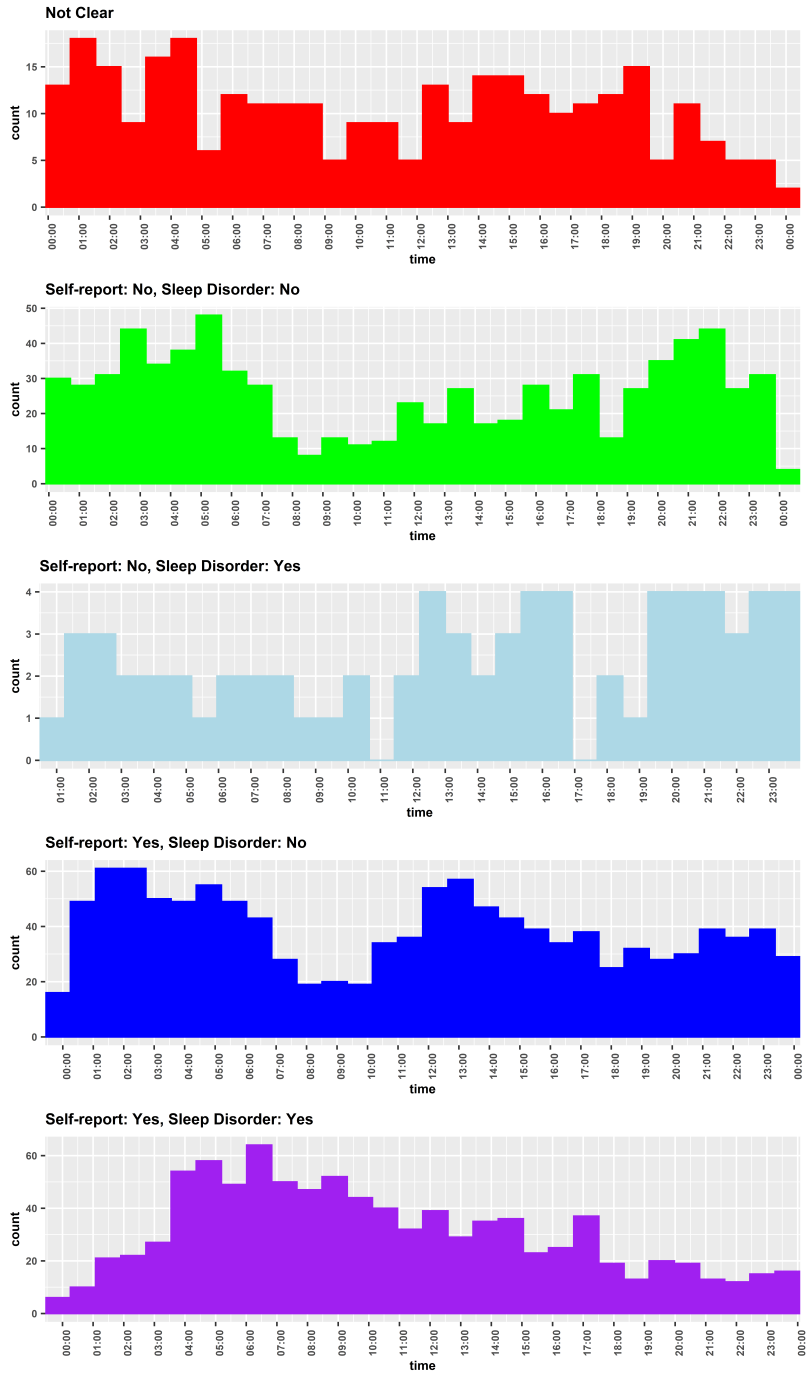


Figure 4: *Isolating daily frequency distributions of different types of tweets throughout the day.*

top two words are "asleep" and "slept". However, across all of the words used throughout the day, there do not appear to be more mentions of definitively sleep-related words in any given timeframe than in the tweets that do not pertain to sleep disorders. And interesting finding is that "insomnia" appears in only one top words graph, and it is for the afternoons, figure 8b.

Results for Addressing Question 2.1:
Are certain words associated more strongly with sleep disorder and non-sleep disorder tweets?

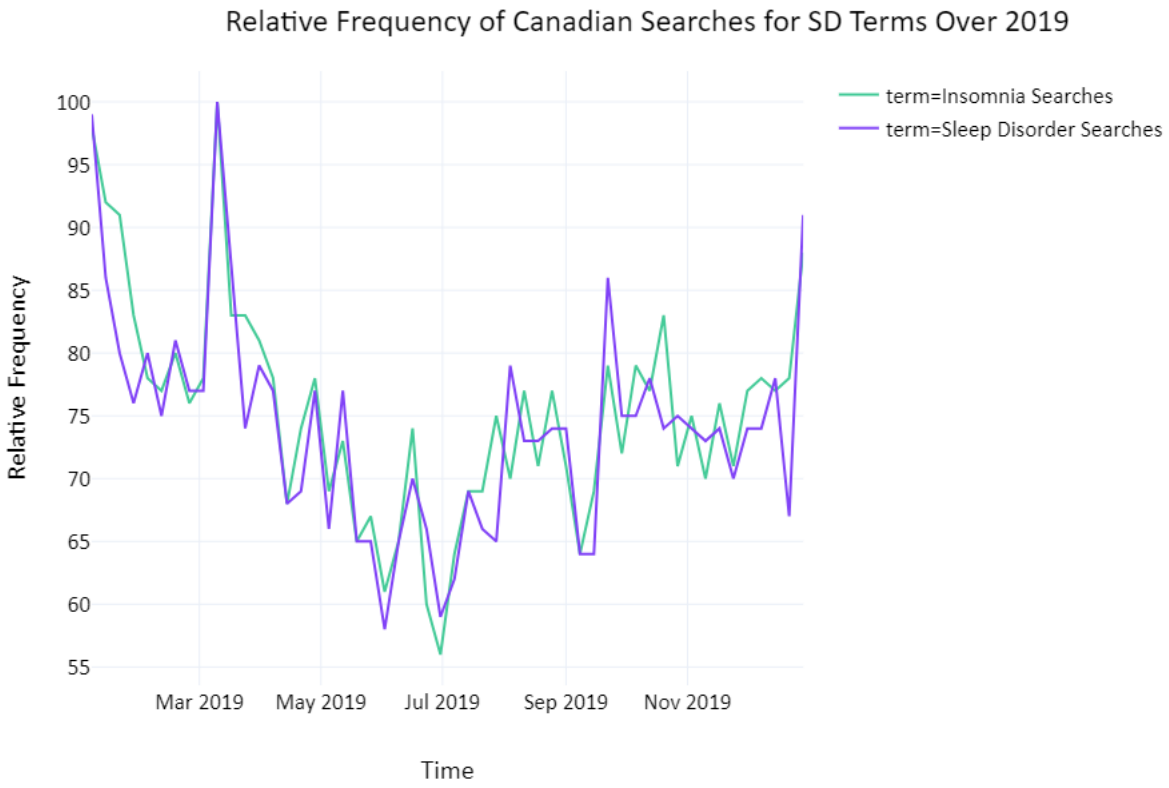


Figure 5: *Google Trends searches regarding sleep disorders throughout 2019*.

In an attempt to further corroborate the type of language used when posting about sleep disorders, we removed the time groupings from the tweets about sleep disorders or non-sleep disorders, then calculated the frequency of each word in every tweet in the dataset, figure 9. Here, we can see that there are simply more words used in the no sleep disorder category. However, highlighted in red are words that do not appear in the other dataset. Both groups are talking about "tire", but the sleep disorder tweets are commonly using words that are not included in the non-sleep disorder tweets.

We then mapped the top words from the sleep disorder tweets onto the non-sleep disorder datasets to quantify how often these words are or are not mentioned. We can see in figure 10 that there are about the same number of words and the two most common ones are "tire" and "nap". However, when we map the no sleep disorder group's words onto the yes sleep disorder groups tweets, we see the number of words available to graph drop by more than half. Therefore, we can conclude that these two groups are using distinctive vocabulary in the tweets.

Results for Addressing Question 2.2:
Are sleep disorders expressed the same way across other social media platforms such as Reddit?

In the titles of Reddit posts (Figure 11) we see specific words related to sleep, which is unsurprising, considering the subreddits from which the posts were drawn. However, here we actually see specific references to insomnia and disorders, and even ambien, a common sleep aid. Also unlike twitter, "sleep" dominates usage, whereas it was not even seen used more than 20 times in the twitter dataset. We can also see an overlap from the twitter dataset, such as the "help" word. In the body of Reddit posts, we also see a very prevalent usage of "sleep", and mentions of insomnia and requests for help. There are also some references to work schedules, similar to the twitter data.

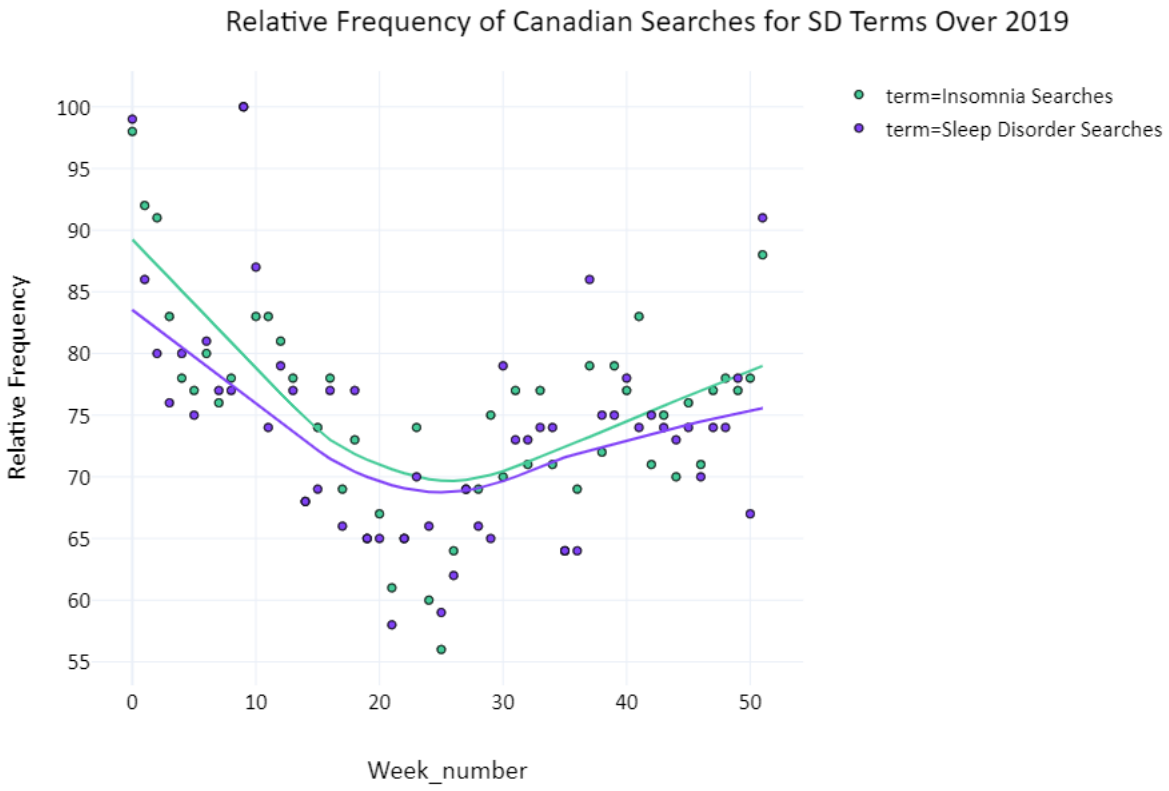


Figure 6: *Google Trends searches regarding sleep disorders by week*.

Similarly, our bigram reveals that, as well as word combinations referring to sleep disorders, pleas for help or advice are common. Word grouping such as 'anybody else', 'how do you', and 'does anyone' occur frequently. This suggests users go to Reddit to give and receive advice on their disorders.

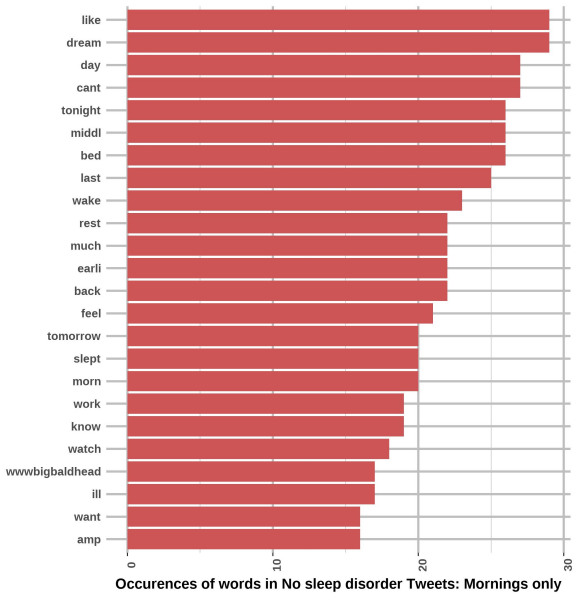
Results for Addressing Question 3:
What cities show the most tweets listed under sleep disorder, and why? (Rank by location?)

From Figure 14, we can see that Toronto has the greatest number of tweets being produced by those who have discussed sleep disorders. Edmonton, Montreal, and Calgary follow distantly behind. Surprisingly, Vancouver (the second most populace city in Canada) is 6th, behind Ottawa. However, the suburbs of Vancouver are coded differently from the dataset, therefore, this is not the Greater Vancouver Area, but Vancouver proper, which is not as populated.

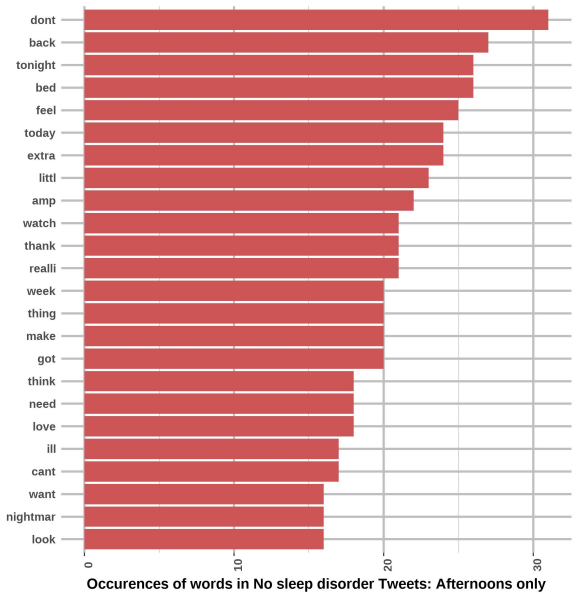
Results for Addressing Question 3.1:
What density distribution appears across Canada for each categorization of tweet?

In Figure 15, the distribution of all tweets from Toronto remains apparent, however, we see that large concentration we would expect in the Greater Vancouver Area. If we filter the dataset for only those tweets pertaining to sleep disorders (figures 16d and 16b), we lose tweets largely from the Maritimes and places further from urban centers. In fact, those tweets which were not self reports but regarding sleep disorders were almost exclusively from regions with cities.

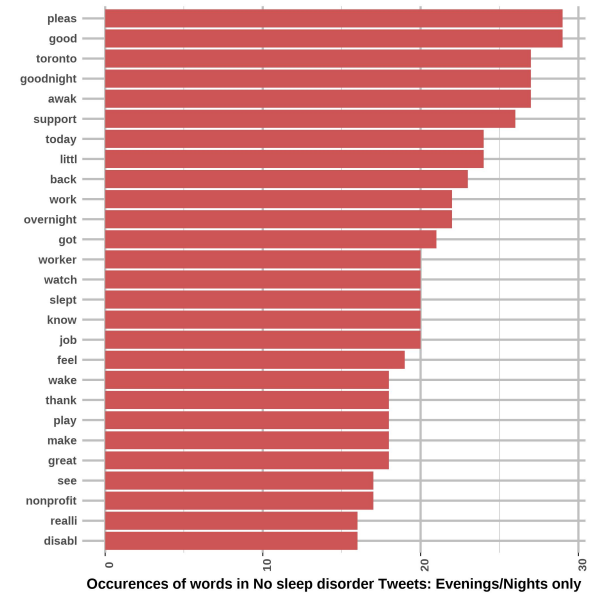
Results for Addressing Question 3.2:
How are Google searches relating to sleep disorders distributed across Canada's provinces?



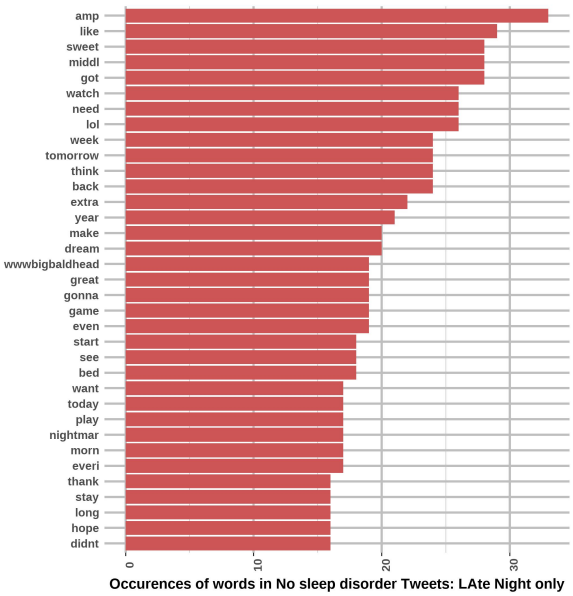
(a) Mornings; No sleep disorder



(b) Afternoon; No sleep disorder

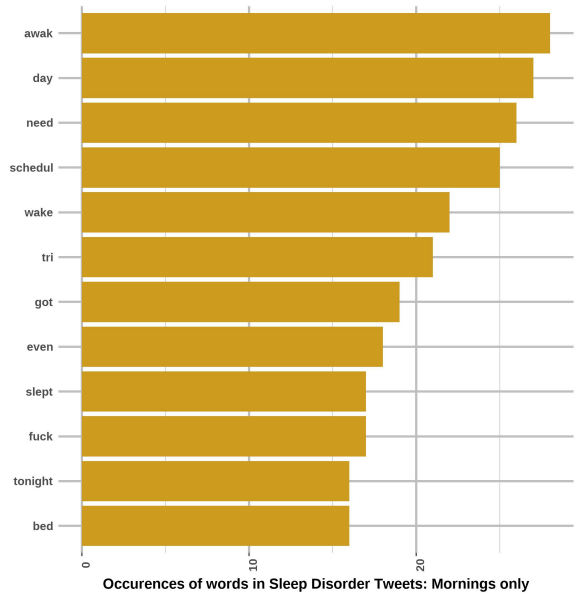


(c) Evenings; No sleep disorder

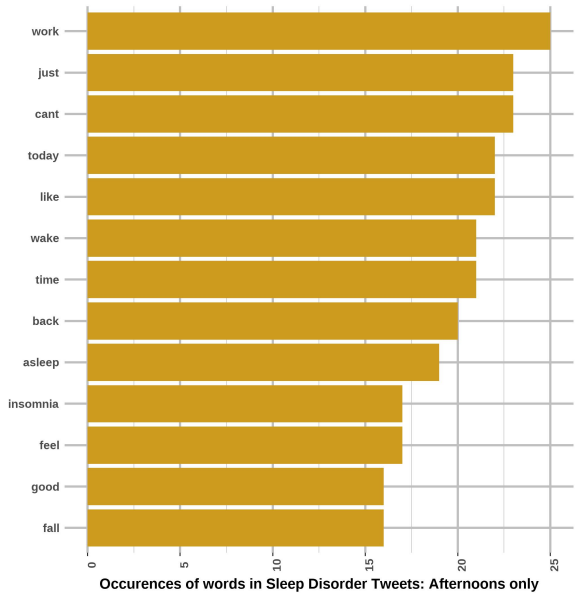


(d) Late Night; No sleep disorder

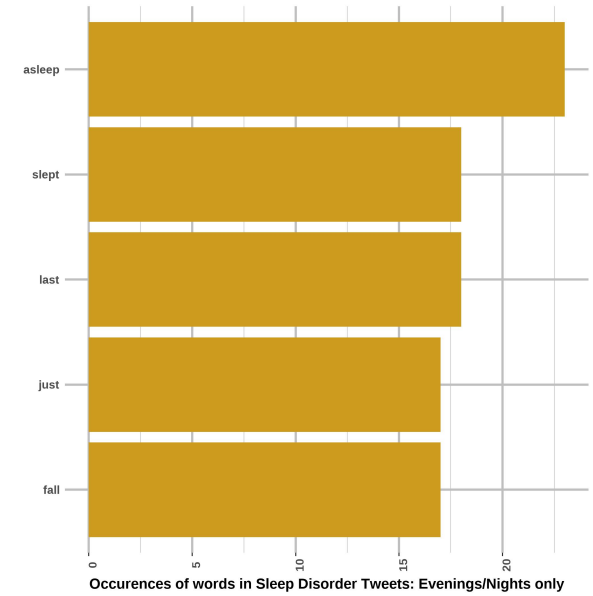
Figure 7:
Temporal relationship to words used in non-sleep disorder tweets



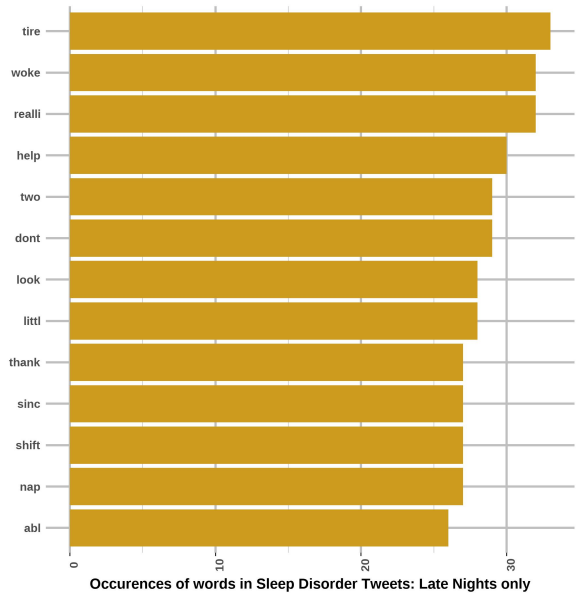
(a) Mornings; Sleep disorder



(b) Afternoon; Sleep disorder



(c) Evenings; Sleep disorder



(d) Late Night; Sleep disorder

Figure 8:
Temporal relationship to words used in sleep disorder tweets

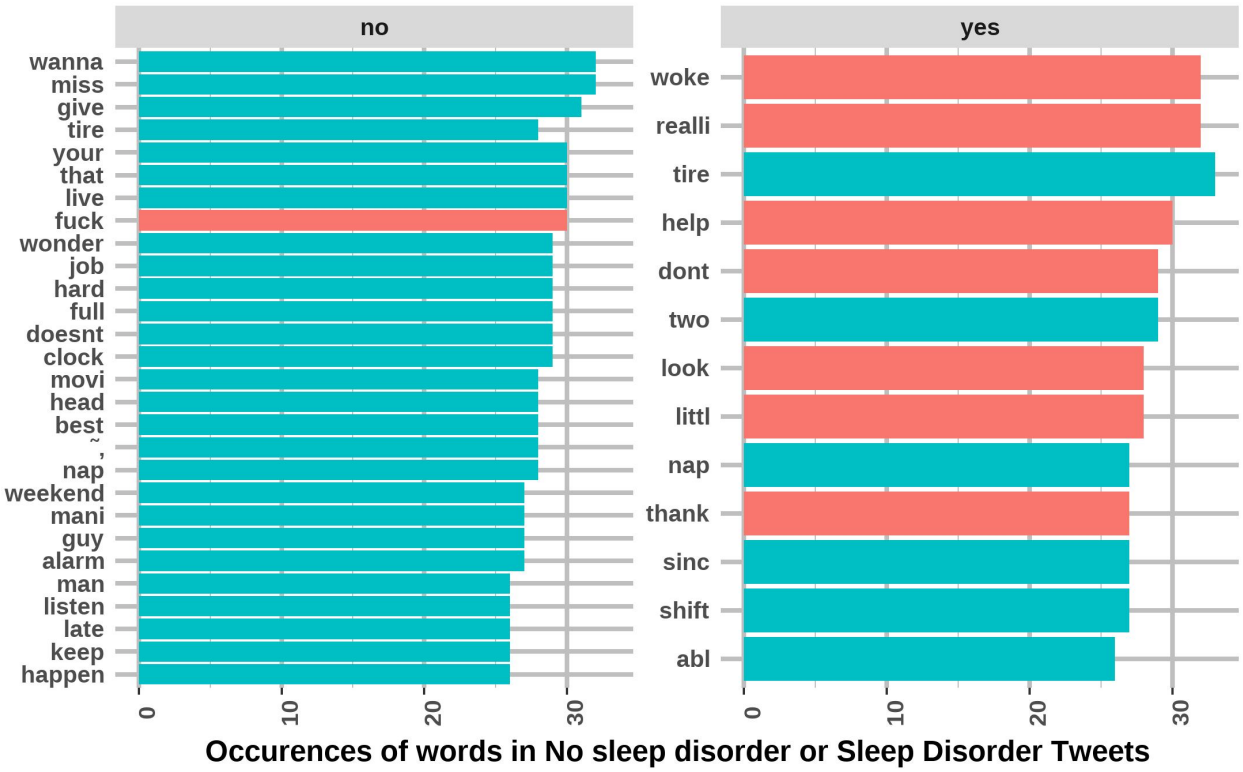


Figure 9: *Words appearing in tweets from non-sleep disorder and sleep disorder tweets.*

In Figure 17, we see that The highest volume of Google searches relating to sleep disorders come from British Columbia, followed closely by Ontario. As Canada's main urban hubs occur in these provinces (Vancouver and Toronto), this may reflect a tendency for those in big cities to be suffering from sleep disorders at a higher rate.

Results for Addressing Question 4:
What are the common themes that emerge when users tweet about sleep disorders (Insomnia, parasomnias, circadian rhythm disorders, etc)?

The AUC for our logistic regression model was 0.52. This is extremely poor, indicating that any subsequent analysis would be better served by using features which summarize the lexical content of tweets in a more complex way than we have here.

When looking at the bigrams, it is unsurprising that word combinations regarding sleep behaviour are common. Word pairings such as 'to sleep' and 'of sleep', are very common, while there appears to be a greater diversity of word pairings used in non-sleep disorder tweets. Interestingly, 'to sleep' is also the most common paining in non-sleep disorder tweets.

Results for Addressing Question 4.1:
Can we classify if a tweet reflects a sleep disorder based on the language used/diction
As can be seen from our model metrics, our model was barely able to classify tweets as pertaining to sleep disorders or not above a chance level.

Our chi-squared tests yielded more fruitful results. As can be seen in figure 21a, certain words appear at significantly different frequencies in sleep disorder tweets and other tweets. As can be seen, many of these

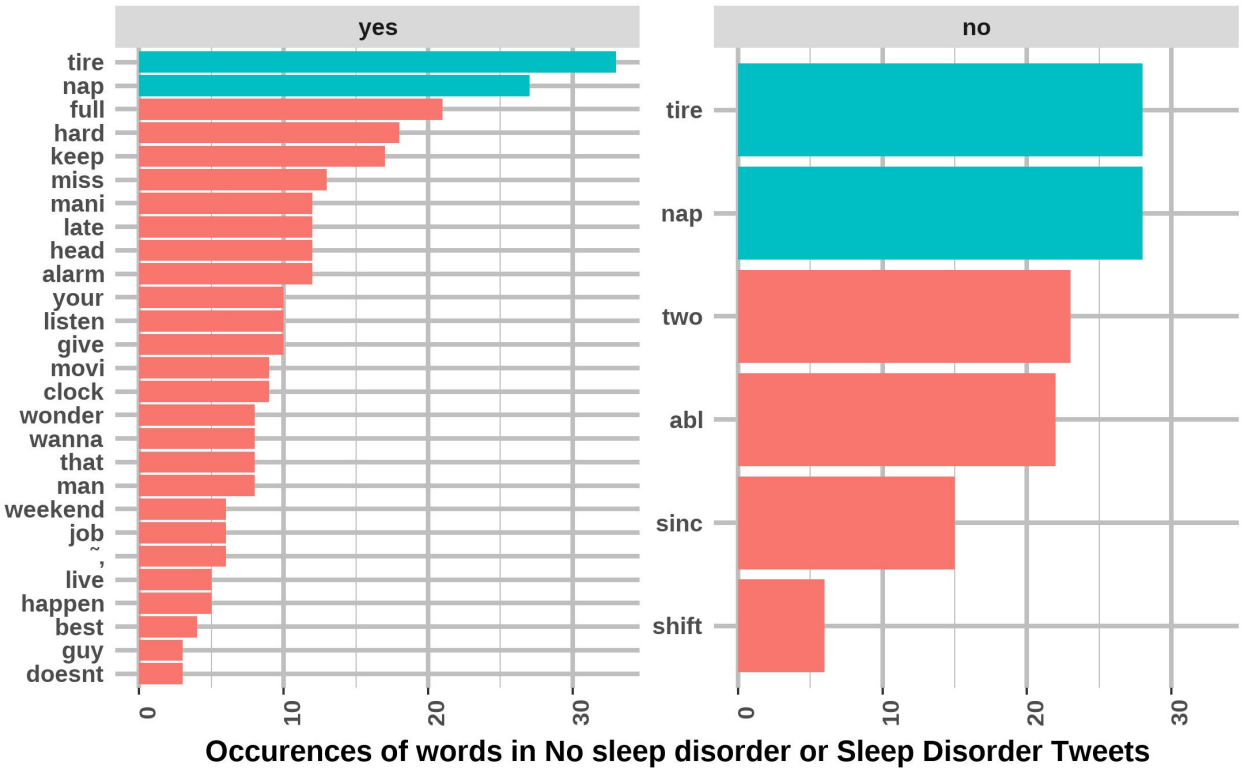


Figure 10: *Cross-referencing words appearing in tweets from the other sleep disorder group of tweets.*

relate to sleep. Interestingly, the word 'shift' occurs at significantly different frequencies, perhaps pointing to shift work as a source of sleep disturbance.

---

5 Timeline

- Are there any **deadlines** to be aware of? talk with your clients to address this question
  - Feb 5, 11:59pm: Phase 1 complete
  - Feb 26, 11:59pm: Phase 2 complete
  - Mar 18, 11:59pm: Phase 3 complete; Delayed to March 25, 11:59pm
  - Apr 1, 5:00pm Project presentations; Delayed
  - Clients have not listed any other deadlines. Will provide an approval draft at least 5 days before presentation.
- When do you need to see the first results (both visual and analytical results)? By the completion of Phase 2, tentatively
- When do you want to have a finished solution (please check the course outline to answer this question)? Within 5 days of the presentation, tentatively.

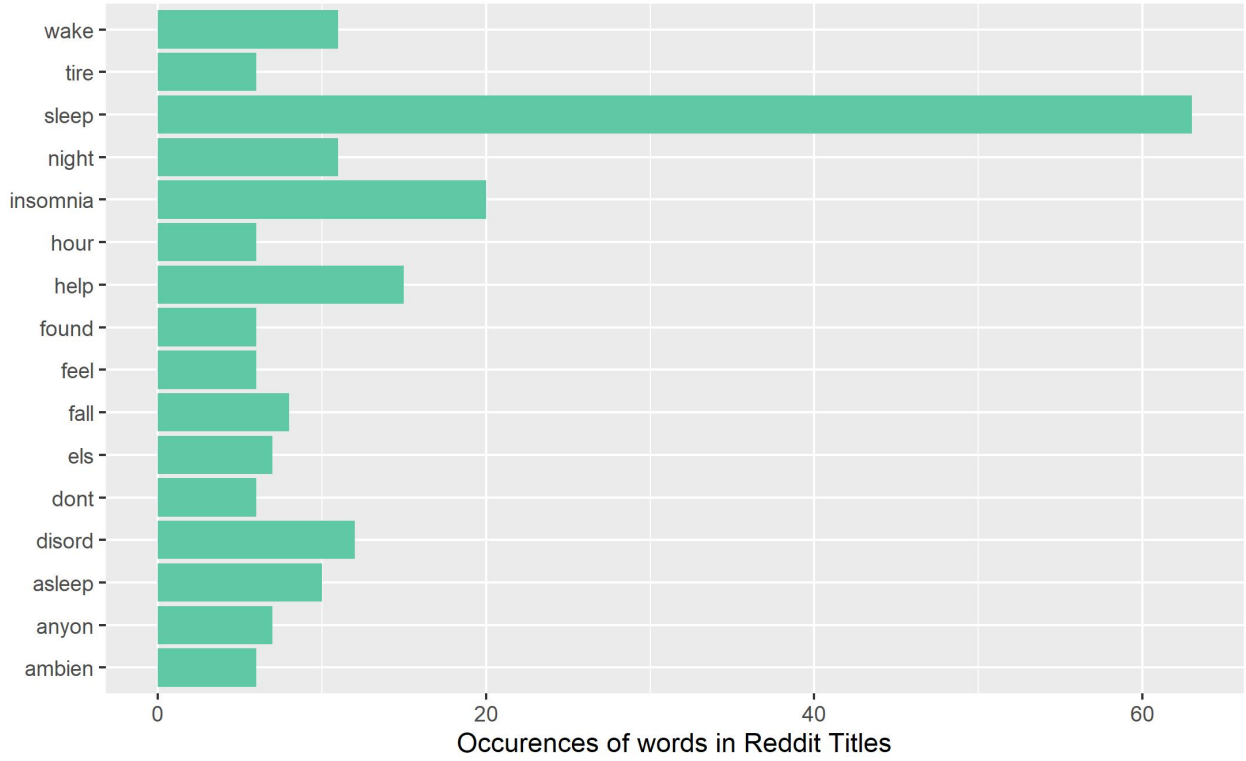


Figure 11: *Words appearing in Reddit titles*.

6 Collaboration

We will meet with Shilpa and Sid every Monday at 1pm to update them on our progress, show preliminary findings and visuals, and discuss difficulties encountered and potential ways to face them.

- Who should be involved?
  - at least one member of the client team and the development team
- What should they learn?
  - Progress on each question itemized, any issues faced and what solutions/workarounds have been found

7 Conclusion

In this section please address the key findings of your project and mentions at least two limitations of your work.

Current Conclusions: From our visualization for Question 1, it is clear that tweets regarding sleep disorders, both self report and second-hand report, show a different temporal pattern to tweets about other subjects. They seem to be most frequent around times when individuals would be waking or going to sleep, respectively. Going forward, this knowledge could be used to better predict which tweets relate to these phenomenon with further machine learning approaches. It could also inform interventions to

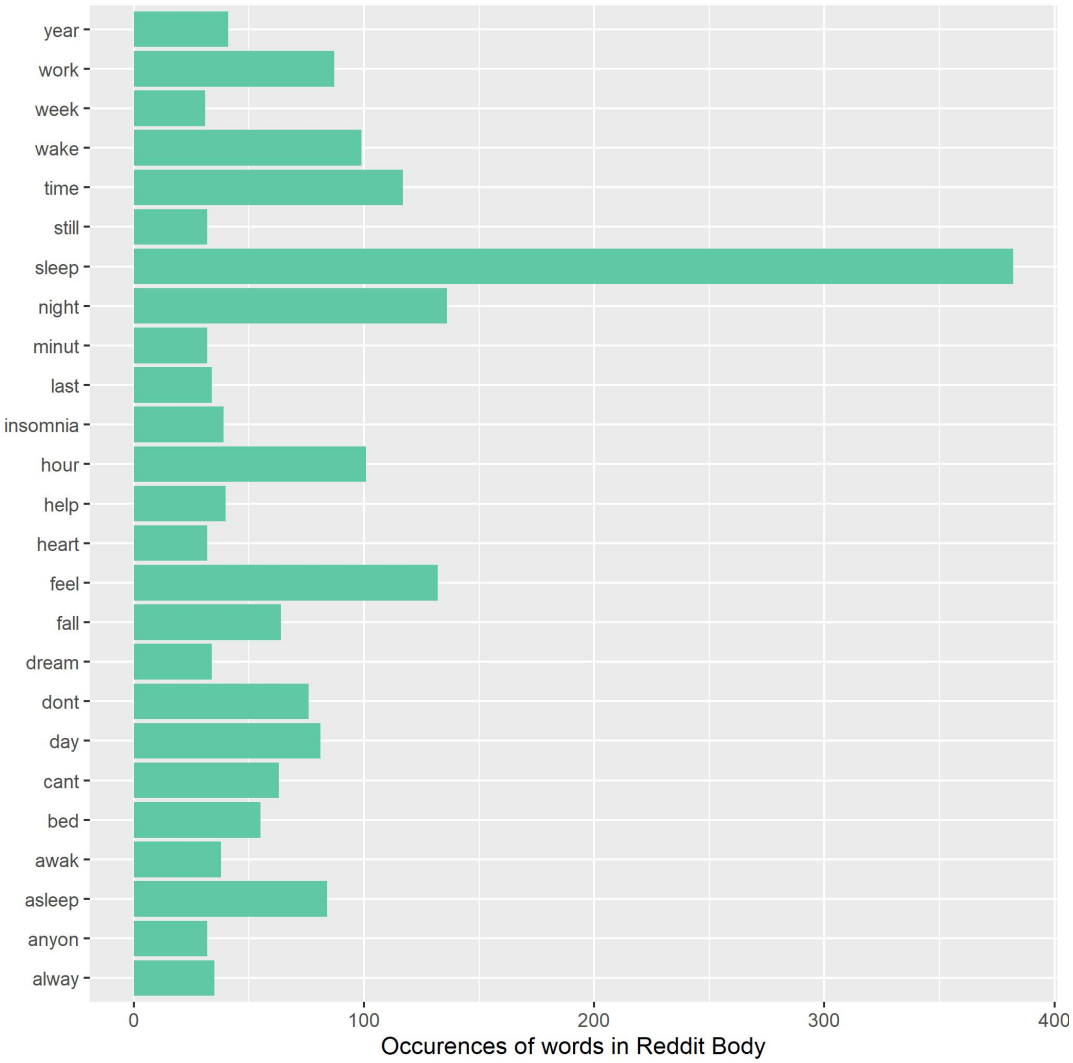


Figure 12: *Words appearing in Reddit body paragraphs*.

curb sleep disorders. If people are the most conscious/anxious regarding their sleep disorders at these hours, then this anxiety regarding sleep may be inhibiting sleep. As such, making individuals aware anxiety-relieving practices to do before bed (e.g. meditation), could reduce anxiety and promote sleep in themselves and those they know.

We also see temporal trends over the course of the year. Google queries for sleep disorders were most common at the beginning and end of the year. This could be due to weather effects, or perhaps the occurrence of stressful holidays. This implies that these times of year would be ideal for educating sleep disorder sufferers via pamphlets or by holding workshops.

Though our chi-squared tests, and other visualizations, revealed that certain words occur more commonly in sleep disorder tweets, we were unable to use these features to classify tweets as relating to sleep disorders based on these words. Our initial model performed quite poorly, indicating that either presence or absence of top words is a poor, or at least insufficient metric on which to classify tweets pertaining to sleep disorders for those that are not. Were we to do this project again, a more profitable way to summarize the lexical content of tweets for classification would be to perform vectorization on

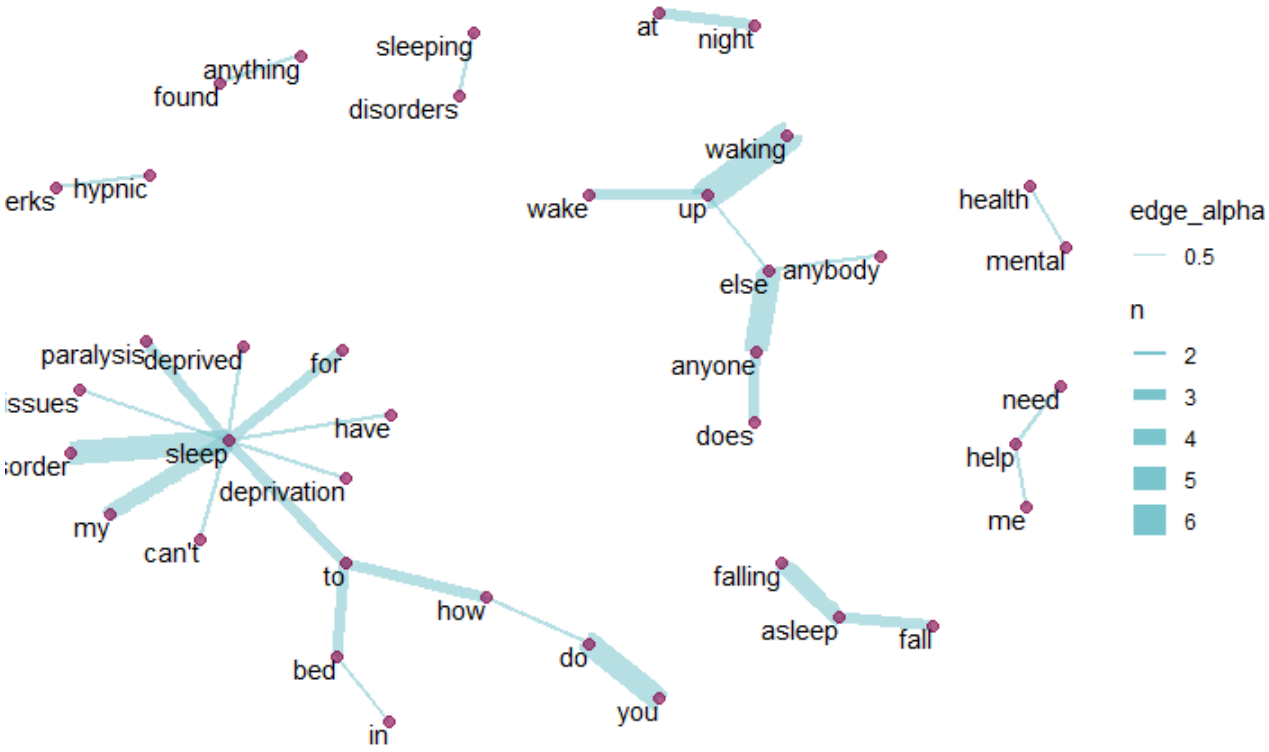


Figure 13: *Bigram showing the relationship between the appearance of certain words in titles of sleep disorder Reddit posts*.

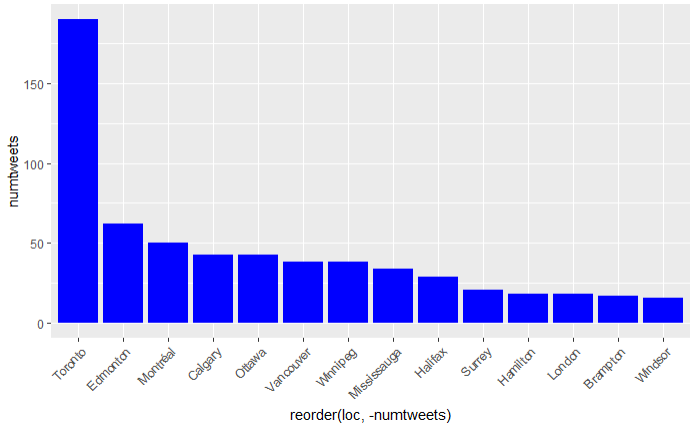


Figure 14: *Number of sleep disorder tweets for different Canadian cities (not adjusted for population size)*.

the tweets. This creates a continuous feature which summarizes the frequency of certain words or word groupings in each tweet. This approach is much more sophisticated than the one we implemented here, and it got good results when employed by other students.

Additionally, with data triangulation from Reddit, we can see that people are concerned with their sleep across different social media platforms. However, from the top words used it seems that users are engaging with others on these platforms regarding their problems quite differently. In Twitter, many of

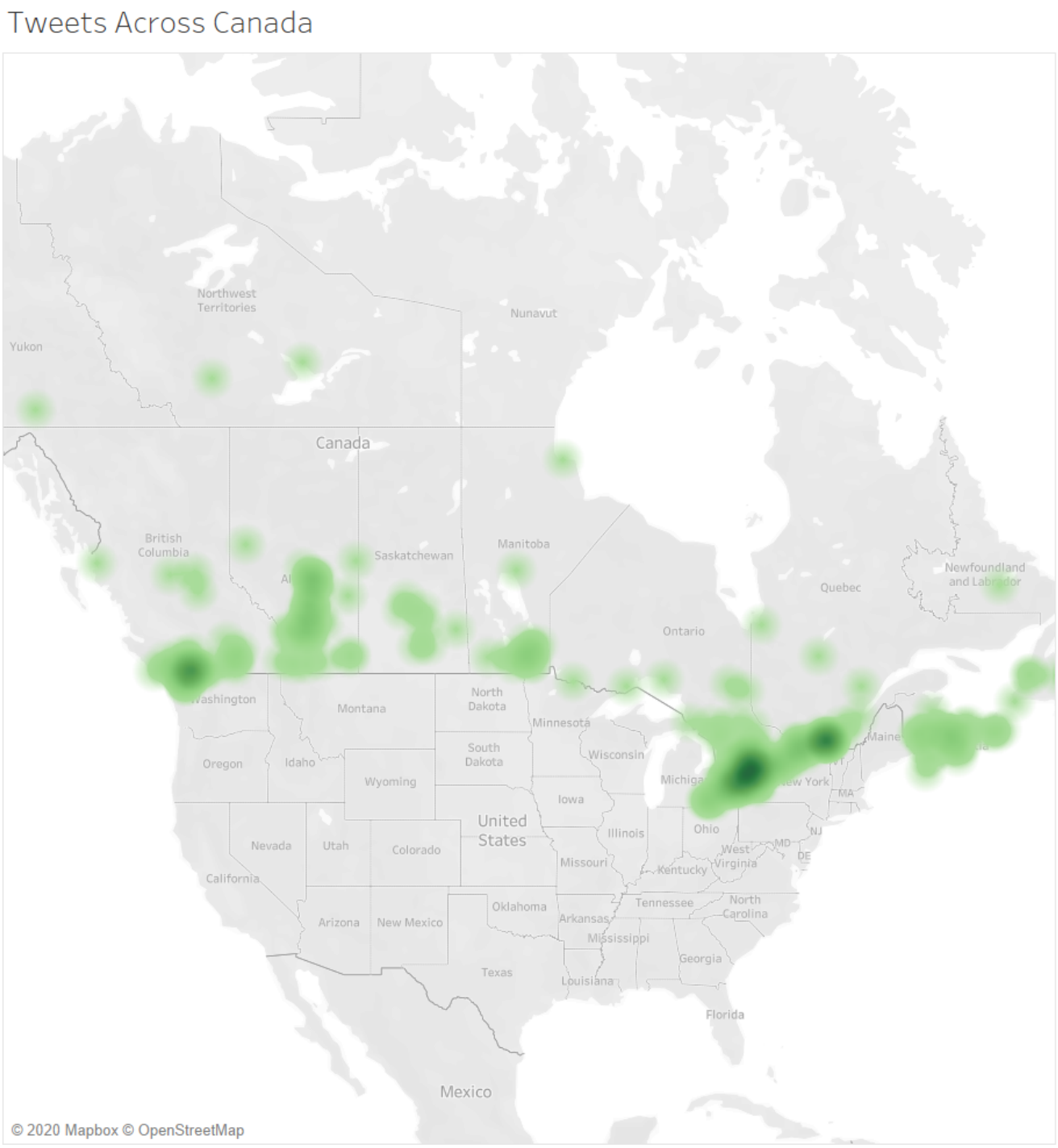
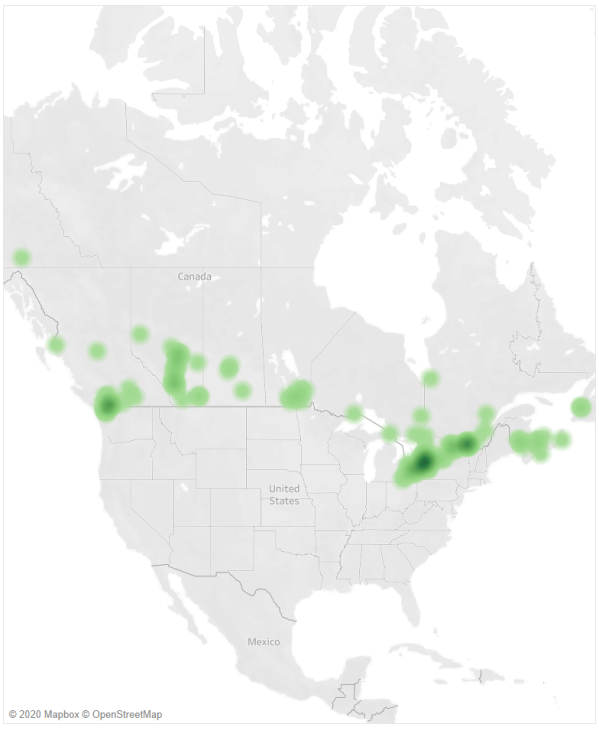


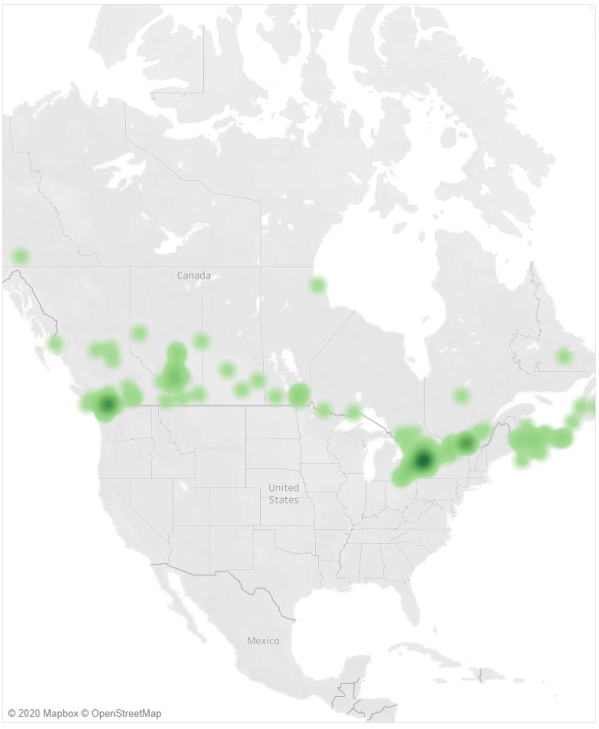
Figure 15: *Geographic locations of all Tweets*.

the top words seem to be describing sleeping issues, e.g. 'woke', 'tire'. Similar words also occur in the titles and bodies of sleep disorder-related Reddit posts, but there are also words that seem to be directed towards seeking advice from others, or seeing who else shares their problems; e.g. phrases like 'anyone else'. Though more thorough analyses are needed, this suggests that Twitter users may

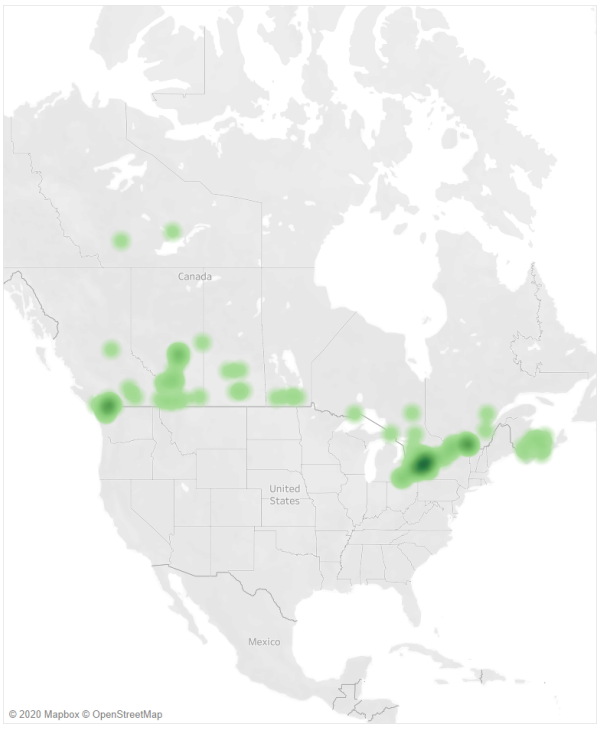


(a) No Self Report; No Sleep Disorder

(b) No Self Report; Yes Sleep Disorder



(c) Yes Self Report; No Sleep Disorder



(d) Yes Self Report; Yes Sleep Disorder
Yes Self Report; Yes Sleep Disorder

Figure 16:
The Geographic Distribution of Tweets Across Canada by Tweet Label

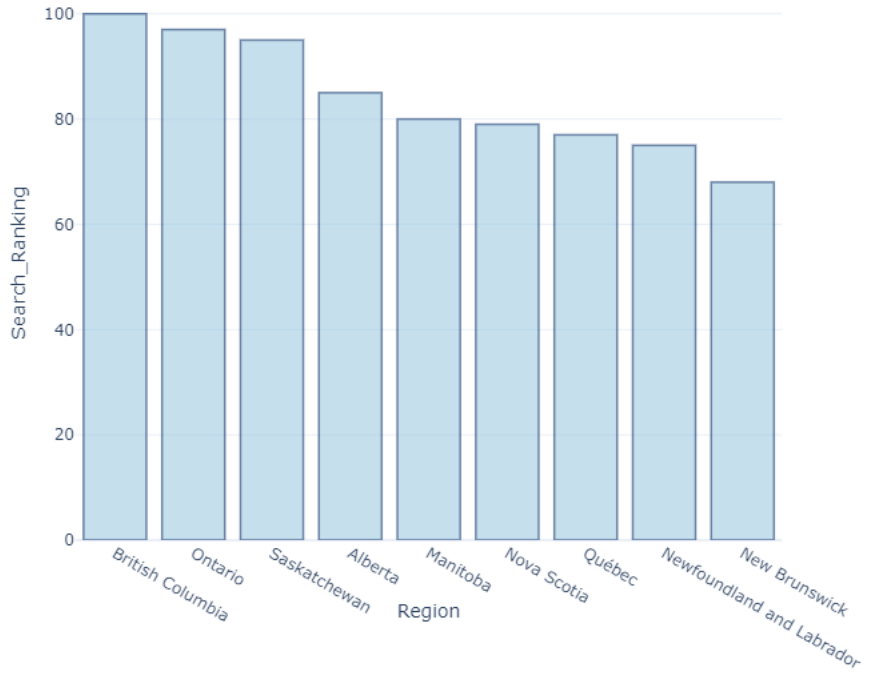


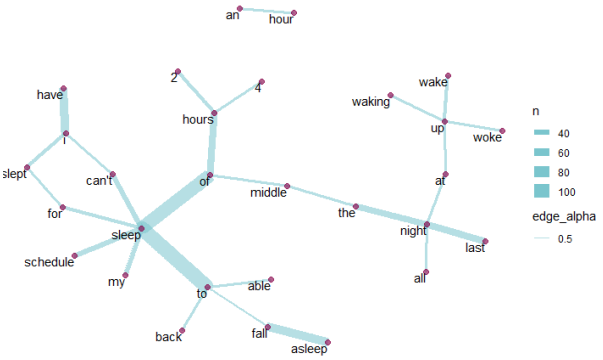
Figure 17: *Relative rankings of provinces for volume of tweets relating to sleep disorders*.

be more inclined to lamenting their sleep disorders to the internet without specifically seeking advice or help from others. On the other hand, Reddit users seem to be more geared towards sharing and seeking advice. Reddit's structure of subreddits related to specific topics may be what fosters this kind of engagement.

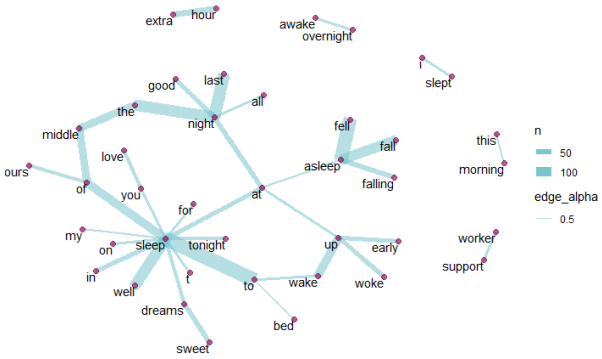
From our combined geographical analysis, we see that Tweets relating to sleep disorders seem to primarily occur in the larger urban centres, though they occur at some level all across Canada. These larger urban centres presumably have larger populations, so an additional step of controlling for population would reveal more insight into these patterns. When looking at Google searches for sleep disorders at the provincial level, we see that British Columbia and Ontario show the highest number of searches for sleep disorders. This may be because these provinces are home to the largest urban centres in Canada (Toronto and Vancouver). However, Google Trends data controls for population, which indicates that this is not only an effect of large population, but may indicate that living in these larger cities can cause difficulty in maintaining proper sleep habits.

8 Figure

References



(a) Bigrams showing the relationship between the appearance of certain words in sleep disorder tweets



(b) Bigrams showing the relationship between the appearance of certain words in non-sleep disorder tweets

Figure 18:
Twitter Bigrams

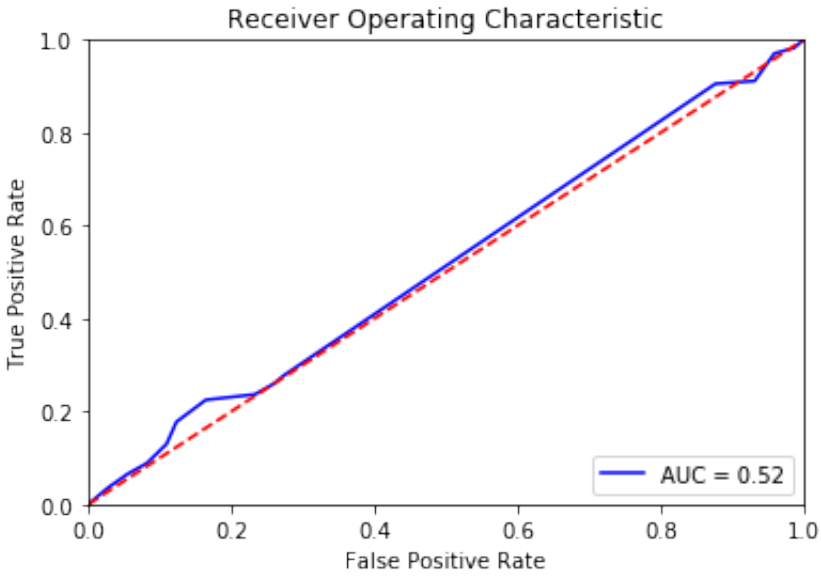
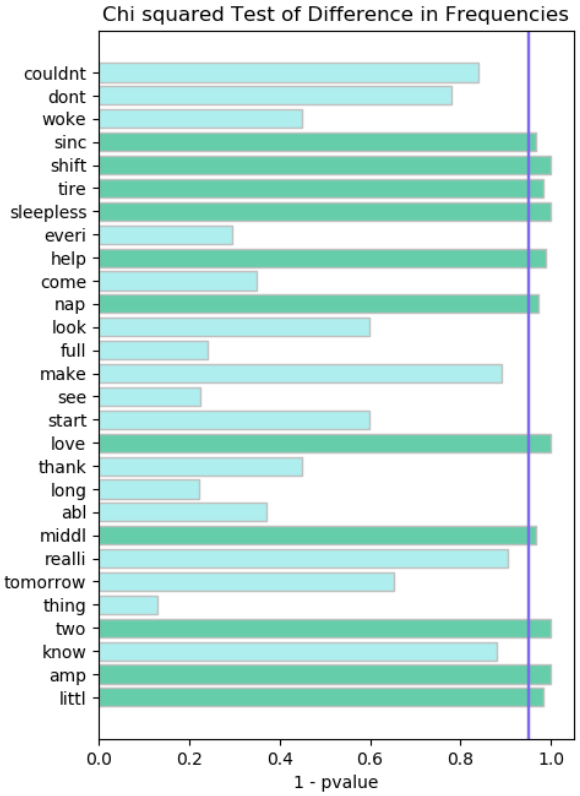


Figure 19: *ROC curve for our XGBoost model.*

```
Accuracy: 0.4049586776859504
f1: 0.3793103448275862
AUC: 0.5161708681202886
precision: 0.6984126984126984
recall: 0.2603550295857988
rand: -0.0028289267658011506
```

Figure 20: *Performance metrics for our XGBoost model.*



(a) 1-p values for frequency differences of top words across tweet category

words	pvals
littl	0.016363
amp	0.001760
two	0.000552
middl	0.033985
love	0.001034
nap	0.028243
help	0.010566
sleepless	0.000235
tire	0.017341
shift	0.000053
sinc	0.032267

(b) p values for words with significantly different frequencies

Figure 21:
Results of chi-squared tests