# Magnitude of the Sleep Disorders Problem in Canada Analyzed from Web Sources

**Maruthi Kumar Mutnuri, Dina Labib**

**Data Science 624-Winter 2020**

**April 19, 2020**

## Key Findings of the Project

We can list the key findings of this project as:

**Key Findings**

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

1. Google Trends was the most valuable source of information regarding sub-categories of sleep disorders in Canada; Twitter and Governmental data were useful only in evaluating insomnia. The provided geographical distributions for different categories of sleep disorders would be useful to guide provincial policy makers to direct resources for specific disorders.
2. Regarding insomnia - a surrogate to psychological problems such as depression - different results were obtained regarding geographical distribution from Google Trends and Twitter, with the former showing Prince Edward Island and British Columbia to be the top provinces with insomnia, while the latter (using normalized counts) shows the Northwest Territories and Nova Scotia to be the most commonly affected.
3. Obstructive sleep apnea - an important risk factor for cardiovascular disease - is most prevalent in Saskatchewen and British Columbia. Distributions for parasomnia and narcolepsy are also provided.
4. Twitter could be a valuable source to predict the problem of sleep disorders, mainly insomnia, based on Tweet text.
5. Governmental data reveals a high prevalence of insomnia (over 20%) among all categories of age, gender, and household income of Canadian adults. This specific problem warrants dedicated efforts from the Government and general practitioners.

# 1 Introduction

Sleep deprivation can cause major ailments including heart-related problems, hypertension, and diabetes mellitus [SLGM+09]. Accordingly, there is a significant public health risk with the seemingly innocuous habit of neglecting getting the required amount of sleep. There are various types of sleep disorders, such as insomnia and sleep apnea. We would like to study the sleep disorder patterns using a Twitter dataset complemented by other web sources, namely Google Trends and Physical Activity, Sedentary Behaviour and Sleep (PASS) Indicators Data Tool provided by the Public Health Agency of Canada.

## 1.1 Project Motivation

Sleep disorders have a major impact on work productivity and individual/public safety; it is well known that sleep disorders are a common cause of traffic accidents. We are seeking to explore the magnitude and common types of sleep disorders in Canada overall and at a provincial level. Hopefully, this will serve the ultimate goal of guiding policy makers to allocate more resources to reduce common sleep disorders in provinces where they are more prevalent. In an attempt to overcome the limitations of relying only on the Twitter dataset that spans a limited time interval of approximately three months, we will use data source triangulation - defined as testing the validity of our results through the convergence of information from other sources.

## 1.2 Problem Definition

- **Output:** We aim to answer the following questions:
  1. What is the geographical (per province) and time distribution of sleep disorders? In the first part of the project, we will analyze the volume of tweets about sleep disorders by Canadian provinces and time (week day and day hour). In the second part, we will implement data source triangulation to validate our initial results regarding geographical distribution. To this end, we will complement our Twitter data with that collected through searching Google Trends and the PASS data tool provided by the Public Health Agency of Canada.
  2. Using unsupervised learning, what are the common topics and words in self-reported tweets about sleep disorders compared to non-sleep disorders.
  3. Using supervised learning algorithms, can we classify self-reported tweets into those reporting sleep disorders versus not.
- **Input data for algorithms/solutions:**
  1. Twitter dataset
     - Number of rows (tweets): 6006
     - Columns (variables) to be explored: tweet ID, tweet time (created at), tweet location (place.full name and place.name), tweet text, and Twitter user-related data, including number of followers and friends.
     - Time-span of the dataset: March 25, 2019 - May 9, 2019 and October 12, 2019 - December 1st, 2019.
     - Challenges regarding this dataset: The 2 main ones are data quality as a many tweets (1811 in total) were labelled as "not clear" or "non self-reported", as well as the relatively small number collected over a short time span.
  2. Data from Google Trends

     Google Trends tool has the advantage of providing the flexibility to customize search regarding sleep disorders' subcategories, time span, and provinces. We aim to extract data for the past 12 months.

3. Data from the PASS tool

   The Public Health Agency of Canada provides data on the percentage of Canadian adults reporting having trouble going to sleep or staying asleep "most of the time" or "all of the time" for the years 2014-2015.

- In our opinion, the most relevant factors (variables) in the Twitter dataset that can help to address the project's objectives are tweet text, time, and place, as well as Twitter user variables including number of friends and followers and duration of Twitter account activity.
- Task of labelling the Twitter dataset: Our Twitter dataset was one of 3 Twitter datasets, collectively constituting around 9000 rows. Tweet labelling was done collaboratively by course students, where each student was responsible for labelling around 800 tweets. For the sleep disorders dataset, one of 5 labels were selected: self-reported yes/sleep disorder yes, self-reported yes/sleep disorder no, self-reported no/sleep disorder yes, self-reported no/sleep disorder no, and not clear.

---

# 2 Methodology

## 2.1 Analyzing the geographical and time distribution of sleep disorders

- Twitter dataset:
  To study the geographical distribution of tweets on sleep disorders, we chose to report the counts of these Tweets (self reported + non-self reported) per province. Tweet location on a per province basis was extracted from the place.full name entered as (city, Province). We also considered reporting on city level; however, the numbers were relatively small even at a provincial level (maximum of 1585, followed by 409, absolute tweet count for combined self- and non-self reported tweets on sleep disorders). A geographical map of the provinces was the most suitable visualization (figures 1 and2). Since provinces with larger population count would be expected to have a larger tweet count, it would make sense to normalize the tweet counts by population count per province. The latter was obtained from Statistics Canada website reporting yearly and quarterly population counts ([Can19]). Normalized counts were given per 10,000 population to avoid otherwise difficult to understand counts with many decimal places. A "normalization toggle" button is shown to the right of the map to give the option of displaying absolute tweet counts ("Don't Normalize" option) or the normalized ones ("Normalize" option). Finally, to allow meaningful comparison with the Google Trends dataset (further details mentioned below) in which data are on a scale from 0-100, the final step was to convert Tweet count into the same scale, by dividing with the largest provincial count (1585 for non-normalized values and 1.782 for normalized values), and multiplying by 100. A value of 100 denotes the province with the highest count, with all other provinces having smaller values relative to the highest. The final value for Tweet counts for sleep disorders per province on a scale from 0-100 was represented both as the intensity of background blue color for provinces, as well as the height of the blue bar.

  In order to study the timing of tweets, we chose a rose chart to show the total self-reported tweet counts per hour for each of sleep/non-sleep disorders indicated by the radial height of the bars. This kind of chart is more suitable than conventional bar charts as the 24 hours would be displayed as a 24-hour clock (figure 3). Next, we showed the distribution of tweets per hour of the day for each of the 7 days of the week in a modified Gantt chart (figure 4), with different colours for tweets on sleep disorders versus non-sleep ones. The absolute count of tweets at a particular time, such as 6:27 am, is represented as a square, with the area of the square being proportional to the Tweet count.

  It should be noted that Tweet hours were originally provided as UTC timing, prohibiting derivation of meaningful conclusions when combined for analysis. In order to mitigate this, we we re-coded thtthe the e hours according to the individual provincial time zones.

Starting from the timing of tweets and moving forward to all further analyses of the Twitter data, we chose to report only self-reported tweets as these reflect the emotions of the person who Tweets, ultimately influencing Tweet timing and topic characterization, as opposed to tweets by a person reporting another person's problem.

- Data from Google Trends:
  These were extracted as a csv file through searching for sleep disorders in general, as well as the individual subcategories, including insomnia, parasomnia, sleep apnea (overall and obstructive subcategory), and narcolepsy across Canadian provinces over the past 12 months. We chose this particular time span to provide adequate data compared to the that collected over the limited time span of the Twitter dataset. In Google trends, numbers reflect "*popularity score*",defined as the search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 denotes the term is half as popular. A score of 0 means there was not enough data for this term. Visualization for this dataset is combined with Twitter dataset and shown in the same geographical map (figures 1 and 2) for easy comparison, and were displayed as colored bars for different categories side by side with Twitter bars, with the height of the bar corresponding to the search term value (popularity score).

- PASS indicators data:
  Data were extracted on the percentage of Canadian adults aged 18-79 who reported having trouble going to sleep or staying asleep "most of the time" or "all of the time" for the years 2014-2015, with values reported for categories of age, gender, and household income level.([fSAR18]) Data for other years were not available on the website. Simple, easy to understand, visualizations were provided for this part, including bar charts for age and income categories (figure 5). It was not possible to perform formal statistical tests to compare different categories as only the percentages, not absolute raw counts, were available.

### 2.2 Unsupervised learning-based extraction of common topics and words in tweets

To gain insight into topics in tweets on sleep disorders which would hopefully help us identify common sleep disorders, we used latent Dirichlet allocation (LDA) for our topic modeling - a technique which facilitates the automatic discovery of themes in a collection of documents. We created a new column for the tweet text, given as the extended tweet text in case the original one was truncated and the original one if complete.

Topic modelling was performed in R Software, using packages: “tm” and “SnowBall”, "topicmodels", and "tidytext. "SnowBall" package was used for stemming, defined as reducing relevant words to their common root, such as reducing “slept” to “sleep”. Steps other than stemming included converting hyphens and colons to white spaces, removing punctuation and numbers, converting upper case characters to lower case, removing stop words, and then finally stripping white spaces. Stop words removed included those from the English dictionary included in the “tm” package, as well as others relevant to the context, such as "up", "go", and "https". Multiple iterations were performed until the final output was satisfactory.

The results of topic modeling were presented as bar charts (figure 6) showing beta (per-topic-per-word probabilities) for each of sleep / non-sleep disorders. Two topics were chosen as this is the minimum number and it later turned out the majority of sleep disorder tweets are on insomnia, theoretically only one topic.

Next, we used R packages igraph and ggplot2 to create word graphs (networks) to visualize the relationships among words (bigrams) simultaneously for self-reported tweets on sleep disorders and non-sleep disorders separately (figure 7). These graphs show the word node an edge is coming from (first word of bigram) and the node an edge is going towards (second word of bigram), with the bigram frequency (weight) represented by the thickness of the edge. We chose a minimum threshold of 10 and 5 for bigram frequency for sleep disorder and non-sleep disorder tweets, respectively, due to the relatively small frequencies in our dataset.

### 2.3 Using supervised learning to predict sleep versus non-sleep disorder related tweets

We were interested in predicting whether the self-reported tweets were related to sleep disorders or not. We chose random forest modelling for this classification problem due its multiple advantages, including its well-known performance with high dimensional data, robustness to outliers and non-linear data, as well as capability of handling imbalanced data. Our data are high-dimensional as we planned to use tweet text as our main predictor, with features corresponding to the many words per document (tweet) corpus.

Our first step was to run a random forest classifier (model 1) from Python's Scikitlearn on Tweet text, using TF-IDF vectorizer approach. Through this approach, the corpus of tweets is converted into a sparse matrix, with each row representing a tweet, columns representing the total words in the corpus, and individual cells containing TF-IDF values. Term frequency (TF)is defined as the number of times a word appears in a document divided by the total number of words in the document. Inverse document frequency (IDF) is defined as the log of the number of documents divided by the number of documents that contain the word. IDF determines the weight of rare words across all documents in the corpus. Lastly, the TF-IDF is simply the TF multiplied by IDF. We split the dataset into 80% training and 20% testing. We applied random grid search, with 3-fold cross-validation and F1 score as the performance metric, to tune three hyper-parameters: n-estimators (number of trees in the forest), max-depth (maximum depth of the tree), and max-features (number of features to consider when looking for the best split). The model with the best hyperparameters was run on the entire training data and evaluated on the test portion. Since our dataset was imbalanced (2683 sleep disorder tweets/1502 non-sleep disorder ones), we tested 2 approaches for handling imbalanced data: random forest with balanced sub-samples and balanced random forest classifier. The first method changes the class weighting based on the class distribution in each bootstrap sample, instead of the entire training dataset; the second one implements random undersampling of the majority class in each bootstrap sample. Predictions of class labels

The second step was to create a new dataset containing the predicted class labels from the first step as well as three continuous features: number of user's friends, number of followers, and influential ratio (number of followers divided by friends), with the first two normalized to the account duration in days. The 3 numerical features were scaled using scaling to the range approach (MinMax scaling). The motivation for adding these features was the the research by [MHC+15] that concluded that Twitter users with sleep disorders were less socially active on Twitter as denoted by less friends and less followers, when normalized to the lifetime of the Twitter account. We applied a second random forest classifier (model 2) to this new dataset after tuning the hyperparameters using the same methodology as model 1.

All steps performed on the training set were repeated for the test set.

## 3 Results

### 3.1 Geographical Distribution of Sleep Disorders

#### 3.1.1 Twitter Datset

The maps in figures 1 and 2 show the distribution of non-normalized and normalized Twitter and Google Trends values on a scale from 0-100 in Canadian provinces, with the former data collected over approximately 3 months and the latter over one year. Regarding non-normalized values of sleep disorder-related tweets (both self-reported and non-self reported), Ontario has the highest count (absolute count 1585, scaled value 100), followed by Alberta (401, 26) and British Columbia (409, 25). However, looking at the normalized values, Northwest Territories had the highest count (count 1.78, scaled value 100), followed by Nova Scotia (1.39, 78).

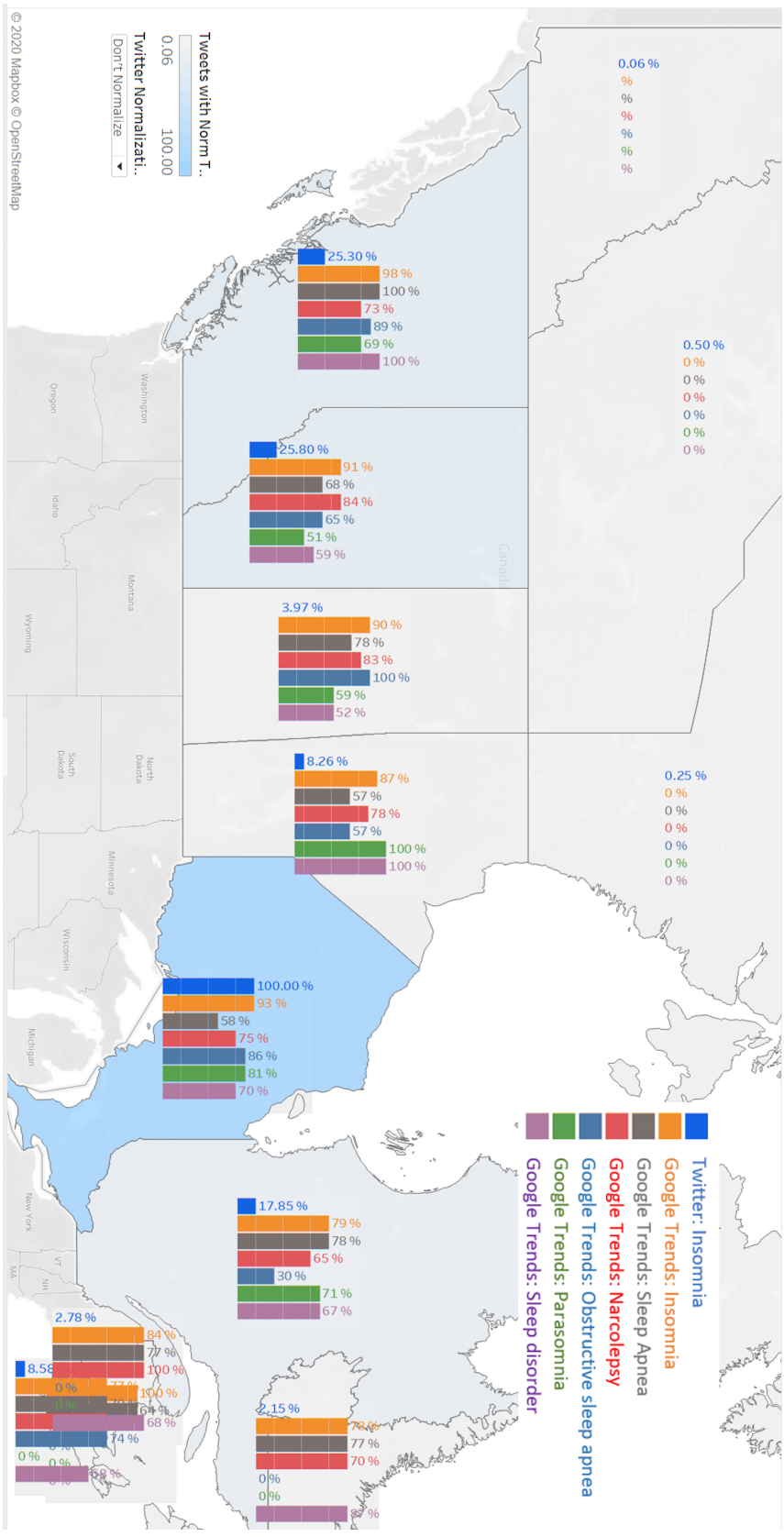


Figure 1: Google Trends popularity score (March 2019-March 2020) & scaled Tweet counts unnormalized by population count (collected over 3 months of 2019) for sleep disorders by province

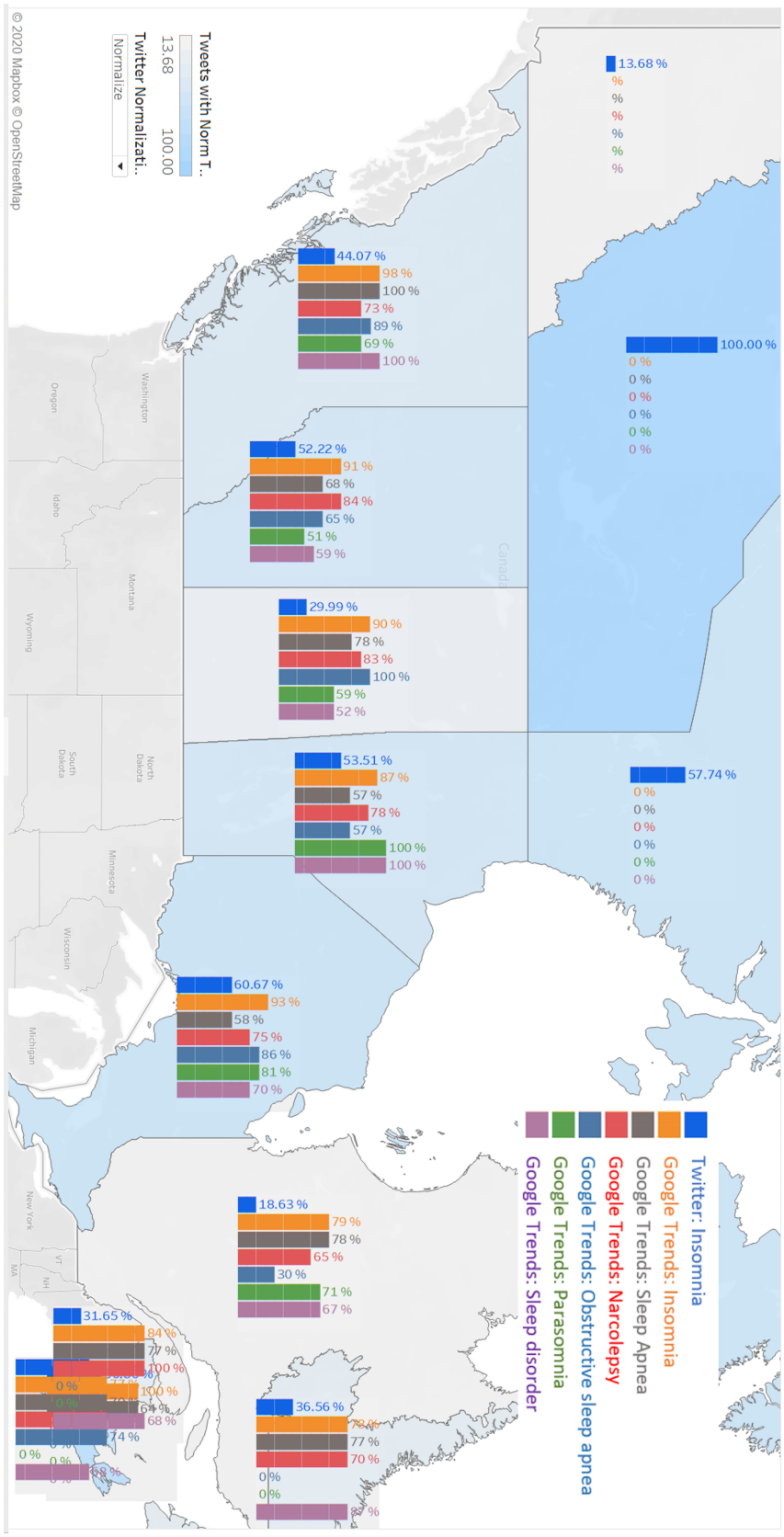


Figure 2: Google Trends popularity score (March 2019-March 2020) & scaled Tweet counts normalized by population count (collected over 3 months of 2019) for sleep disorders by province

### 3.1.2 Google Trends Data

Compared to the Twitter data, where insomnia was the dominant theme as we will detail later, Google Trends show other common sleep disorder types, as well as a different geographical pattern for insomnia (figures 1 and 2). The top 2 provinces with insomnia are Prince Edwards Island (popularity score 100) followed by British Columbia (98), while the top ones for obstructive sleep apnea are Saskatchewen (100) followed by British Columbia (89). Looking at individual provinces, in Alberta, the most commonly searched sleep disorder is insomnia, followed by narcolepsy, with popularity scores of 91 and 84, respectively. There are not enough data for sleep disorders generally and the different categories in Northwestern territories and Nanuvut (popularity score of 0).

## 3.2 Analysis of timing of self-reported tweets

Figure 3 shows a rose chart of the total counts of tweets per hour of the day, with sleep disorder-related tweets in orange and non-sleep disorder-related ones in blue. Tweets on sleep disorders are most common between 10 pm - 1 am, while those related to non-sleep disorders are scattered throughout the day, but are relatively more common at 8 am and 9-10 pm.

Figure 4 shows a further breakdown of Tweet time by week day and day hour for each of the sleep disorder-related tweets (orange) and non-sleep disorder-related ones (blue). Tweets on sleep disorders are more common between 8 pm - 6 am on weekdays, but spread out over the entire day during weekends. There are no obvious patterns for non sleep disorder-related tweets.

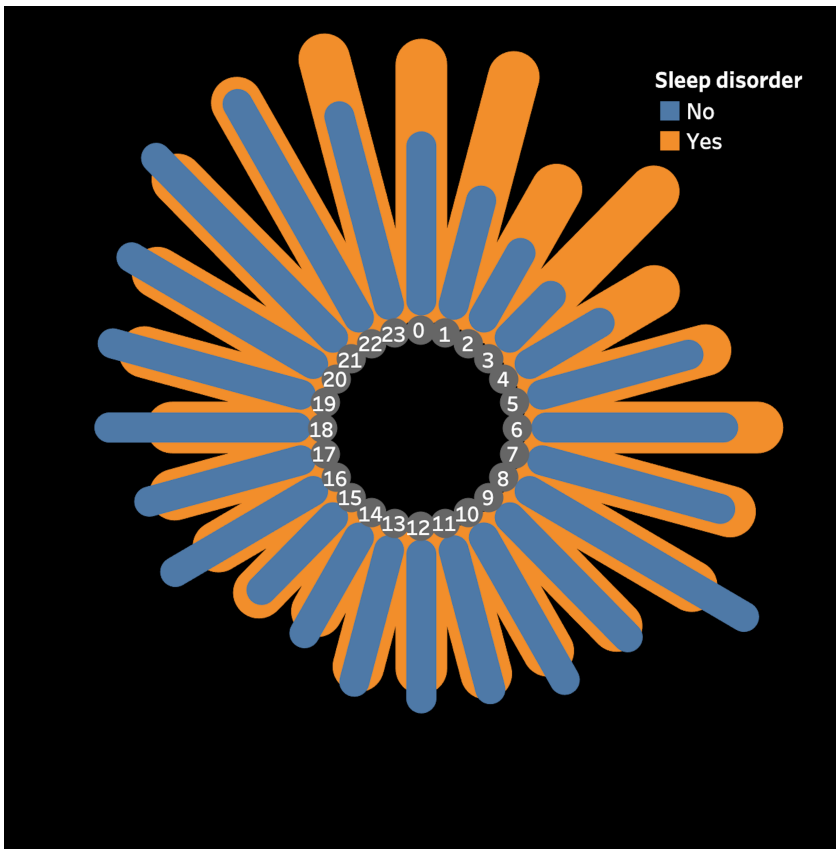


Figure 3: Distribution of count of self-reported tweets by hour

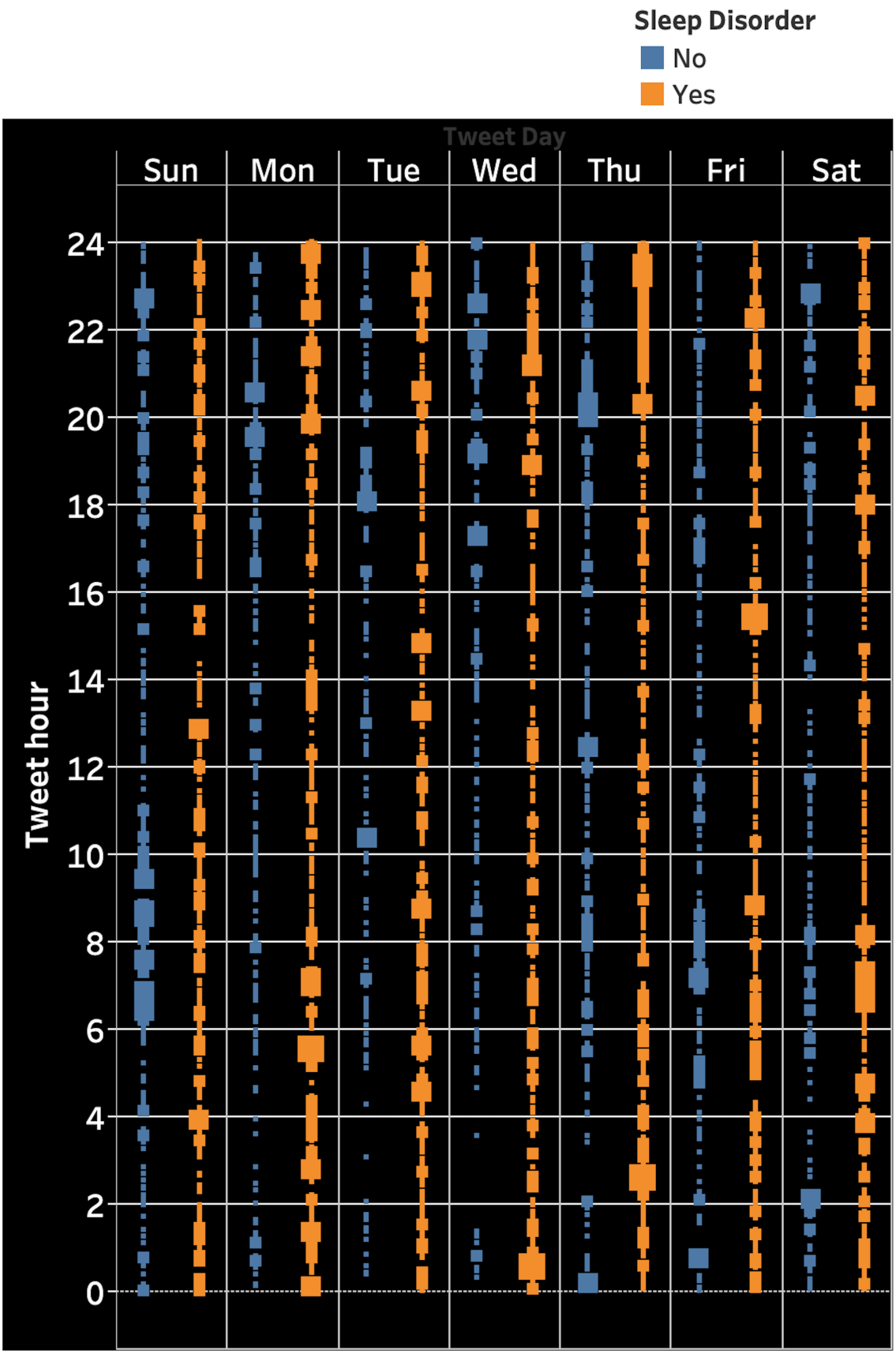


Figure 4: Distribution of count of self-reported tweets by day and hour

### 3.3 Analysis of PASS indicator datatool

Similar to Twitter data, this dataset provided only information about insomnia. The prevalence of insomnia was high across all categories of age, gender, and household income, with a value always greater than 20% (figure 5). Insomnia was detected in 30% and 20% of males and females, respectively. Insomnia was more common in age group 50-64 and low household income categories.

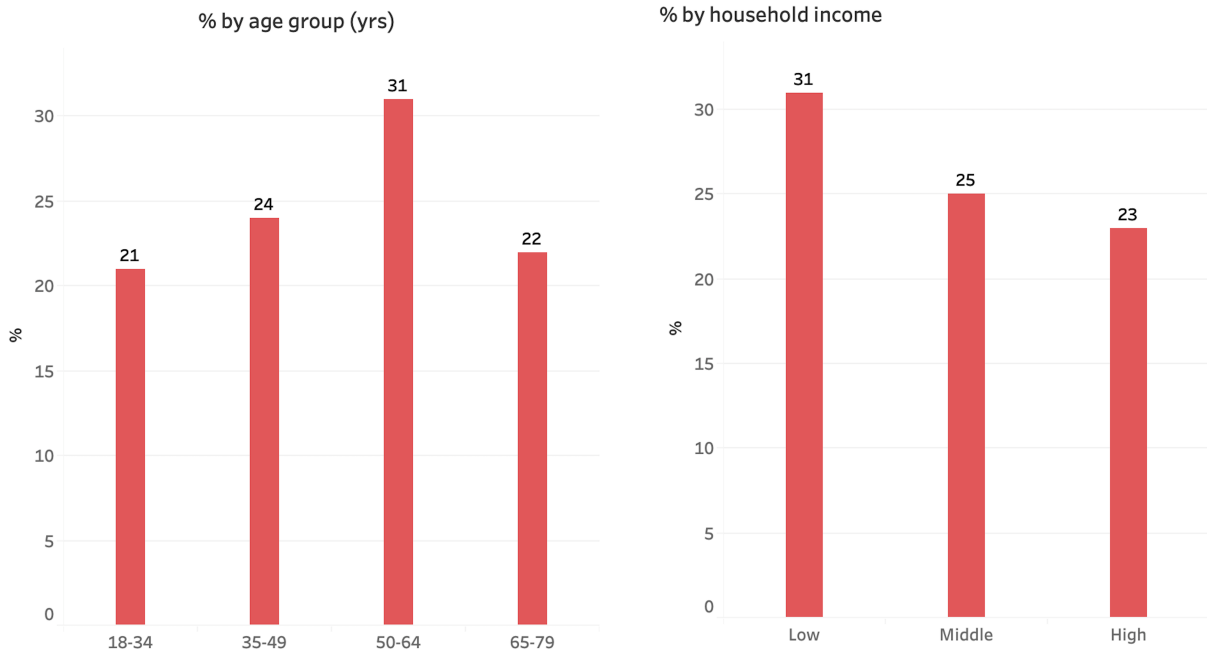


Figure 5: % of Canadian adults reporting having trouble going to sleep or staying asleep most/all of the time (2014-2015) per gender, age and income categories (Source: Canadian Community Health Survey - PASS indicators)

### 3.4 Unsupervised Modeling

Figure 6 shows the results of topic modeling with the frequency of the top 15 common words in each topic. We believe that it is not possible to derive 2 topics for either of the 2 tweet categories as the sleep disorders ones are most likely related only to insomnia, and the non-sleep disorders are generally related to sleep. commonly occurring words in sleep disorder-related tweets include "insomnia", "nightmare", and "melatonin". Interestingly, "work" is a common word in sleep disorder-related tweets (probability of 0.018 in topic 1), pointing to work problems being a common cause for insomnia. Non-sleep disorder tweets generally convey positive emotions, such as "good", "goodnight", and "love".

Figure 7 shows graphs of word bigrams. Commonly occurring bigrams are "sleep schedule", "wide awake", "haven't slept", again emphasizing the problem of insomnia in sleep disorder tweets. Another common bigram in sleep disorder-related tweets was "sleep paralysis"; this was the only hint of another sleep disorder besides insomnia from Twitter dataset analysis. Similar to topic modelling analysis, non-sleep disorder tweets convey good emotions, with common bigrams including "sweet dreams", "deep sleep", and "sleep goodnight".

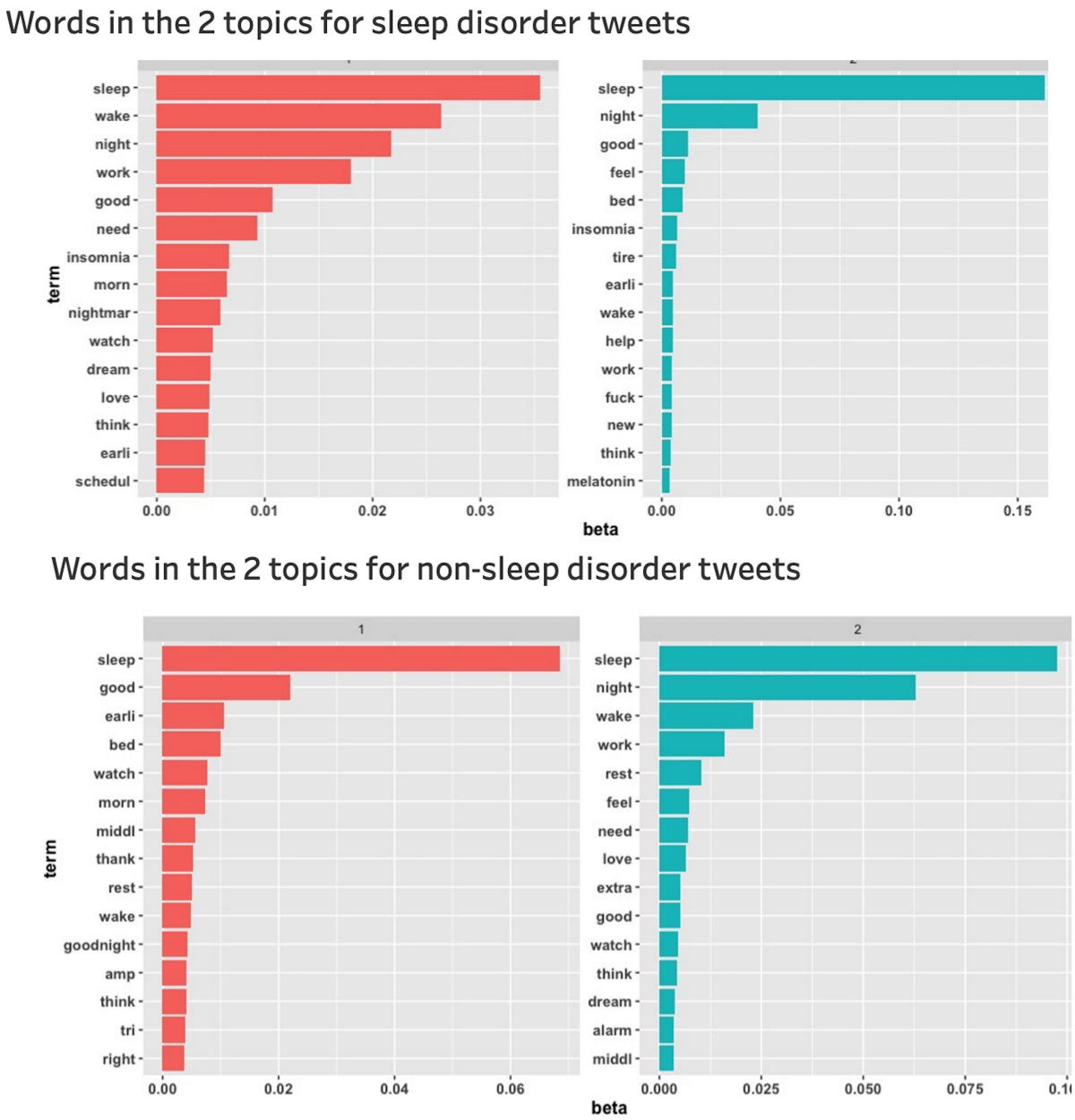


Figure 6: Results of topic modelling

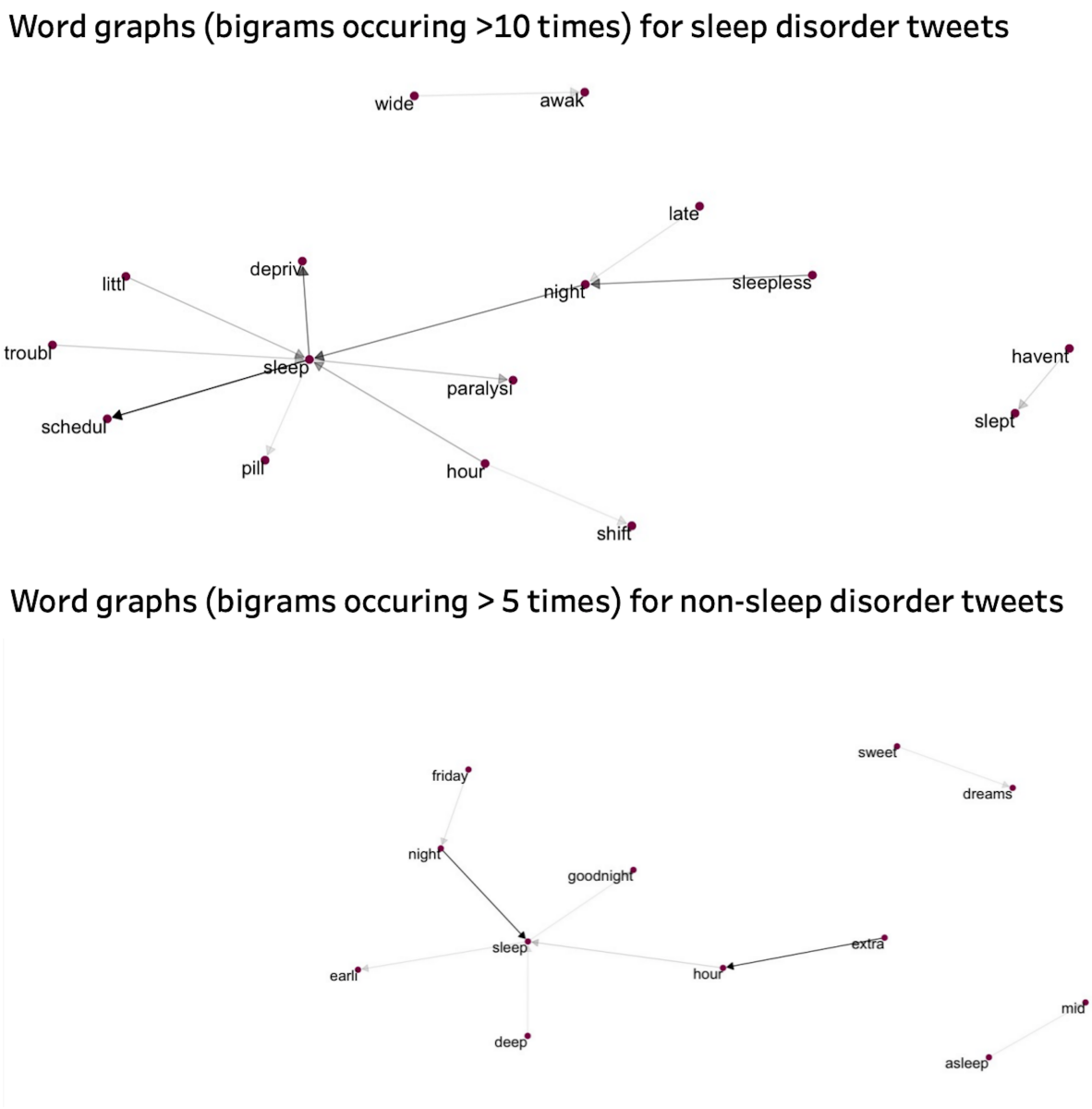


Figure 7: Word graphs of bigrams for sleep disorder/non sleep disorder-related tweets

## 3.5 Supervised learning for classification of self-reported tweets into sleep disorder-related / non-sleep disorder-related

### 3.5.1 Random forest model including only Tweet Text (Model 1)

Table 1 shows performance metrics for the 2 random forest (RF) models with different approaches for handling class imbalance when evaluated on test dataset. Although overall model performance with both approaches is reasonable, with an F1 score of 0.73-0.74, and AUC under ROC curve of 0.72-0.73 (figures 8 and 9), the performance of both approaches with the minority class of non-sleep disorder tweets is poor, with an F1 score of 0.47 and 0.58 for balanced subsample and balanced RF approaches, respectively. From the table, all metrics are relatively better for minority class prediction with the balanced RF approach, however overall model performance id=s better with the other approach.

| | **RF with balanced subsample** | | | | **Balanced RF** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **Specificty** | **F1 score** | **Precision** | **Recall** | **Specificty** | **F1 score** |
| **Non-sleep** | 0.72 | 0.35 | 0.93 | 0.47 | 0.51 | 0.66 | 0.65 | 0.58 |
| **Sleep** | 0.72 | 0.93 | 0.35 | 0.81 | 0.78 | 0.65 | 0.66 | 0.71 |
| **Overall score** | 0.72 | 0.72 | 0.55 | 0.72 | 0.68 | 0.66 | 0.66 | 0.66 |
| **Overall ROC AUC** | | | | 0.74 | | | | 0.73 |

Table 1: Performance metrics for random forest (RF) model to predict sleep disorder based on Tweet text (model 1), using 2 approaches to handle class imbalance

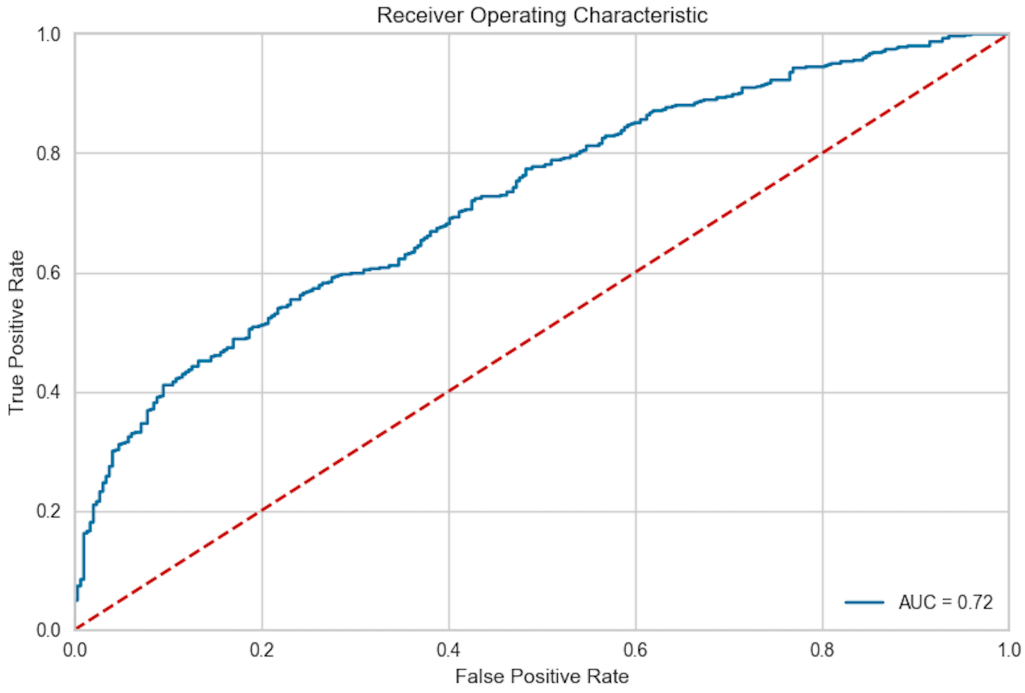


Figure 8: ROC curve of random forest Model 1 using balanced subsample approach

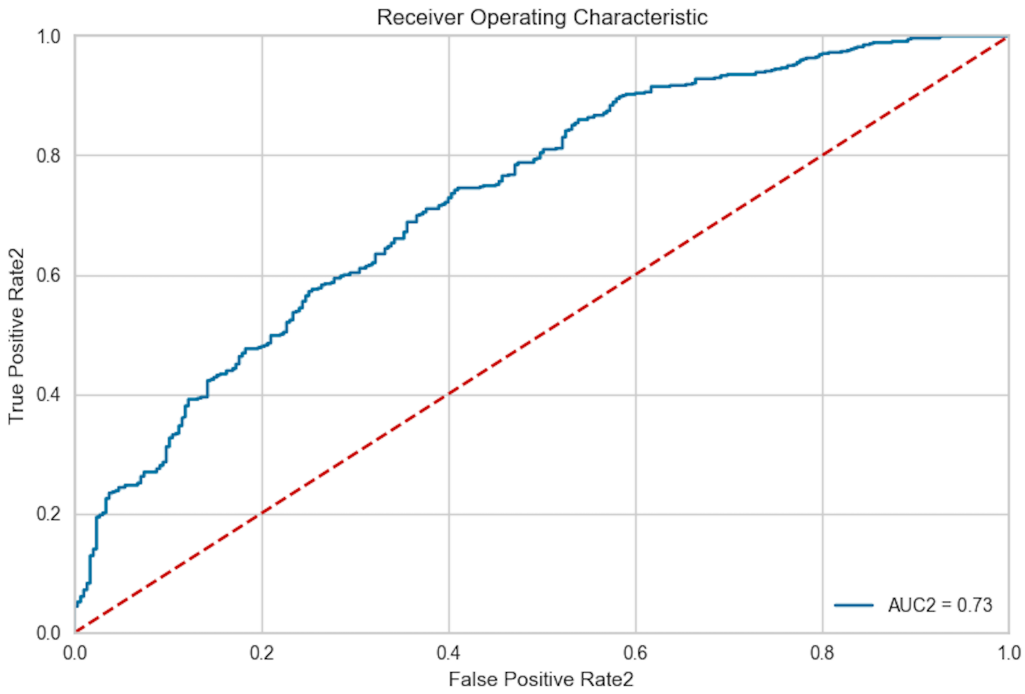


Figure 9: ROC curve of random forest Model 1 using balanced random forest approach

### 3.5.2 Adding other features to RF model (model 2)

Constructing a second RF model including user-related features and the label predictions from Model 1 yielded very poor results with both approached for handling imbalanced data, with relatively better performance with the balanced subsample approach (results shown in table 2). This step turned out useless to improve our predictions.

| | **RF with balanced subsample** | | | | **Balanced RF** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **Specificity** | **F1 score** | **Precision** | **Recall** | **Specificity** | **F1 score** |
| **Non-sleep** | 0.31 | 0.61 | 0.25 | 0.41 | 0.35 | 1 | 0 | 0.52 |
| **Sleep** | 0.54 | 0.25 | 0.61 | 0.34 | 0 | 0 | 1 | 0 |
| **Overall** | 0.46 | 0.38 | 0.49 | 0.38 | 0.12 | 0.66 | 0.66 | 0.35 |
| **Overall ROC AUC** | | | | 0.42 | | | | 0.43 |

Table 2: Performance metrics for random forest (RF) model to predict sleep disorder based on predictions from Model 1 and numerical features, using 2 approaches to handle class imbalance

## 4 Discussion and Conclusion

Main findings:

- Google Trends was the most valuable source of information regarding sub-categories of sleep disorders in Canada; Twitter and Governmental data were useful only in evaluating insomnia. Re-

garding insomnia - a surrogate to psychological problems such as depression - different results were obtained regarding geographical distribution from Google Trends and Twitter, with the former showing Prince Edward Island and British Columbia to be the top provinces with insomnia, while the latter (using normalized counts) shows the Northwest Territories and Nova Scotia to be the most commonly affected. Obstructive sleep apnea - an important risk factor for cardiovascular disease - is most prevalent in Saskatchewen and British Columbia. Distributions for parasomnia and narcolepsy are also provided. These results would be useful to guide policy makers to direct resources, such as specialized health personnel and specialized clinics for specific sleep disorders.

- Twitter could be a valuable source to predict the problem of sleep disorders, mainly insomnia, based on Tweet text, with no role for Twitter user-related features from this dataset.
- Governmental data reveals a general trend of high prevalence of insomnia (over 20 %) among all categories of age, gender, and household income of Canadian adults. This specific problem warrants dedicated efforts from the Government and general practitioners.

These are the main limitations:

- Different results were obtained regarding the geographical distribution of insomnia from analysis of Twitter and Google Trends datasets. This warrants further exploration. One possible explanation is the different purpose of using each platform. A Twitter user would want to express his feelings about a topic or problem, while a person who "googles" is seeking information about a topic/problem. The latter does not necessarily have the problem, but might want to get information out of curiosity, as part of an educational project, or because he has an acquaintance (family member/friend) with this problem. We believe that a self-reported tweet is usually a more important indirect indicator of having a particular personal problem compared to Google search. To prove this, we propose future research with a study design aiming at comparing the usefulness of both platforms to predict having a sleep disorder problem. Moreover, Twitter is a public social media platform, while Google Trends platform reflects private searches, yielding different topics. Lastly, the Twitter dataset was collected over a short duration of approximately 3 months, while Google Trends data were derived from a one-year duration; therefore the results might not be comparable.
- It was surprising that Northwestern Territories had the highest prevalence of insomnia according to Twitter data. We would have expected this problem to be more prevalent in more industrialized, busy provinces, such as Ontario and Quebec. One possible explanation is that North Western territories are one of the lowest income (lowest GDP) Canadian provinces, with higher financial stress and, theoretically, higher sense of insecurity.
- In order to predict sleep disorders from tweets, our comparison cohort was tweets searched using the term "sleep" and thus could be biased. It would have been better to choose a comparison cohort of all tweets reported over a similar, long enough duration.
- While we adjusted the tweet hour timing to address UTC time zone, we did not consider the problem of daylight saving time.
- Our random forest model analysis for the tweet classification problem performed sub-optimally when using Tweet text, with no benefit when adding user-related features. We believe class imbalance was still a problem, even though we explored two approaches to correct it; other techniques might be more helpful. Additionally, the relatively small number of tweets included in the model collected over a short interval might have contributed to the problem.

## References

[Can19] Statistics Canada. MS Windows NT kernel description. https://doi.org/10.25318/1710000901-eng, 2019. Accessed: 2020-02-23.

[fSAR18] Center for Surveillance and Public Health Agency of Canada Applied Research. Physical activity, sedentary behaviour and sleep (pass) indicators data tool, 2018 edition. https://health-infobase.canada.ca/pass/data-tool/?index=1241, 2018. Accessed: 2020-03-24.

[MHC+15] David J McIver, Jared B Hawkins, Rumi Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P Fitzgerald, Sachin H Jain, and John S Brownstein. Characterizing sleep issues using twitter. *J Med Internet Res*, 17(6):e140, Jun 2015.

[SLGM+09] Fabien Sauvet, Georges Leftheriotis, Danièle Gomez-Merino, Christophe Langrume, Catherine Drogou, Pascal Van Beers, Cyprien Bourrilhon, Geneviève Florence, and Mounir Chennaoui. Effect of acute sleep deprivation on vascular function in healthy subjects. *Journal of applied physiology*, 2009.