Visualization & Text Analysis of Twitter

Comparing Sleep Disorder and Physical Activity Tweets

Team 3: Jenny Kim, Atlanta Liu

Data Science 624-Winter 2020

June 8, 2020

Key Findings of the Project

Key Findings

- The Western provinces show a slightly higher proportion of physical activity tweets and Eastern provinces have higher proportion of sleep disorder tweets. Overall, the proportion of self-reported sleep disorder and physical activity tweets are analogous across Canada.
- Increases in related google web search queries can be weakly associated with higher occurrences of self-reported tweets.
- We achieved an overall accuracy of 71% using a random forest to classify relevant, self-reported tweets.
- Bi-grams comparing Twitter with Reddit revealed a difference in usage of adjectives, with Twitter being more likely to use them when describing physical activity engagement.

Contents

List of Figures

1 Introduction

The government of Canada has recently introduced a Physical Activity, Sedentary behavior, and Sleep Disorder (PASS) surveillance to better understand Canadian health on a population level. Most of the data collected for this framework has been conducted through national-level surveys such as the Canadian Health Measures Survey (CHMS). While this provides a standardized measure of understanding the broader scope of Canadian health, the data is static and does not allow policymakers to analyze or monitor continuous population-level changes in habits.

Social media platforms such as Twitter can be leveraged to understand how the broader scope of Canadian health is changed throughout time and major events. With the use of natural language processing tools, we hope to gain a better insight on the reactions of Canadians as they share and post about perceptions regarding the state of their own physical and mental health. Data triangulation will also be used to help compare our results against other platforms such as Google Trends and Reddit to help validate our inferences. This report aims to supplement the PASS framework by comparing the overall context and trends of self-reported sleep disorders with physical activity tweets through an exploratory data analysis.

2 Methodology

2.1 Research Questions

1. What is the distribution of self-reported physical activity and sleep disorder tweets by province across Canada?
2. Using bigrams, what are the common themes found in self-reported physical activity and sleep disorder tweets for Canadians?
3. How well can we classify the type of self-reported labels using supervised learning algorithms (Naive Bayes & Random Forest Classifier)?
4. Can rises in Google Trends searches be correlated with the frequency of self-reported tweets for either sleep disorder or physical activity?
5. How well does text analysis on web scraping subreddit posts (fitness and insomnia) compare to self-reported Tweets?

2.2 Dataset

Our Twitter data set for both sleep disorders and physical activity is provided in a CSV format, with approximately 6000 and 8800 rows respectively. While the raw twitter dataset has approximately 500 columns, only the name of the city, coordinates, text, and date created were used for our analysis. The time span of our initial dataset included about 3 months, ranging from October to December 2019. A separate data set containing manually labeled self-reported conditions was later added, once the classmates finished with the data entry. For the final project we also included the extra dataset provided to us on D2L, which consisted of an extra 3000-4000 row of labeled tweet posts.

2.3 Methods

Distribution of self-reported tweets across Canada

To begin our data exploration, we mapped the distribution of self-reported sleep disorders and physical activity tweets across Canada through multiple pie charts layered on top of the map. The cities were grouped based on their provinces and territories to analyze the tweet counts for each province and territory. Tweets from New Brunswick, Nova Scotia, and Prince Edward Island (PEI) were grouped as a maritime province and Northwest Territories, Nunavut and Yukon were grouped into territories for this visualization since they had relatively few data points. The proportions were calculated by the taking the total sum of self-reported tweets for sleep disorder and physical activity for each province and divided by the total number of tweets per province. This method allowed us to normalize the percentages so that they would not be drastically skewed due to the population size of a province.

Comparison of Google Trends and Twitter posts

We then investigated the association between the public's search interest in Google Trends and compared it with the frequency of self-reported tweets to see if there was a correlation between a higher number of tweets and a higher number of related search queries. To do this, we linked the raw twitter post with the manually labeled dataset from our class in Python and extracted the dataframe in a CSV format to be visualized in Tableau. Using multiple stacked area charts, we plotted the total number of tweets and label type (Self-report: Yes/No, Sleep Disorder: Yes/No and Self-report: Yes/No, Physical Activity: Yes/No) from October 2019 to December 2019. To compare this with Google Trends, a similar line graph was plotted with the total number of public web searches in Google. The search terms that we compared were "Sleep Disorder" and "Physical Activity" and the location of the search was filtered to Canada.

Supervised Learning (Binary Classification)

Naïve Bayes

As part of the supervised learning requirement for this report, we conducted a Naive Bayes binary text classification on the type of label for each tweet. This was done on both data sets, physical activity as well as sleep disorder related tweets. The Naive Bayes classification involves computing the conditional probabilities of a tweet belonging to either a self-reported tweet and identifying themselves with the condition, or any other labels as explained below.

For example, our binary classification for sleep disorders would involve either "Self-reported:Yes, Sleep Disorder:Yes", or they would belong to the "Other" category. The "Other" category consists of all other labels not of interest, including: "Self-reported: No, Sleep disorder: Yes", "Self-reported: Yes, Sleep: disorder: No", "Self-reported: No, Sleep disorder: No", and "Not Clear". It is important to note that there is a bit of data imbalance present in our binary classification. Approximately 30% of the data was labeled with "Self-report:Yes, Sleep Disorder: Yes", whereas the rest were labeled as "Other". The classification scheme and presence of data imbalance is similar for physical activity dataset. To address this issue, we down sampled the majority group so that we had the same number of tweets in both groups. It should be noted that this is not the most ideal way of addressing data imbalance since we lose out on a lot of information. However, we were unfamiliar with how to up sample text-based data.

After the text for each tweet was compiled into a single data frame with the corresponding labels, the data was preprocessed through Scikit-learn in Python. We conducted a 30% test - 70% train split, where the text data was then compiled into a matrix of features through a term frequency-inverse document frequency (TF-IDF). After fitting it on a multinomial Naive Bayes classification, a simple K-Folds (k=5) cross-validation was used to determine if data imbalance had an impact on the effectiveness of our model. Finally, our trained model was evaluated against the test data based on its accuracy. Results from the classification algorithm were then plotted on a confusion matrix.

Random Forest Classifier

We also decided to use a Random Forest Classifier from Yellow Brick in Python to see if it provided better predictive performance than our Naive Bayes. Following the same 30%-70% test-train split, we conducted a randomized search with a 5-fold, 100 iteration, cross validation on a variety of hyper parameters. Since we didn't have any domain expertise going into Twitter text analysis, we sampled without replacement from a set list of parameters found in Koehsen's (2018) [1] Random Forest Classification guide. The best parameters found in our randomized search is listed below:

- "n_estimators" = 2000, "min_samples_split" = 5
- "min_samples_leaf" = 1, "max_features" = "auto"
- "max_depth" = 100, "bootstrap" = False

The same hyper parameters are used for analysis of both datasets. Afterwards, the tuned model was then evaluated on the test dataset based on the accuracy of its predictions. Results are plotted on a Receiver Operating Characteristic (ROC) curve, which helps to illustrate the diagnostic capabilities of our final classification model.

Text Mining

Web Scraping Reddit

To meet the phase 3 requirement of our report, we applied web scraping applications to compare a different social platform such as Reddit to our exploratory text analysis. Using the PRAW package in Python, both subreddits ("Fitness" & "Insomnia") were scraped to find the first 1000 hot posts. For our text processing, only the body of the post is used to build a bigram model. The dataframe was exported into a CSV file for subsequent text processing in R. While it may be preferable to use the exact search terms for our subreddit such as "sleep disorder" and "physical activity", there is a much smaller community for these subreddits. As such, we chose to look at fitness and insomnia to compare with physical activity and sleep disorder respectively.

Text Preprocessing

To initialize the text in our Twitter data, we first filtered the labels to only include self-reported conditions that the user engaged in (ie "Self-report: Yes, Physical Activity: Yes"). Characters from the body of both Reddit and Twitter posts are interpreted through UTF-8 encoding. Using the TidyText package, we built an initial bigram and began text preprocessing. Commonly used stop words were removed through the built-in package, along with special characters, punctuation, numbers, stems, and context-specific words ("sleep", "asleep", "slept" for sleep disorders). Any words not caught by our stop-word package was manually removed, along with other words that had no apparent relevance to our topic of interest (i.e. tagged usernames). In cases where stemming did not work as intended, selected words were also manually replaced (I.e. "fallen" > "fall").

Regarding the preprocessing for the text data of hot posts in the fitness subreddit, we realized that many of the posts are made by bots. These bots are either reinforcing community guidelines, encouraging routine discussion, or highlighting weekly announcements. As a result, we dove into the term frequency document and manually removed words related to these posts. A more ideal way would have been to filter them out directly at the pandas dataframe level, however,

Bigram Modelling

A total of four bigrams were produced, two from our original Twitter dataset (sleep disorders and physical activity) and two more from scraping through subreddit posts (insomnia and fitness). The number of nodes that were plotted varied by the filter size, which was determined by visual inspection. Some bigrams such as Twitter sleep disorder had sparse matrixes since common words appeared less frequently. As a result, bigrams with more sparsity required a lower filter to display a meaningful number of nodes. By plotting the sleep-related and physical activity-related bigrams side-by-side, we can compare the overall context between the Twitter and Reddit to see if interesting similarities or differences appear between the words used.

Topic Modelling - Latent Dirichlet Allocation

To model the topics associated with tweets in sleep disorder and physical activity, we built a document term matrix (DTM) from our corpus of words for both datasets. Using the same text preprocessing methods as described above, we applied Latent Dirichlet Allocation to split the DTM into two topics, two for sleep disorder and two for physical activity. Each topic listed the top 10 words with the highest beta values. A higher beta value indicates a higher probability of that word being associated with the topic (the word occurred more times in that topic). These beta values were then extracted as a dataframe to be visualized as a bar plot dendrogram in Tableau to complete our project requirement for an advanced visualization.

Several design principles are combined together to help improve the clarity of this visualization. The length of the bars in each branch reflect the magnitude of the beta values. Additionally, since several of the same words appeared in both topics, we altered the color scheme to reflect their presence. Redundant words (such as "night", "hour", "sleep", etc.) are revealed as dark blue (sleep disorder) or dark red (physical activity), whereas the rest of the words followed a continuous color scale. Adding on a black background also allowed us to emphasize the colors used for each word.

3 Results

3.1 Mapping Canadian Distribution of Tweets

Our pie-chart map visualization (Figure 1) displays the distribution of the proportion of self-reported tweets on sleep disorder and physical activity across Canada. Looking closer, we see that the pie charts are quite similar to one another when compared across all the provinces and territories in Canada. Physical activity ranges from 42%- 56% and sleep disorder tweets ranges from 43%-57%. Moreover, self-reported sleep disorder tweets are higher in the Eastern Canada region (Ontario, Quebec, Newfoundland, and Maritime provinces) and self-reported physical activity tweets are higher in the Western Canada region (Alberta, Manitoba, Saskatchewan, and Territories).

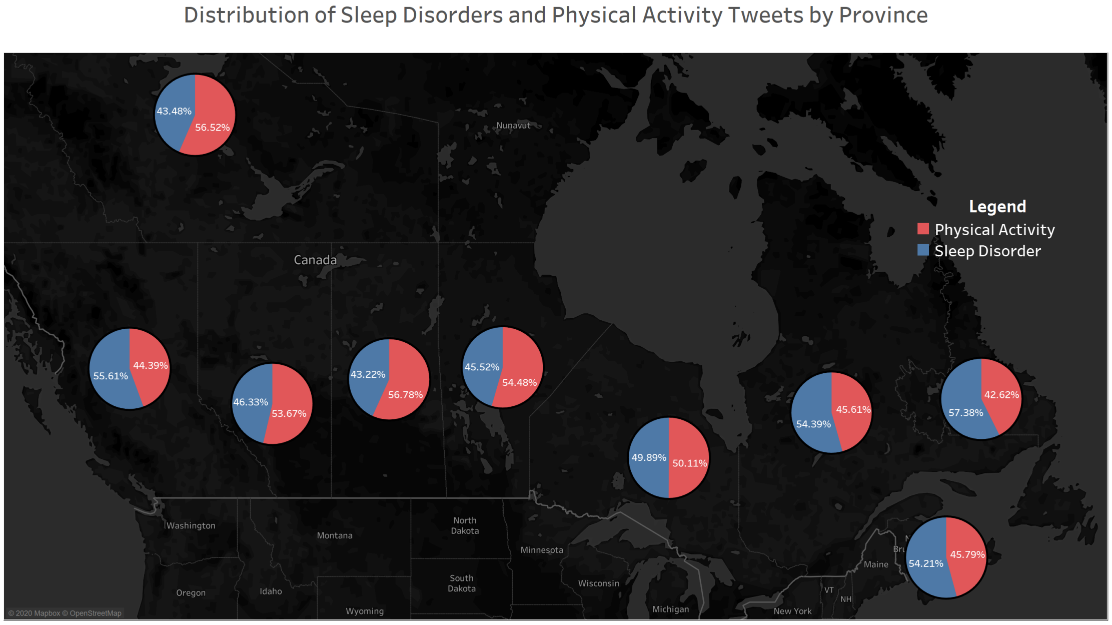


Figure 1: Distribution of self-reported tweets for each condition

3.2 Twitter Posts & Google Trend Comparison

For our Google Trends comparison as shown in Figure 2, the number of tweets on sleep disorder and physical activity were compared weekly over the period of three months (October to December) by counting the number of tweets of each category of the labels. Within the span of three months, both behaviors had two large jumps in the on October 27 and November 17 due to increases in the number of tweets. When the number of web searches of sleep disorder and physical activity in Google was compared, the number of web searches for both terms were high on October 13, November 3 and 17. When the number of tweets is compared to the number of web searches in Google, we can observe some correlation between the public's search interest and their behavior on Twitter. Thus, this infers that higher public web search of these two terms is also associated with higher tweets of sleep disorder and physical activity.

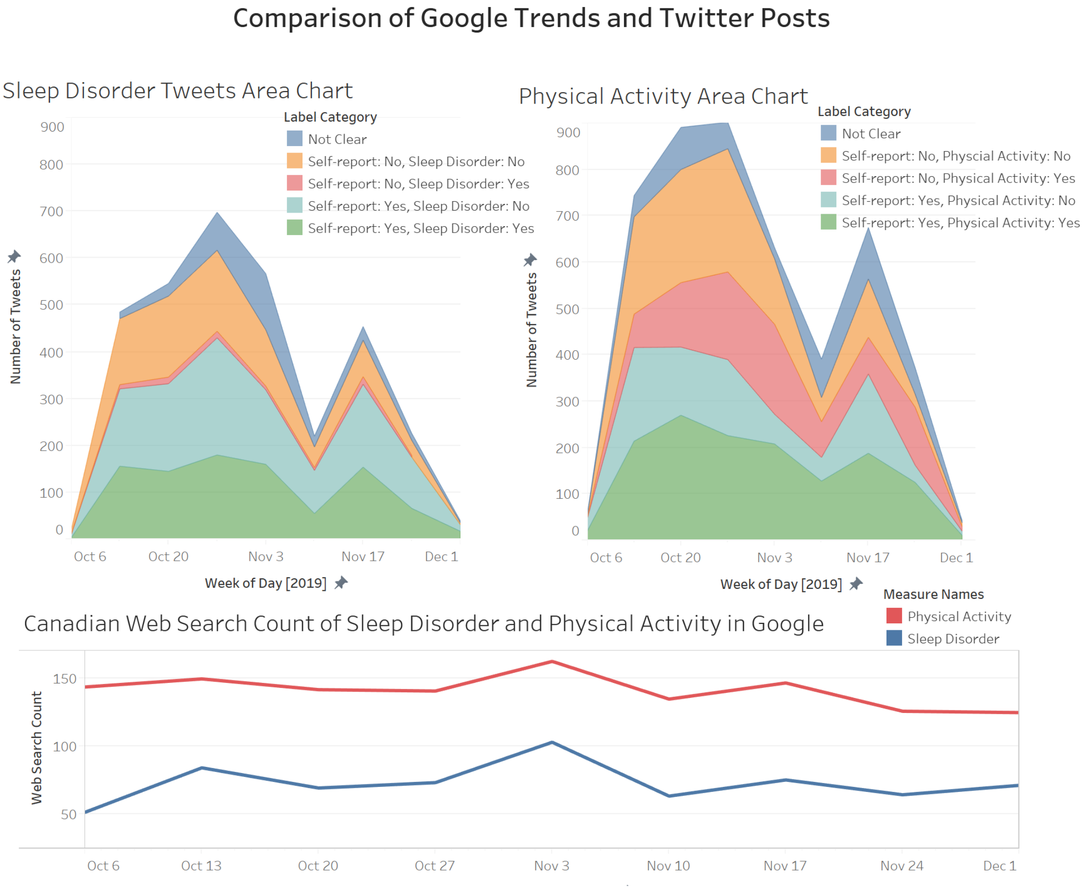


Figure 2: Temporal trend analysis over three months

3.3 Unsupervised Learning - Bigrams

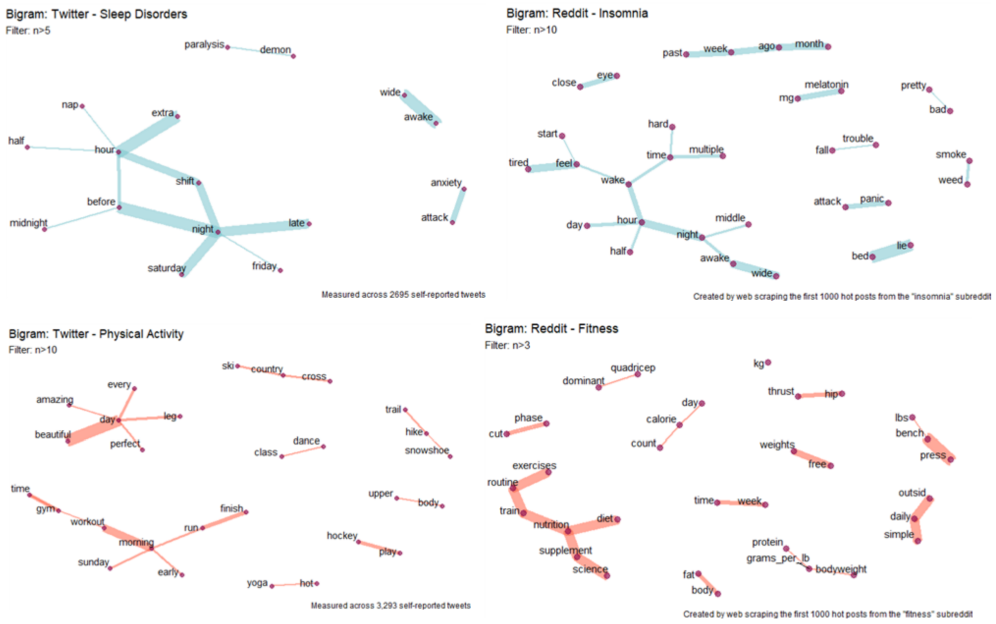
3.3.1 Sleep Disorder (Twitter) vs. Insomnia (Reddit)

Comparing the two bigrams, sleep disorder tweets (left) and insomnia subreddit posts (right) as shown in Figure 3a, both platforms had few similarities in frequently occurring words which were "hour", "night", "attack", "wide", and "awake". Moreover, words that are commonly used together in both Reddit and Twitter included "wide" and "awake". Twitter bigram is based the topic sleep disorders; thus, this bigram includes a wide variety of topics that influence sleep such as "night shift", "nap", and "anxiety". In contrast, Reddit bigram only explores the topic insomnia and as a result, the words are less sparse, and the set of words are used much more frequently compared to Twitter bigram. As well, some words are highly related to insomnia such as "melatonin". Both platforms display a combination of words that are linked with disorders such as "panic" - "attack" and "anxiety" - "attack". These combinations of words are associated with sleep disorder symptoms and highlights commonly used words on Reddit and Twitter that is related to sleep disorder behaviors.

3.3.2 Physical Activity (Twitter) vs. Fitness (Reddit)

Diving into our second set of bigrams, Figure 3b displays Twitter physical activity (left) and Reddit fitness (right). We see that there are different overarching topics discussed across both platforms. The Twitter bigram in physical activity includes a large variety of sports and athletic activity such as yoga, hockey, running, and dance. There is also a much greater use of adjectives (“amazing”, “beautiful”) in Twitter, indicating a stronger presence of individuality across each post. On the other hand, much of the context in Reddit involves weightlifting and workout at the gym. Discussion about nutrition and physical health is much more prominent in Reddit, particularly regarding self-image (“body fat”, “bodyweight”). There are less adjectives used here as the terms used tend to be much more specific and neutral in connotation. Looking at the similarities, both set of bigrams tend to involve time of day or week (I.e. “morning”, “time”, “week”, “day”) in the nodes. Both platforms also discuss specific types of workouts (“upper body”, “bench press”) albeit to different extents.

Bigram Dashboard - Twitter vs. Reddit



Bigram: Twitter - Sleep Disorders
Filter: n>5

Measured across 2695 self-reported tweets

Bigram: Reddit - Insomnia
Filter: n>10

Created by web scraping the first 1000 hot posts from the "insomnia" subreddit

Bigram: Twitter - Physical Activity
Filter: n>10

Measured across 3,293 self-reported tweets

Bigram: Reddit - Fitness
Filter: n>3

Created by web scraping the first 1000 hot posts from the "fitness" subreddit

Figure 3a: Bigram comparison of Twitter sleep disorder (n > 5) vs. Reddit insomnia (n > 10) posts.

Figure 3b: Bigram comparison of Twitter physical activity (n > 10) vs. Reddit fitness (n > 3) posts.

Caption: Comparison of bigrams between Twitter and Reddit based on topic. There does not appear to be a substantial difference in the terms used comparing sleep disorders and insomnia posts. Comparing physical activity and fitness, we see that Reddit terms have a more neutral connotation, whereas Twitter has more uses of adjectives. Differences in word usage across social platforms is reflected by the style of writing in a post.

Figure 3: Cross-platform word usage comparison

3.4 Supervised Learning - Binary Classification

From our Naive Bayes binary text classification as shown in Figure 4, our mean K-Fold (k=5) cross-validation score was approximately 71%. Similarly, the accuracy of our final model is also approximately 71%. In comparison, our Random Forest Classifier (Figure 2) received a 78% accuracy and an F1-Score = 54%. The addition of a cross-validated randomized search improved our accuracy by approximately 2%, increasing from 76% to 78%. The same number of folds are used in both algorithms. Overall, we see that the Random Forest Classifier outperforms the Naïve Bayes in terms of accuracy.

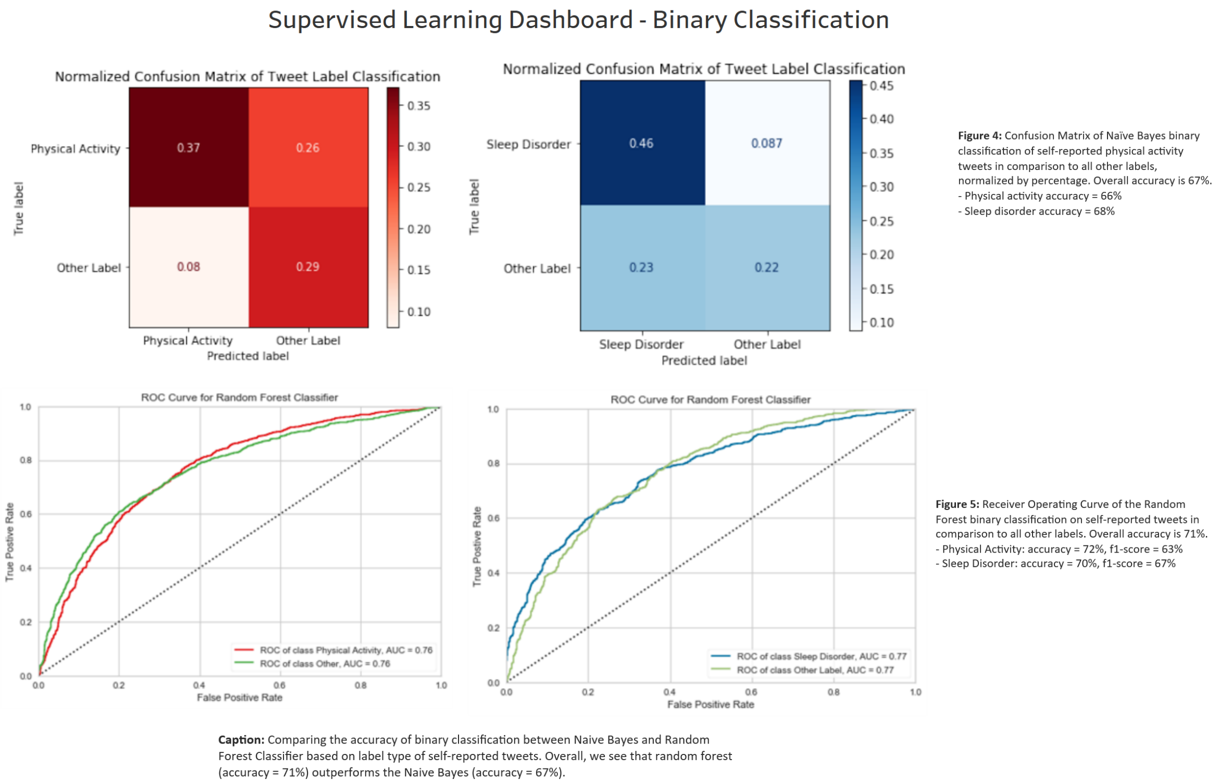

Supervised Learning Dashboard - Binary Classification

Figure 4: Confusion Matrix of Naïve Bayes binary classification of self-reported physical activity tweets in comparison to all other labels, normalized by percentage. Overall accuracy is 67%.
- Physical activity accuracy = 66%
- Sleep disorder accuracy = 68%

Figure 5: Receiver Operating Curve of the Random Forest binary classification on self-reported tweets in comparison to all other labels. Overall accuracy is 71%.
- Physical Activity: accuracy = 72%, f1-score = 63%
- Sleep Disorder: accuracy = 70%, f1-score = 67%

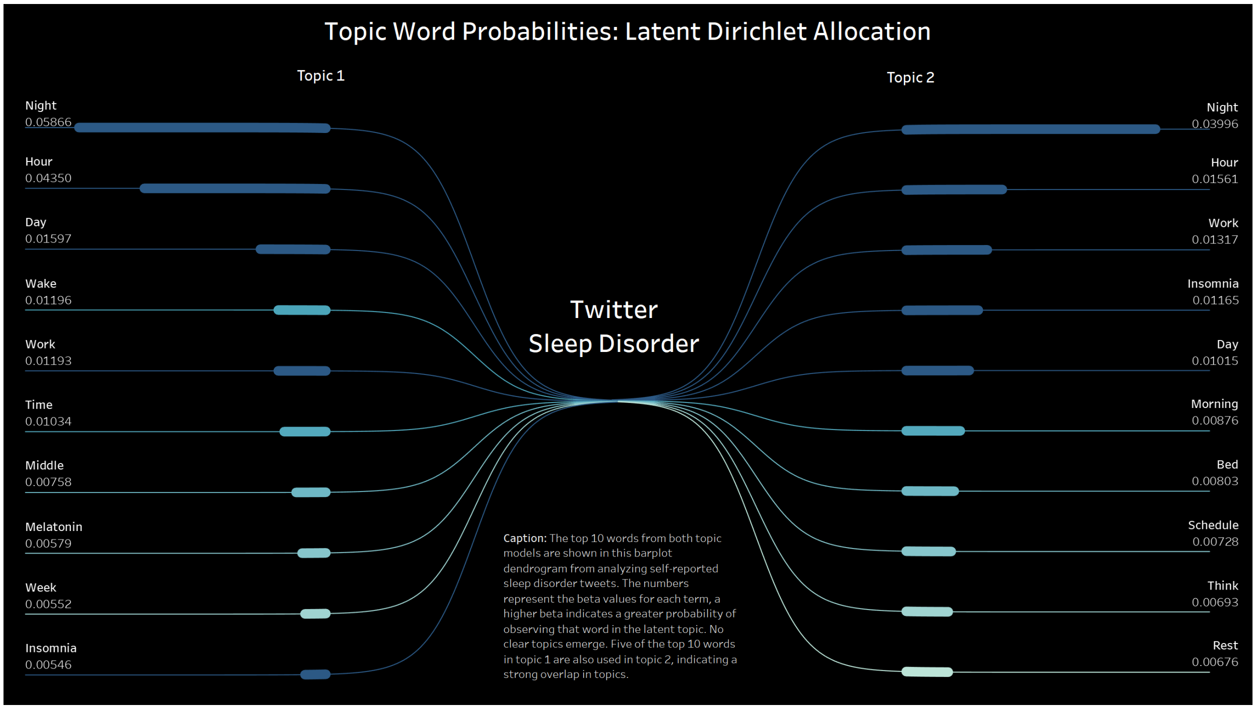
Caption: Comparing the accuracy of binary classification between Naive Bayes and Random Forest Classifier based on label type of self-reported tweets. Overall, we see that random forest (accuracy = 71%) outperforms the Naive Bayes (accuracy = 67%).

Figure 4: Naive Bayes & Random Forest Classifier

3.5 Topic Modelling - Latent Dirichlet Allocation

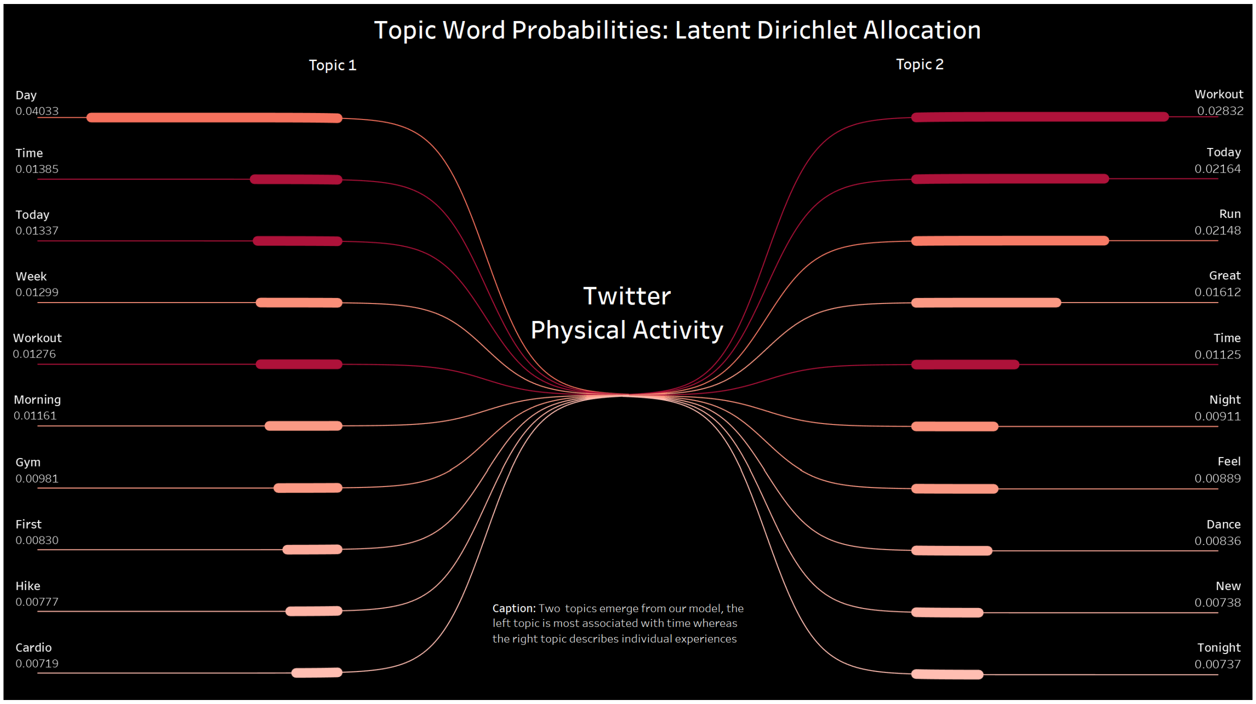
Finally, we have our bar plot dendrogram as shown in Figure 5 and Figure 6. Looking at sleep disorders, we see that there are five words that appear in both topic, including: "Night", "Hour", "Day", "Work", and "Insomnia". The presence of such a large overlap in the number of words suggests that there is not a clear distinction between the two topics. As a result, applying LDA for this topic does not provide any meaningful insights. However, when we look at the LDA in physical activity, we see that there are only 3 words that overlap ("Time", "Today", and "Workout").

Focusing our attention on the first topic, we see a several words relating to the time spent engaging in a physical activity (Morning, Week, First). We can infer that the first topic of words relates to how long or when individuals prefer to exercise. The second topic has several words associated with the subjective experiences of engaging in physical activity, such as "Great", "Feel", or "New". We can also infer here that people on twitter tend to use words to describe their experiences while posting about physical activity.


Topic Word Probabilities: Latent Dirichlet Allocation

Caption: The top 10 words from both topic models are shown in this barplot dendrogram from analyzing self-reported sleep disorder tweets. The numbers represent the beta values for each term, a higher beta indicates a greater probability of observing that word in the latent topic. No clear topics emerge. Five of the top 10 words in topic 1 are also used in topic 2, indicating a strong overlap in topics.

Figure 5: Sleep Disorder LDA


Topic Word Probabilities: Latent Dirichlet Allocation

Caption: Two topics emerge from our model, the left topic is most associated with time whereas the right topic describes individual experiences

Figure 6: Physical Activity LDA

4 Limitations

The Twitter dataset had information on when the tweets were posted. Therefore, we attempted to conduct a long-term temporal analysis to look at how behaviors changed over time on Twitter. However, there were missing timestamps in sleep disorder and physical activity datasets and these timestamps did not align when the two datasets were compared to each other. Another limitation is the lack of subreddits on the topics of sleep disorder and physical activity. Not many users discuss on the topics of ‘Sleep Disorder’ and ‘Physical Activity’ therefore, search terms ‘Insomnia’ and ‘Fitness’ were studied instead for our text analysis.

5 Conclusion

Studying social media platforms is important for analyzing user’s behaviors towards their physical and mental health. In this report we compared self-reported sleep disorder with physical activity tweets in Canada and determined that the two topics behave similarly across Canada. In addition, Twitter was compared with Google to understand the association between the public’s web search and behaviors of sleep disorder and physical activity in tweets. Based on the stacked area chart visualizations we saw that there was a slight correlation between the two platforms. However, more data will be needed here to reach a compelling conclusion.

The results of our binary classification showed that we can classify the type of self-reported labels of each tweet to a respectable accuracy of 71% through a random forest classifier. This indicates the potential of using machine learning algorithms to classify tweets of interest for further research in this area. Our report only used a simple RandomSearchCV, so further emphasis on hyper parameter tuning could bring more improvement to the overall performance of the random forest classifier. We also performed text analysis on tweets and public web forums from Reddit to compare the commonly used words in the two platforms and between each condition of interest.

When we compared Twitter posts with Reddit regarding sleep disorders and insomnia, there does not appear to be a strong difference between the two, apart from Reddit consisting of more terms. On the other hand, Twitter and Reddit indicate a much bigger difference in the words used regarding physical activity and fitness. Specifically, we see that terms used across tweets reflect more individuality as more adjectives are used to describe their experiences. Reddit posts also have a unique attribute in that there is a lot more discussion pertaining to one’s image. Altogether, our analysis reveals that even though the two platforms share some similarities, there can be a striking difference in terminology depending on the topic of interest.

6 Bibliography

1. Koehrsen, W., 2018. Hyperparameter Tuning The Random Forest In Python. [online] Towards Data Science. Available at: <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74> [Accessed 26 March 2020].