Digital Surveillance of Physical Activity in Canada

Phuong Uyen Nguyen, Chee-Lam Tam

Data Science 624-Winter 2021

April 15, 2021

Key Findings of the Project

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

1. **Canadians actively tweet about physical activity throughout the day.** Peak Twitter activity occurs in the evening (5-8pm)
2. **Twitter activity provides a good representation of different provincial physical activity levels.** Based on Twitter data, Quebec had the lowest rate of physical activity and Northwest Territories had the highest rate.
3. **Physical activity trends change over time.** Changes were particularly evident when comparing fitness trends from before versus during the Covid-19 pandemic.
4. **Machine learning models to label physical activity tweets is promising.** User descriptions may be associated with physical activity levels.

1 Introduction

1.1 Project Motivation

People are more prone to sedentary behaviours than physical activities, especially when many have to work from home during the ongoing pandemic. Being inactive can lead to many health problems, such as obesity, cardiovascular diseases, or mental health disorders. Surveillance of people's frequency of physical activity, sedentary behaviour, and sleep quality (PASS) will provide insights for constructing public health strategies that help to raise awareness of adopting good habits to improve people health. With the robust availability of data from social media, using natural language processing (NLP) to analyze people's texts can detect the pattern of PASS more time- and cost-effectively, compared to traditional self-report surveys.

This project supports the strategic goal set out by the Data Intelligence for Health (DIH) Lab's PASS Surveillance project. Particularly, in this project we aim to understand the digital behaviours of physically active Canadians. Also, we will examine if Twitter user information (e.g., profile descriptions and number of followers) can predict their tendencies to tweet about self-reported physical activity.

1.2 Problem Definition

At phase 2, we aimed to build an unsupervised model to explore the most commonly tweeted topics about physical activity and a supervised model that can classify if a Twitter user was actually engaged in an activity or just mentioned it in their tweet. Via topic modelling, we can extract information about the physical activities which were tweeted about the most. This method requires manually filtered or labeled data of physical activity to narrow down the vocabulary and reduce the chances where physical activity related topics being overlooked due to irrelevant words with higher frequencies. Simultaneously building a classification model based on those labeled data will help to accelerate the data collecting process, resulting in more up-to-date data to improve the performance of the unsupervised learning model as well as to catch up with the evolution of the tweet topics over time. Additionally, from these models we can expand our findings to answer the research questions from our client. During phase 3 of the project, we pursued the following goals: identifying the peak time(s) of day when users tweeted about their physical activities, comparing the frequency of physical activity related tweets among provinces to see if there were any visible differences in levels of physical activity levels, analyzing any potential association between the physical activity tweet frequency and count of a user's followers, and detecting common themes in profile descriptions which could be linked to a user's tendency to have at least one physical activity related tweet.

The DIH data set used in our analysis is a selection of 4694 tweets in 2019, filtered by the criterion of containing words that are potentially related to physical activity. Among 717 columns of the data set, the most relevant factors that can help to address our project's objectives are the tweet creation date/time, tweet full text, location/place, user id, user name, user description, and user followers count. The data were labelled by the DIH Lab. Each tweet was manually classified whether it was self-reported and/or about a physical activity.

Public data from Google Trends and government data on PASS surveillance were used for data source triangulation to complement the findings from the Twitter dataset. Google Trends provides numerical scores to indicate popularity of web searches about physical activity topics according to different geographical regions and time spans. National PASS surveillance data in 2018 from the Center for Surveillance and Applied Research, Public Health Agency of Canada [1] reflects the proportion of adults who reported to have at least 150 minutes of physical activity per week. This data will help to evaluate whether our Twitter data can be representative of the Canadian population.

Working on social media data had some challenges. Since we analyzed personal social media posts, relevant data could have been missed if the tweets have spelling errors, shorthand, sarcasms or uncommonly

[1] Center for Surveillance and Applied Research, Public Health Agency of Canada. Physical Activity, Sedentary Behaviour and Sleep (PASS) Indicators Data Tool, 2020 Edition. Public Health Infobase. Ottawa (ON): Public Health Agency of Canada, 2020.

uses of languages which cannot be recognized by text mining and NLP software. There were also limitations in understanding the context in which words are used. Since these problems may hinder the validity of our findings, employing data source and method triangulation is necessary to overcome the challenges.

2 Methodology

2.1 Data Analysis

Unsupervised information extraction

Latent Dirichlet Allocation (LDA) was used to discover the top physical activities that were often tweeted about. This technique can be applied to this data set because it is a simple and effective way to find and group the most common topics. Since LDA is a probabilistic model, the results provide a good indication of the relative importance of the topics in the data set. It also provides information about the common words used in conjunction with the identified physical activity topics, and LDA have often been used in similar NLP projects. We tuned the model with a range of k topics. Once the topics were extracted, we labelled each topic based on the observed results. Since LDA information extraction can result in overlapping topics or inaccurate groupings,tuning k is also useful in selecting the best themes and removing redundant or irrelevant information.

We then used the findAssocs() function in R to extract word associations in the document term matrix, based on the topic labels discovered through LDA. This provides a named list, where each list component is named after a term and contains a named numeric vector. Each vector holds matching terms with their rounded correlations. To obtain only the highly correlated words, the lower correlation limit was set to 0.1.

To examine how the Twitter users often described themselves, we used R to convert the user description text into a text corpus which would be cleaned and stemmed before our text mining steps. Other than the common stop-words, after manually assessing the words and their frequencies, we created an additional word list to remove from the documents. Those words were temporal terms, spatial locations, or related to personal contact information and thus irrelevant to what we wanted to analyze in the user description. The description written by those who had at least one self-reported physical activity related tweet over the time period covered in the data set were compared with that from the users with no physical activity related tweet at all. How often the physical activity related keywords (specified in the result session) appeared in the two groups can reveal an association between the tendency to self-report at least one activity and the pattern in which people identified themselves. Since different numbers of users in the two groups can cause inflation in the frequencies of the extracted terms, instead of the raw count of each word we used the word's percentage of total occurrences. This allows for a more accurate comparison in case where, for example, we would find a word 100 times out of a total of 200 word occurrences within a group versus 200 times out of 1000 word occurrences in another group.

Supervised information extraction

Based on the labels provided in our data set, we created a new binary variable to classify if a tweet was about self-reported physical activity. Since we are interested in whether people were engaged in physical activity, the main task is not only detecting keywords about activities being discussed on the social media platform, but monitoring if people claimed to actually carry on the action is also essential. Hence, only tweets with a label of "Self-report: Yes, Physical Activity: Yes" received value "1" in that new variable, and the remaining tweets were marked as "0", meaning that the Twitter user was not doing the activity being tweeted or the tweet was not relevant to any physical activity at all.

The tweet texts were vectorized using the CountVectorizer() function from the scikit-learn's text feature extraction package. This function transformed each tweet text into a data frame consisting of 500 columns, which recorded the presence of top 500 highest frequency words within the text. These words must not contain those that appeared less than 3 times or more than 90 percents of the text as well as those that are not informative for the context of physical activity and profile description (i.e., stop-words). We manually constructed the stop-words and tuned the model to determine the best values for the minimum document frequency, maximum document frequency, and maximum highest frequency feature parameters. The tweet text and binary self-report activity variables were then ravelled into arrays and used to train several classification models that could predict if a tweet referred to a self-report physical activity. The performances of those models were then evaluated. Note that although TF-IDF (term frequency-inverse document frequency) analysis is better than count vectorizer for the ability to capture the rare words which can significantly improve our model, the function could not improve our model performance. Therefore, only the results produced by the CountVectorizer() function will be reported.

Besides identifying the self-reported, physical activity related tweets, we also built two more classification models to predict whether a user would post at least once (1) or would not post anything at all (0) about a physical activity. In one model, we use the counts of user's followers, favourites, friends, listed groups, and statuses as our predictors. For the other model, we vectorized the user description texts using the CountVectorizer() function so that the predictors of this model were the frequencies of words found in the user profile description. Both models were built on the scikit-learn's random forest classifier algorithm.

To optimize the performances of our classifications models, the imbalanced class weights was addressed. For each model being tested, we also examined if stratifying and/or Synthetic Minority Oversampling Technique (SMOTE) could alleviate the imbalance and thus improve the performance. Data were partitioned; 80 percent of them were used to train and cross validate, and the rest were set aside for final testing steps. Via cross validation and grid searching (on the train and validate sets), we tuned the hyperparameters and evaluated the model performance.

Data aggregation by province

The 'place' information in the Twitter data set was cleaned so that a new dimension could be created for province. Place names in the data set could refer to cities, points of interest, or neighbourhoods. Most of the places had the province names listed in addition to the place names. Where the province was not provided, it would be manually added by using the Twitter user's location. If their location still did not provide the province, then the province would be inferred by manually determining where the city or landmark is located. There were ten remaining tweets where the province was unknown and these tweets were filtered out for the geographical analysis.

To account for differences in population in each province, the relevant numerical data needed to be normalized by dividing the value by the provincial population in Q4 2019 according to Statistics Canada. For example, the number of tweets per province would be represented as number of tweets per 100,000 people living in the province.

Google Trends provided data source triangulation to support Twitter findings about activity level by province. The Google Trends interest by geographical subregion data was filtered to match the time span of the Twitter data set (October 12 to November 30, 2019). The data provided a score of the relative popularity for web searches about physical activity in each province. This data was joined with the Twitter dataset so that each province would have and activity level score according to Twitter as well as the popularity score from Google Trends. The Google Trends data are provided in files "geoMap.csv" and "relatedEntities.xlsx".

Temporal analysis

The temporal data analysis was based on the tweet creating date and time. The date and time stamps in the raw data were in Greenwich Mean Time and needed to be converted to the correct time zones. Since Canada observes six different time zones, the conversion of the timestamp would depend on the province. A calculated field in Tableau was created so that the tweet timestamp was in the local time zone according to the province of the tweet. To represent the data at an appropriate level of detail, another calculated field for the hour (0-24) of the local time stamp was required.

Analysis of physical activity trends before and during the Covid-19 pandemic

The Google Trends related entities dataset provided a list of topics related to physical activity with the corresponding score of search popularity. The data was filtered to compare two time periods. The first time period of October 12 to November 30, 2019 aligns with the Twitter data set time span and represents web search trends before the Covid-19 pandemic. The second time period of March 11 to April 30, 2020 represents the beginning of the Covid-19 pandemic.

User's number of followers by number of physical activity related tweets

To see if users with large numbers of followers would tweet more about physical activities, we visualized the average numbers of user's followers by their counts of self-reported activity tweets. This also allows us to detect any potential association between the two variables, which could help to guide the variable selection process for our supervised machine learning models.

Validation of Twitter physical activity data using national PASS surveillance data

Since not every Canadian has a personal Twitter account and not all Twitter users want to report every single physical activity, it is essential to evaluate whether the patterns of our findings are consistent with other reliable data sources. With an assumption that every time someone tweeted about a physical activity was when they actually exercised, we want to compare the proportions of active and inactive people to what had been reported in the national PASS surveillance data by the Center for Surveillance and Applied Research for 2018, which was the closest year to our Twitter data that we could find. Particularly, we expect the proportion of Twitter users who had at least one physical activity tweeted during October and November in 2019 to be relatively similar to the average proportion of active people who were engaged in at least 150 minutes of physical activities in 2018.

2.2 Visualization

The results from the unsupervised information extraction are visualized as term versus beta plots for each topic. The visualizations were first generated in R to perform exploratory analysis and observe results in order to select the best number of k topics, determine if any additional data cleaning was required, and see if there were any overlapping or irrelevant topics. By visualizing the 15 most important terms per topic, we could easily read the word groups of the most common physical activities. A selection of the topics were re-plotted as term versus beta plots in Tableau so that only the findings relevant to physical activity are presented.

The user's descriptive words were visualized as word clouds in which the size and color of a word indicated how often the word has occurred. After preliminary data cleaning and exploration process in R, the words and percentages of occurrence were exported into Tableau to generate two interactive word clouds, which illustrate the differences between users without any physical activity tweets and those with at least one tweets, in terms of how they had identified themselves. To make the differences more visible, the most important words and their percentages of occurrence were displayed as bar charts, which can provide a more effective size contrast than word clouds.

The performance scores of the supervised classification models to detect physical related tweets are visualized as a scatter plot with marginal density plots in R and Tableau, with a table of accuracy and Receiver Operating Characteristic/Area Under the Curve (ROC AUC score) produced in Python source code for each model. For models that predict whether a user will have at least one or no physical related tweet at all, these metrics scores are reported in text and confusion matrices produced by the best algorithm. A slight modification in the presenting of the confusion matrices is that we use packed bubble charts instead of the classical matrix tables, since the sizes of the bubbles can quickly provide comparison between correct predictions and false positive/negative.

The geospatial tile grid map presents the physical activity level per province. The colour gradient represents the proportion of tweets which are about physical activity. To enable this visualization, a csv file ("hexmap-canada.csv") was created to code the location of each province tile, which would later be joined to the main dataset.

The bar chart is used the compare the relative interest in Google web searches about physical activity topics. The bars are sorted in descending order, with the first bar representing the province having the maximum score of 100, and all other values are assigned relatively to the score of that first province.

The clock face visualization shows the number of tweets about physical activity throughout the day. This visualization technique suggests to the audience that the data is temporal and has been aggregated by hour. Two concentric rings of data points represent two 12-hour periods. Each circle marker provides data for one hour in a day. The location of each circle marker was determined by calculated the x and y coordinates. The daytime hours are coloured yellow and the nighttime hours are coloured blue. The size of the circles correspond to the number of tweets.

A scatter plot of follower counts by user's numbers of physical activity tweets was chosen to illustrate the association between the two variables, because the data points can visualize any potential trend of how a variable changes in accordance with the other. Each data point represents the average amount of people who followed the users with the corresponding number of physical activity tweets. The color hue of a data point was coded by the amount of users having that number of tweets.

To compare and validate our finding with other data source, we calculated the proportions of people with and without physical activity tweets and compare them with the proportion of Canadian adults who met physical activity guidelines by accumulating at least 150 minutes of moderate-to-vigorous physical activity each week, in bouts of 10 minutes or more, in 2018. For that purpose, another bar chart visualizes the magnitudes of those proportions and thus helps to evaluate if the discrepancy between the two data sets is noticeable.

3 Performance Measurements

The unsupervised topic modelling requires the LDA model to be run iteratively, for which we need to manually tune the number of k topics and identify additional words that are considered stop words in the context of physical activity tweets. A value of k which is too small will provide overly broad topics, and a value of k which is too large will result in similar topics. For this project, we assessed the models over a range of k between 5 and 25, and determining the best number of k topics based on the observed groupings of terms. A perfect solution would provide well-distinguished topics, in which the term having the highest beta value in each topic describes a type of physical activity and the top 10-15 words in the topic are related to that activity. The topics should neither be too broad nor too specific. We also want to ensure that the top terms in each topics are meaningful by manually identifying and removing stop words. We determined that k=20 provided the best model.

To choose and evaluate our supervised classification models, we extracted the average accuracy and the ROC AUC scores from 10 fold of stratified cross validation and/or the metrics produced by the test data. The higher these metrics (especially the ROC AUC score) are, the more correctly a model can detect physical related tweets or predict the user's tendency to tweet. For the two models that use the frequencies of the most occurring words (in all tweet texts or all user description) as predictors, we manually tuned the models by adjusting the word lists to remove and include, in order to maximize the metrics scores.

4 Results

4.1 Characteristics of Physically Active Canadians

Physical Activity Level Per Province/Territory

Ontario had the highest number of tweets about physical activity (1055 tweets) and the Northwest Territories had the highest normalized number of tweets about physical activity (15.5 tweets per 100,000 people). It can be interpreted that the greatest amount of physical activity occurred in Ontario, but the greatest proportion of residents of Northwest Territories were physically active. Yukon Territory had the lowest number of tweets about physical activity (3 tweets) and Quebec had the lowest normalized number of tweets about physical activity (1.5 tweets per 100,000 people).

This finding is supported by Google Trends relative interest scores. Extracting web search data from the same time period as the Twitter data set (Oct-Nov 2019), most of the provinces/territories had fairly high interest levels. Quebec and Nunavut had relative interest scores below 50.

Time of Tweets

Twitter users actively tweeted about physical activities throughout the day. The least amounts of Twitter activity were observed from midnight to 5 am. From 6 am, the Twitter activity increased, peaking from 5-8 pm, and then ramping back down.

Physical Activity Trends

The Google Trends related entities data suggest that Canadians approached physical activities differently in Q4 2019 (before the Covid-19 pandemic) and late Q1 / early Q2 2020 (at the beginning of the Covid-19 pandemic). The data suggests that in 2019, web searches about physical activities taking place in public or group settings were increasing in popularity. The rising topics included ’Fitness boot camp’, ’Orangetheory Fitness’, and ’Physical medicine and rehabilitation’. It also suggests that in 2020, web searches about working out from home increased in popularity. The rising topics included ’Bodyweight exercise’, ’Chloe Ting’, ’Insanity’, ’Fitness Depot’ and ’Calisthenics’.

The percentages calculated for Twitter "physically active" and "physically inactive" users reveal that during the period of October to November of 2019, 31 percents of our data had tweeted at least once about carrying on a physical activity, while 69 percents had not had any activity related tweets.

4.2 Information Extraction

Unsupervised Topic Modelling

The LDA using 20 topics provided the best results. The preliminary output is visualized in Figure 1. The final output is visualized in the Exploratory Visualization section below.

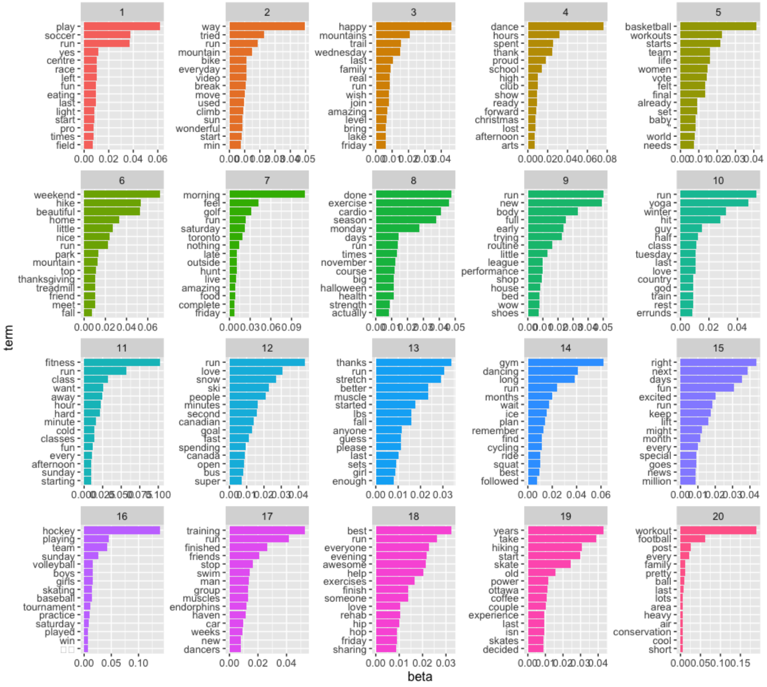


Figure 1: Preliminary results of the LDA information extraction

The unsupervised information extraction resulted in twenty topics. Analyzing the 15 most important terms (highest beta value) per topic allows us to name the most common physical activity themes. Nine out of the twenty topics were selected as the best representation of the physical activities which were tweeted about the most. The remaining topics were either redundant or irrelevant to physical activity. We could further group the nine topics into three categories. The final categorization and topics are:

- Fitness: Gym, Running, Training
- Sports: Soccer, Football, More Team Sports
- Fun and Leisure: Dancing, Mountain Activities, Fall Hiking

Extracting the word associations to the above listed topics resulted in common words about physical activity, shorthand, slang words, and emojis. Below is a list of some examples of the highly correlated words to the topics.

- Gym: "regular", "gymselfie", weighlifting - emoji
- Football: "communities", "nfl", football - emoji
- Fall Hiking: "mountain", "lake", "happythanksgiving", mountain - emoji

Words that were highly correlated with football imply that the related tweets were about both playing and watching football. The Twitter users often mentioned the names of teams or players, snack foods, and emotions. This information can also be extended to the other sports which were popular in tweets, such as soccer, hockey, volleyball, baseball.

Tweets about health and fitness imply some regular routine or schedule, whether it was running, going to the gym, or other type of training. The gym associated tweets included many emojis of facial expressions or body parts, different types of exercises, and selfies.

The terms in the fun and leisure category are related to communities (schools, clubs, friends) and family. Those who tweeted about fall hiking mentioned park names and natural landscapes.

Some topics were excluded from the presentation of results as they were unrelated to physical activity but were associated with politics (i.e., running for office) or daily tasks (i.e., running errands).

Supervised Classifications of Self-Reported Physical Activity Related Tweets

The table in Figure 4 shows the ROC AUC and accuracy scores calculated for different machine learning methods that we have tested to classify self-reported physical activity, and the scatter plot shows these metrics scores produced in each stratified cross validation fold for the top four algorithms with the best performances. Among all, logistic regression and random forest appear to be the best algorithms with the highest ROC AUC scores of 0.79 and accuracy score at 0.78 on the testing data, and the mean metrics scores computed for the individual validation sets of these algorithms mostly cluster at the upper right corner of the scatter plot. Additionally, logistic regression and random forest classifiers were computationally fast compared to the support vector machine and extreme gradient boost classifiers, which are not too far behind in term of performance.

Supervised Classifications of User's Tendency to Have At Least One Physical Related Tweet

Our clients want to see if it is possible to predict how many physical activity related tweets a user would have, using the number of followers they had. At first, we expected that fitness influencers who had lots of followers would have the highest numbers of tweets. However, this is not the case, as we can see that the users who were followed the most actually tweeted about physical activity from 0 to 3 times, whereas the only one account with 13 tweets only had 302 followers. When we look at the number of tweets as a binary variable, we see that only one third of users had more than one physical activity tweet, and the majority had none. This is not surprising as from the national PASS surveillance data in 2018, there are much less people who exercised regularly than those who did not.

For predicting user's tendency to tweet at least once about their physical activity, random forest classifier out-performs simple logistic regression. Unfortunately, the ROC AUC and accuracy scores from our best random forest models are much lower than those of the best models that detect self-reported physical activity. The ROC AUC scores are 0.58, while the accuracy score is 0.68 for model that uses quantitative variables of user information (e.g., number of followers, likes, etc.) and is 0.64 for the model that uses profile description as predictors. Additionally, the confusion matrices show that both models were better in classifying users with no physical activity tweet than those with at least one tweet.

Descriptive Words by Users With (1) and Without (0) At Least One Physical Activity Related Tweet

The last Figure shows two word clouds of profile description of users with and without physical activity related tweets, which were made not by word frequency but by how many percents a word occupied in all user description, within a user group. The color shade and size of a word represent its percentage or weight. For example, "fan" was the most used word, which accounted for 18 percents of the words used in the description of user with no physical activity tweet and 16 percents in the other counterpart group.

Most Twitter users identified themselves by disclosing their family roles, genders, nationalities, habits, personal interests, and professional occupations in their profile descriptions. In general, many common words used in the profile descriptions were seen in both groups of users, including indoor/outdoor activities and interests in travelling, photography, food/cooking/baking, movies, and sports. The word occurrence changes between the two word clouds. To further evaluate the user descriptions, we chose some physical related keywords to represent the physical activities, which are more likely to be carried on than passively watched on TV or at a stadium, such as "run"/"runner", "gym", "golf", "dance" and "bike". In a closer look into how often these words appeared in each group's descriptions (using word percentages of total occurrence), the bar charts in the last dashboard reveal that there were some discrepancies in the appearances of those physical activity keywords. These words account for 1-6 percents in the no-tweet group, but for 2-15 percents in the at-least-one-tweet group. Indeed, the percentages of the chosen keywords' frequencies found in the description of users with one or more physical activities self-reported were at least two times higher than those found for the description of users with no physical activity tweet.

4.3 Exploratory Visualization

Digital Surveillance of Canadian Physical Activity Levels: This dashboard displays the provincial activity levels based on Twitter data, the interest level in physical activity by province based on the relative interest score from Google Trends, the time of day when people tweet about physical activity, and a list of the common physical activity themes on Twitter. The visualization techniques used were tile chart, bar chart, and radial chart.

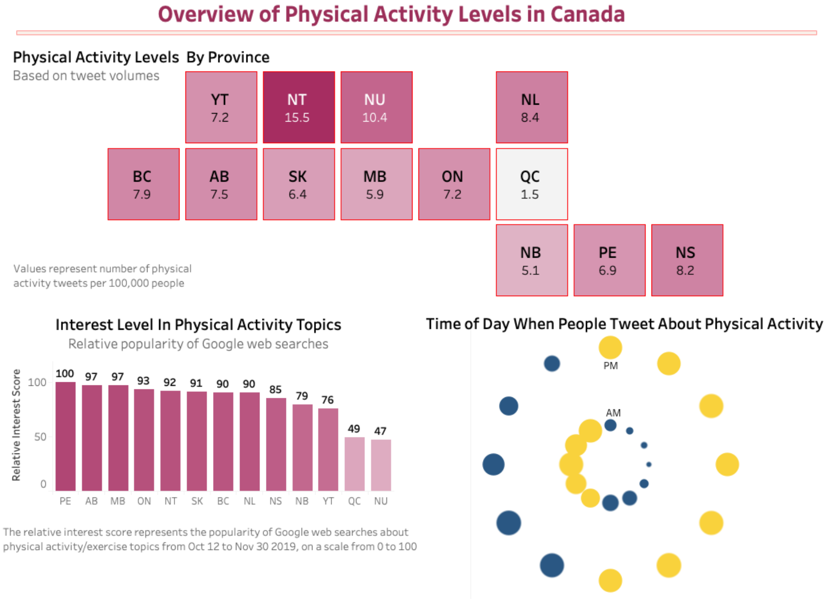
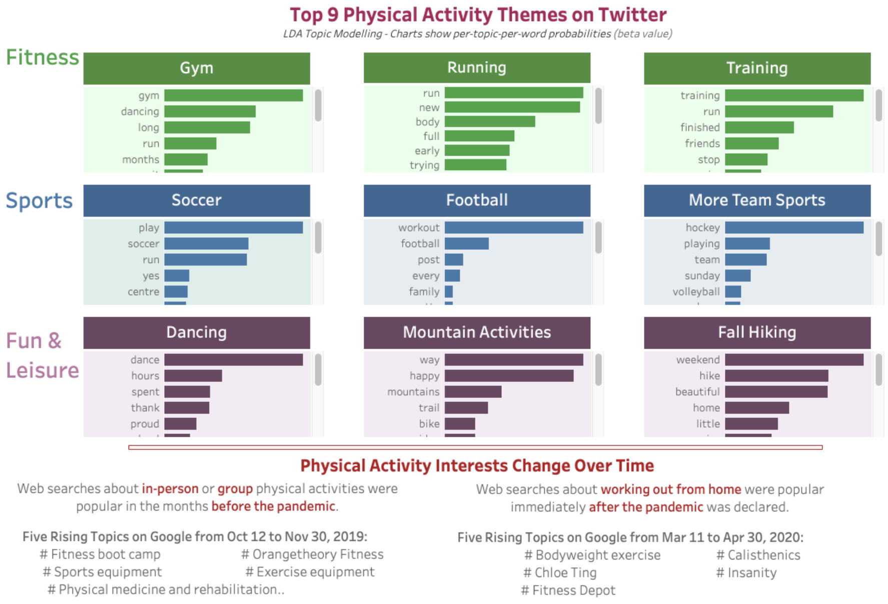


Figure 2: Overview of Canadian Physical Activity Levels

Top 9 Physical Activity Themes (LDA Topic Modelling): The physical activity themes are visualized as separate term vs beta plots per topic. The horizontal bar charts were generated in Tableau based on data generated in R. The Google Trends rising topics are listed to compare the trends in late 2019 and early 2020.



Top 9 Physical Activity Themes on Twitter

LDA Topic Modelling - Charts show per-topic-per-word probabilities (beta value)

Physical Activity Interests Change Over Time

Web searches about **in-person** or **group** physical activities were popular in the months **before the pandemic**.

Five Rising Topics on Google from Oct 12 to Nov 30, 2019:

Fitness boot camp
Orangetheory Fitness
Sports equipment
Exercise equipment
Physical medicine and rehabilitation..

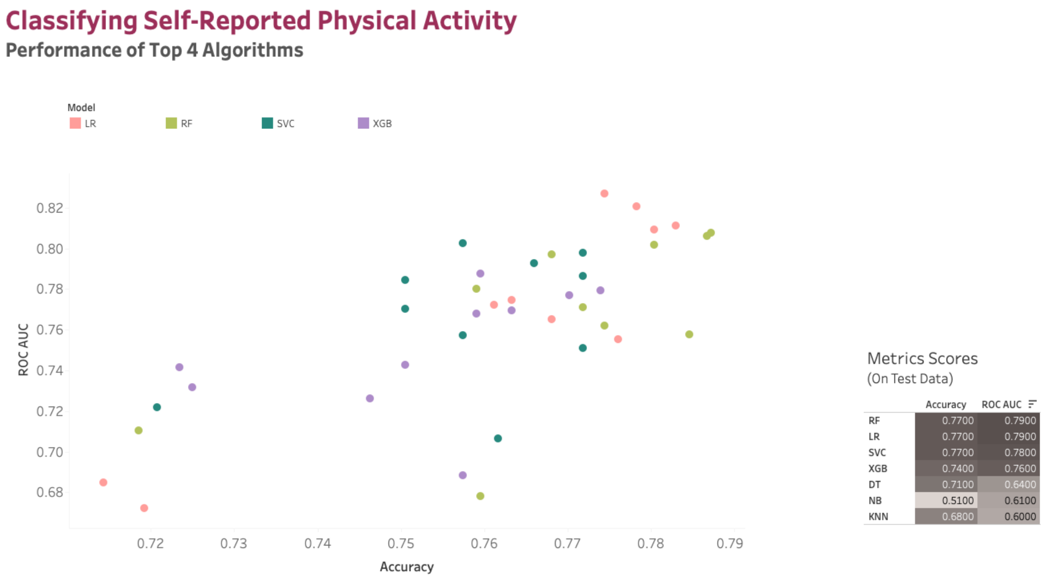
Web searches about **working out from home** were popular immediately **after the pandemic** was declared.

Five Rising Topics on Google from Mar 11 to Apr 30, 2020:

Bodyweight exercise
Calisthenics
Chloe Ting
Insanity
Fitness Depot

Figure 3: Physical Activity Themes

Classifying Self-Reported Physical Activity - Performance of Top 4 Algorithms: This dashboard compares the performances of the top four algorithms used to classify self-reported physical activity tweets. The visualization techniques used were a scatterplot with marginal density plots to illustrate the ROC AUC vs Accuracy as well as a table with the same metrics for all algorithms tested.



Classifying Self-Reported Physical Activity

Performance of Top 4 Algorithms

Metrics Scores
(On Test Data)

	Accuracy	ROC AUC
RF	0.7700	0.7900
LR	0.7700	0.7900
SVC	0.7700	0.7800
XGB	0.7400	0.7600
DT	0.7100	0.6400
NB	0.5100	0.6100
KNN	0.6800	0.6000

Figure 4: Classifying Self-Reported Physical Activity - Performance of Top 4 Algorithms

Average Number of Followers by Number of Physical Activity Related Tweet and Performance of Random Forest Classifier: This visualization helps to see if there is a potential relationship between how many followers a user had had and their numbers of physical activity related tweets. The confusion matrices and other metrics scores of the random forest classification models (which predict the tendency of having physical activity related tweets) are also included in this dashboard. Additionally, the proportion of Twitter users who have posted at least once about their physical activity and the proportion of Canadian adults who were often engaged in physical activities were compared using a bar chart.

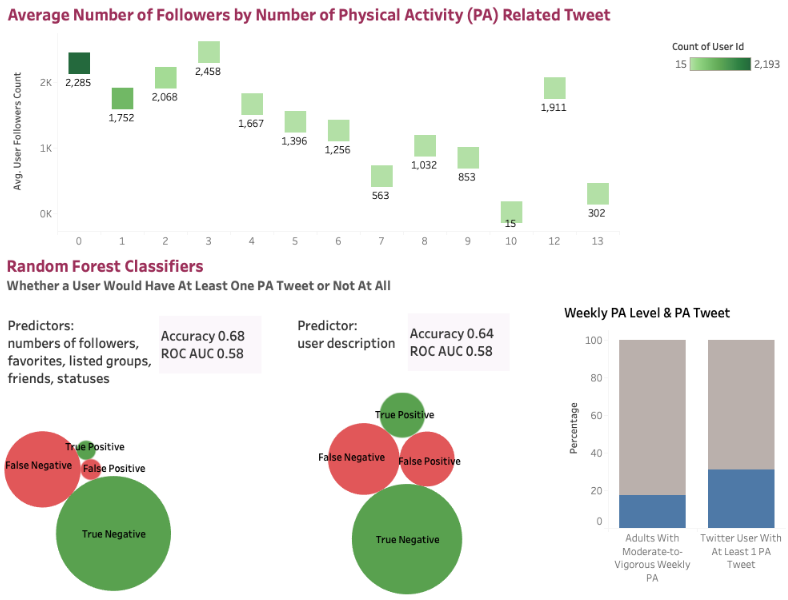
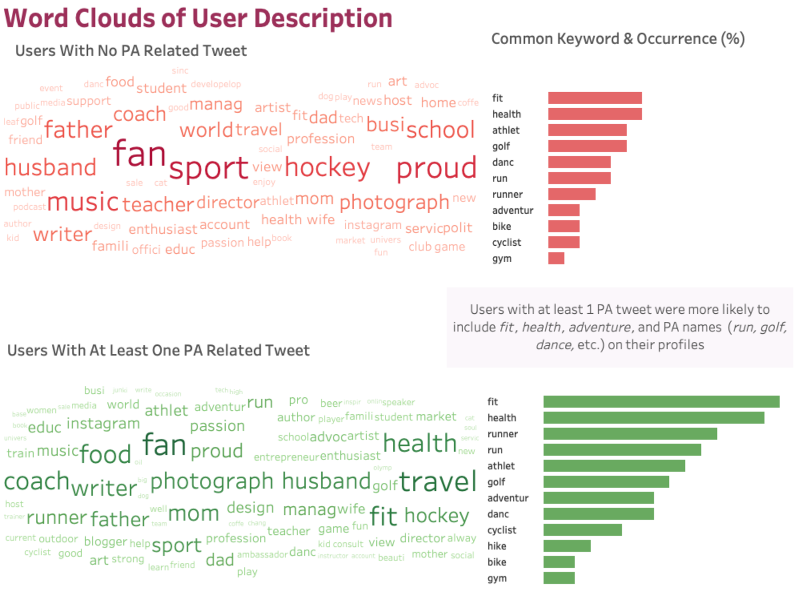


Figure 5: Modelling Prediction of Physical Activity Tweets by User Information

Word Clouds of User Description: Words associated with each user group's profile descriptions are visualized as two separate word clouds in Tableau. The frequency percentages of important physical activity related words found in both user groups were further illustrated and compared via bar charts.



Word Clouds of User Description

Users With No PA Related Tweet

Common Keyword & Occurrence (%)

Users with at least 1 PA tweet were more likely to include *fit*, *health*, *adventure*, and PA names (*run, golf, dance,* etc.) on their profiles

Users With At Least One PA Related Tweet

Figure 6: Words In Profile Descriptions of Users Have 0 vs. 1+ Physical Activity Tweets

5 Timeline

The analysis started on February 10 for some preliminary findings. The first results were finalized by February 24. A presentation that visualizes and summarizes our analysis findings was delivered on April 7. The final detailed report was submitted to the client on April 14.

6 Collaboration

We meet our clients weekly for feedback on the progress of our project and possibly for clarification for their expectation. We plan for all members from both teams to be involved so that everyone shares the same understanding about the project. Throughout the weekly meetings, both teams will gain communication skills, NLP, data processing, and how to create informative visualizations that can deliver the main messages effectively and aesthetically.

7 Conclusion

Based on data from Twitter and Google Trends, it was observed that physical activity levels and interest does vary from province to province. Twitter users actively post throughout the day about their physical activities, with peak Twitter activity from 5 pm to 8 pm.

There were some distinct trends in the user's profile description that could imply an association between how people introduced themselves and whether they would participate in physical activities. Additionally, we have also identified nine most commonly tweeted topics in the categories of fitness, sports, and fun/leisure activities, using an unsupervised LDA model, and tested the performances of several supervised classification algorithms in detecting self-reported physical activity related tweets. Our first findings about the topic associated words give us insights into how people often talk about or describe the activities in each category. These results will provide a great foundation upon which our further study on PASS surveillance will be based. Since the logistic regression and random forest models have shown a promising accuracy in classifying self-report physical activity tweet posts, if we can improve this model during the next stages of our analysis, we will be able to have more data for further unsupervised learning analysis of people's exercising habits. The model development methodology and techniques used in our project can also be leveraged to explore other health and social media topics, such as analyzing self-report concerns about sleep quality and sedentary life style.

There are limitations to the analysis. First, although performing relatively well in identifying topics from big data sets and taking the word contexts into account, LDA method is more successful in analyzing documents with clear structures and formal writing styles such as news articles, research journal, and government announcements. For a collection of personal tweets that can be short and arbitrary, topic modeling requires a considerable amount of manual tuning to navigate and thus relies on the analyst's assumptions and interpretations. Most of the data was retained for the topic modelling via LDA, however a future improvement to this work may be to exclude retweets, mentions, weblinks and/or hashtags to simplify the dataset. Second, the data analysis was based on a labelled dataset and it should be recognized that labelling is subjective and may have errors. By investigator triangulation and assessing coder agreement, future research can overcome this issue. Third, due to the time and resource limit, our project did not analyze non-English tweets, which might have resulted in, for example, the low rate of self-reported physical activity tweets in Quebec where many people speak French. Expanding the scopes to cover other popular languages is necessary in order to obtain a more representative of the Canadian population (i.e., to account for ethnic equity). Lastly, in the classification models to detect whether or not a user will have a physical activity tweet, determining feature importance (e.g., words that were most significant to improve the models) is complicated and computationally expensive, so we could not fully tackle this challenge by any means except manually inspecting the words and trying a couple selections of potential words. Future analysis should perform formal feature importance test as well as dimensional reduction to decide the stop-word and predicting-word lists, which can help to improve the model performance.

8 Suggestions

The breakdown of the project into multiple phases helped to guarantee a high quality deliverable. Starting the project early was appreciated as it gave us many opportunities to refine our objectives and improve our analysis/visualizations. Additionally, timely and detailed feedback/suggestions for improvement as well as clear instructions/requirements contributed significantly to our success.

While it is valuable to simulate a client interaction, it was not considered a priority to conduct on-going reviews with the client. Since the clients are other students in the class, it was not a realistic representation of client reviews. Our classmates tended to be easy-going and agreeable, whereas clients are much more invested in the scope and outcomes. A suggestion for improvement would be to either eliminate the requirement for on-going client reviews so that the focus remains on the project output or provide more expectations on client involvement.