Understanding Population Health
Analyzing Social Media Data

Siddhesh Pawar, Shilpa Gohil

Data Science 624-Winter 2020

May 1, 2020

Key Findings of the Project

Twitter data is a social media platform that affords us a rich source of real-time data that we can transform and harness into useful information for the welfare of the Canadian population. With the advent of machine learning algorithms that continue to improve and evolve, we can get labelled and classified tweets in a manner that aligns with our goals. The impetus of this project was to understand the role of tweets in the context of PASS indicators (physical activity, sedentary behaviour and sleep disorders). Using Twitter data for this project has not only facilitated answering our clients diverse questions, but it has also brought about meaningful insights thus exhibiting the versatility of Twitter as an information resource.

1. **Our project has led to the culmination of Canadian cities we need to focus on to increase physical activity and decrease sedentary behaviour from the standpoint of percentage of tweets per city as well as percentage of tweets per capita. Our project also showcases how the results vary between the two.**
2. **This project also facilitated insights such as smaller cities generally tweet less irrespective of physical, sedentary or sleep disorder tweets. We also see a general decreasing gradient in the prevalence of PASS indicators as you get further from a capital city.**
3. **We found the results from Twitter data to greatly overlap with overall results for CCHS (Canadian Social Health Survey) thus providing the impetus for these data sources to supplement each other. However, we found that triangulation from sources such as Google Trends and Reddit could lead to varying results due to tenuous and hard to match keyword searches and dissimilar scales.**
4. **PASS survey data from national repository used to relate with the Twitter data emphasizes that Canadian adults need to pay more attention to their health than the youth.**

1 Introduction

We live in a digital age where our increasing connection with the Internet has transformed research and information gathering into a process that is convenient and streamlined. Historically, information consumers provided about their behaviour of any regard was gathered through in-person surveys, educational institutions, external marketing, etc. Now, this same information can be found and collected instantly and anonymously in the privacy of one's personal space. Innovation and plethora of developments in the social media avenue and organized structures on mobile platforms, has brought the power to human fingertips. This can come from online form submissions, public forum discussions, social media channels such as Twitter, Facebook, Instagram, etc. This is particularly true when it comes to accessing health information online, talking about personal lifestyle concerns, seeking advise from non-medial experts (i.e.friends, social circle), or even sharing information about one's routine on different health subjects viz., diet, mental care, physical fitness, insomnia, etc. With the advent of new technologies, we find ourselves at the nexus of being able to acquire a dearth of data from these resources. Therefore, public and health institutions are using these new sources of data to conduct more comprehensive analysis and bring forward better health solutions to the markets.

In this report, we will explore and analyze a major Twitter dataset that was collected as a part of an online data scrapping project at the ***University of Calgary's Data Intelligence for Health (DIH) Lab under the supervision of Dr.Zahra Shakeri Hossein Abad and Dr.Joonwu Lee***. The Twitter data provided to us was collected roughly between October of 2019 to November of 2019, and additional data was supplied subsequently in order to carry out more detailed analysis to supplement our findings. Approximately ten-thousand tweets were used by us to synthesize all information we could gain on the three PASS (Physical Activity, Sedentary Behaviour, Sleep Disorder) indicators. The information that can be harnessed from these tweets, will give us some context on the Canadian population's perspective on their health or what generalized topics are discussed regarding health and lifestyle amongst the population in focus.

The data collected from social media channels allows private and public sector to bridge the gap in our Canadian healthcare and set out diverse projects and multi-level roadmap to ensure future generations in Canada to be healthy and strong[Eke11]. By analyzing the social media data and learning about Canadian sentiments, we can enhance health and tax policies to allocate subsidies or credits for encouraging Canadians to partake in more physical activities, access mental health care services with more ease and convenience, provide more resources in the context of stress reduction thus allowing the people suffering from insomnia or other related anxiety issues lead healthier lives.

The healthcare system is constantly challenged to improve the internal functioning, organizational aspects and offerings to the consumers as a whole. If proactive measures are taken to arm ourselves with insights granted from harnessing data from various sources, we could improve the lives of millions of Canadians, make the society healthier, and reduce the economic stress that falls on our healthcare system and all the workers working within this ecosystem. A good example from current affair would be the novel coronavirus *(COVID-19) pandemic*. These hard times, would allow us to utilize the monetary reserves and physical power of our health system to work rigorously towards combating the pandemic rather than common challenges faced by Canadians causing further complications to our medical staff to deal with.

As you follow through the report and visualizations developed, you will find our analyses and the resulting conclusions giving us all a general overview of how Canadian health currently is and how can we use the Twitter data.

1.1 Project Motivation

It is widely discussed, that of all the data mined and collected from the internet, only a tiny portion (approximately 10%) of the data might be of value to use and generate results [MBD+12] [1]. Using our Twitter data, our primary goal is to learn insights about PASS indicators in the Canadian population context. We would like to generate some high-level synopsis of the data and answer certain queries that were part of

[1] Harvard Business Review with former McKinsey & Company global managing partner, Dominic Barton

the project. A note to be taken here is, some queries were provided for our research by our classmates who acted as an external institution *(clients)* asking us to solve some realistic problems. The setting of this research was stemmed from a consulting strategy view where our professor *Dr.Abad* and teaching assistant *Steven Dykstra* were supervising our work which we were to deliver for our clients. In addition, we researched on some critical insights we could draw from, using applications and techniques we learned throughout the semester.

1.2 Problem Definition

Using the aforementioned PASS indicators provided to us, and our Twitter data being classified into, we have the following queries we would like to learn about and answer in our report. We also would perform additional analysis, using all the tool and resources available to us at our disposal.

- Which cities should we be focusing our efforts to increase physical activity and decrease sedentary behavior?
- Does proximity to a major or capital city increase the rate of PASS indicators detected through tweets? Is there a gradient in the prevalence of PASS indicators when you get further from a capital city?
- Do Twitter PASS surveillance results match up with the reported rates of obesity (and related health concerns) for those areas? Is there any pattern present?
- Does the Twitter PASS analysis reveal any association between health and the educational curriculum on health and lifestyle at the provincial level?
- Can we analyse an external data source such as Reddit and Google Trends (data triangulation) and tie back to the Twitter analysis?
- Using a government survey, could we develop or establish some basic relationship with the PASS indicators?

The aim of this work is to demonstrate whether this source of information is one hundred percent possible and feasible to use as a proxy to assess the health of the Canadian population. To that end, we would also like to study trends of how active is the population across different geographical boundaries in Canada. In our opinion, this data source can provide us some really valuable insights that we could verify with national data or survey and relate between each other. Ultimately, we would like to provide some recommendations and make conclusions based on our analyses, to answer each of our above questions, and give our clients some insights to consider, limitations they might have to work with and few recommendations to consider for future development within this scope of project.

---

2 Methodology

With the exponential growth in the field of computer science and data science, we have seen a burst in technology and computational methods that can be applied to work with and develop results from our given kind of dataset. With the knowledge of machine learning, data analysis and ability to comprehensively develop visualizations, our fundamental approach is to render support to decision making system for various stakeholders.

The core technique and methods used in cleaning, processing the Twitter data, tools and software used in performing the analyses, visual tools and methods applied to visually explain our findings, etc. are described as follows below.

2.1 Data Cleaning and Structuring

- Most of the core data cleaning with all the datasets used in the project were done in Python using libraries such as Pandas, Numpy, and SciPy.
- From the labelling that was completed manually [2] was amalgamated with the original datasets for each PASS indicators viz., physical activity, sleep disorder and sedentary behaviour. The five labels from which one was assigned to each tweet were, *"self-reported:yes;indicator:yes", "self-reported:no; indicator:yes", "self-reported:yes;indicator:no", "self-reported:no;indicator:no"*. Here the indicator refers to the PASS indicator on which dataset was based.
- We further cleaned each dataset by deleting the unnecessary and empty columns, arranging the format in Pandas dataframe that would make analyzing the data simple and convenient. The datasets with additional tweets were also used later in our analysis.

2.2 Data exploration

For the purpose of data exploration, we used Tableau to construct a map of Canada and check for the distribution of the tweets for each PASS indicator and the classification labels. This method provides us an overall picture on where are the most tweets concentrated within Canada. We also did a quick visual check for the distribution by proportion for each PASS indicator and its classification. These exploratory visuals and methods are generally good to make sense of the data and get a quick snapshot on what we would be dealing with in the process.

Lastly, once we calculated the proportion of physical activity tweets and sedentary behaviour tweets per city, we plotted the results on a scatter plot in order to check if there were any observable patterns in the data.

2.3 Data Analysis and Visualization

We applied a few different techniques through the process of our research for *Natural Language Processing* (NLP). This was done using different conceptual tools and mechanics that were taught throughout the course of the semester. We further break our analysis and methods in this section to explicitly understand how the analysis flow.

Here we will also describe the visualizations used to showcase our analyses and the reason why we chose them.

2.3.1 Supervised Learning

In this section of method and analysis, we employed two machine learning models which were Binary Naive Bayes and Random Forest. This learning was performed on the physical activity dataset as our interest was primarily to study the classification of tweets using the labels set for the tweets.

Binary Naive Bayes

Naive Bayes classifier typically works on a numerical format data. Since, our raw data was in textual format in form the tweets, we needed to convert the corpus of tweets of different lengths into numerical format. This was achieved by using the process of vectorization from the Term Frequency — Inverse Document Frequency (TF-IDF) technique. By performing this task, we can ensure that the classifier can now interpret the text in our corpus in numerical format.

[2] This was collectively done by the class of DATA 624 - Winter 2020 | University of Calgary.

Random Forest

Random decision forest is an ensemble learning method in machine learning which we used in this analysis for the purpose of classification. When working with the Binary Naive Bayes, we modelled the data on the label column of the data which was made binary for our purpose in the analysis. We classified the label as self-reported or not self-reported. This was to make the label dichotomous. On this data, we then performed random forest for analysis on the self-reporting aspect a given tweet.

A note here would be, even though we performed supervised learning on physical activity alone, we were generally interested in seeing if our model can predict efficiently whether the tweets are posted as self or some other variant in publishing. There many other methods that we could have used from the context of social media text processing[TTSR14] but we found random forest to be more apt for our dataset and purpose.

2.3.2 Unsupervised Learning

We also used unsupervised learning in order to carry out topic modeling, information extraction using sentiment analysis, and we subsequently built a word cloud to depict our findings from the sentiment analysis.

Topic Modeling

We used topic modeling for our physical activity twitter data, specifically on the text column that had all the tweets from different individuals and companies. This data analysis was done using R markdown. By using topic modelling on the physical activity tweets, we aimed to determine the three main topics of the tweets. We experimented with a different number of topics (2,4) as well as a different number of terms associated with those three topics. We ultimately picked 3 topics and 10 terms within that topic as we found that to be most informative in terms of non-repeated and meaningful words. We also added a few more stop words so as to have more meaningful topic analysis. An example of some of the stop words we added are: 'today', 'tonight', 'just'.

Topic modeling is a type of statistical model for discovering the abstract topics that occur in a collection of documents. Therefore, the ideal visualization for topic modelling is a bar graph for each topic depicting not only the topics found but also their quantitative frequency in terms of occurrence. This enables us to get a sense of the topics and at the same time allow for a comparison in how frequently these terms appear in our data.

Sentiment Analysis and Wordcloud

Sentiment analysis was done using the sedentary twitter data set in R markdown. We used sentiment analysis specifically on the 'text' column of our data set that contained the body of all the tweets in that data set. For this, we used the 'tidytext' package that contains sentiment lexicons. As a stepwise process, we generated a count of the most frequently occurring negative sentiments as well as the most frequently occurring positive sentiments from the tweets held in the data set.

We subsequently created a word cloud from the results of our sentiment analysis using the 'wordcloud' package available in R.

Sentiment analysis is the interpretation and classification of emotions (positive and negative) within text data using text analysis techniques. Like topic modelling, sentiment analysis can be depicted ideally using a bar graph showing the frequency of positive words versus negative words in the dataset. The bar graph allows easily interpretable results in the context of the highest occurring positive and negative words (color coded red for negative sentiments and blue for words associated with positive sentiments) and the frequency of each, and a comparison of the frequency of positive in relation to negative words.

Using a wordcloud enables us to visualize the results of the sentiment analysis conveniently and at a glance. It highlights important textual information in terms of words and the size of each word appearing in the wordcloud enables us to determine the relative frequency. Therefore, we used a wordcloud as our medium for visualization to depict the sentiment analysis in a more visually appeasing, meaningful and creative way.

2.3.3 Packed Bubble Charts

All data analysis was done using Jupyter notebook, Anaconda environment. We filtered all the tweets pertaining to physical activity and sedentary behaviour. For all physical tweets, we only used tweets labelled, "Physical Activity: Yes, Self-report: Yes". However, for sedentary behaviour tweets, we took not only tweets labelled, "Sedentary Behaviour: Yes, Self-report: Yes" but we also took tweets labelled, "Sedentary Behaviour: Yes, Self-report: No". The reasoning behind this was that whereas we found most non-self reported physical activity tweets to be advertisements, we found most non-self reported sedentary tweets to still be about sedentary behaviour however, referring to other people's sedentary behaviour (such as son, friend). Therefore, we decided to include this in our analysis as well. We subsequently calculated the proportion of tweets per Canadian city for each category (physical activity tweets and sedentary tweets).

We used a bubble chart to depict the Canadian cities that had the lowest number of physical activity tweets and another bubble chart to depict the Canadian cities that had the highest number of sedentary tweets. Bubble charts are a fun and visually appealing but at the same time meaningful method to depict these results as the colors represent the different cities, and the size of the bubbles allow a quick visual quantification and comparison between cities. It is an attention grabbing visual that can convey important information.

2.3.4 Scatter Plot

In order to explore the results of the percentage of sedentary tweets per city and percentage of physical activity tweets per city, we decided to plot the results on a scatter plot on Tableau in order to establish any observable and knowledgeable insights from these measures.

A scatter plot is a useful visualization when trying to unearth patterns in the data and check for the occurence of a relationship between two measures (physical activity tweets and sedentary tweets). It allows us to quickly visualize the results of each city from the standpoint of both measures as opposed to looking at only one measure at a time and allows us to see trends in data that we would not otherwise notice.

2.3.5 Dendrograms

All data analysis was done in a Jupyter notebook, Anaconda Environment. We calculated the total counts of physical activity and sedentary tweets per city and instead of calculating the percentage as a function of the total number of tweets in the dataset, we actually calculated using each city's respective population in order to see whether population size played a factor in the number of physical activity and sedentary tweets per city. The population data per city was obtained by web-scraping the Wikipedia page for Canadian city population.

Dendrograms allow for a beautiful and artistic way to depict information. The order of the levels (cities) allows us to infer a quantitative relationship (from highest to lowest or lowest to highest). The color of each bar allows us to easily distinguish each city and the size of each bar easily enables us to quantify the proportion of tweets per city.

2.3.6 Funnel Charts

For this question we used both sets of Twitter data (one we labelled ourselves as well as the one that was labeled for us). This was done as we did not have enough data from just one Twitter dataset. Therefore, we merged the two sets of data and grouped by province and calculated the percentage PASS indicators for all capital cities in each province along with other smaller cities in that province that were at different distances away from the capital city. Smaller cities from each province were chosen at random and at varying distances from the main city. We subsequently plotted the percentage of PASS indicators from each city in a province in one funnel chart and hence a funnel chart for each province.

Funnel charts were ideal as they allowed us to check whether there was a gradient in the PASS indicators as one went further away from the capital city. The percentage of PASS indicators from each city in a province was plotted on a singular funnel chart and hence a funnel chart for each province with its respective

capital city strategically placed at the top of the funnel chart and associated cities formed the base of our funnel chart. This visualization was ideal for answering our question as it enabled us to quantify the PASS indicators gradient from the capital city to other cities at varying distances in the different provinces. Adopting this methodology of having the capital city occupy the very top of the funnel chart allowed us to see not only when cities maintained the decreasing gradient but also when cities did not maintain this gradient in PASS indicators as one went further from the capital city.

2.3.7 Correlation Plots

In order to carry out triangulation to compare our results from the Twitter dataset and CCHS data (Canadian Community Health Survey) in order to see if reported rates of obesity matched up with information from the Twitter data, we first calculated the obesity rates per province (total number of people that reported being obese in the survey divided by total number of people that responded to the survey). We subsequently calculated the rate of sedentary tweets (total count of sedentary tweets per province divided by total number of tweets from that province) and the rate of physical activity tweets (total count of physical activity tweets per province divided by total number of tweets from that province) from the Twitter dataset. Rate calculations were done for both datasets in order to make comparisons possible. Pearson's correlation was also calculated to compare the obesity rate from CCHS data and rate of physical activity tweets as well as Pearson's correlation between obesity rate from CCHS data and rate of sedentary tweets as a means to quantify the relationship between the two data.

Scatter plots were ideal for this question in order to see trends or patterns in the data between the two sources. The scatter plots along with the Pearson's correlation value for each plot enabled a quick communication of the quantitative relationship between these sources of data and a visual inspection of this correlation can be directly seen on the respective scatter plots thus making it easy to understand.

2.3.8 Web Scraping

Web scraping for more social media data from other platforms was performed to perform some analysis and develop some understanding on how do the data on different social platforms blend with each other. We scrapped more data from Reddit and Google Trends with the sophistication of knowledge provided to us by Dr.Abad. We sourced the data to closely generate relevant information for our PASS indicators. Following describes our methods.

Reddit

For the Reddit data, we scrapped it using the Python Reddit API Wrapper(PRAW). The authentication needed to access the Reddit platform for scrapping the data were obtained by us while being compliant with all their policies. Further more, using the *praw* library in Python, we scrapped the top one-thousand discussion titles for three different subreddits viz., r/Exercise (for physical activity), r/insomnia, and r/sleep (for sleep disorder).

The reason we chose to use the discussion titles against the comments themselves is because, we wanted to scrape the actual subject of the entire discussion body rather than comments in discussions which might be pulled from different context. This would greatly cause a mix of topics which might not be discussed by people for health and fitness, but rather for promotions, nutrition supplements, sports conversations, etc. Similar caution was exercised with the insomnia conversations. We wanted to ensure we were able to get apt discussions and texts from them rather than irrelevant chatter about nightmares, and overnight work.

Since, the availability for sedentary behaviour data on Reddit is very dilute and sparse, and doesn't have a dedicated subreddit from where we could reliably get good information, we omitted this PASS indicator and rather got some data from Google Trends.

Google Trends

Pie-charts

To compare PASS indicators from different provinces, we made use of Google Trends compare functionality. This enabled us to compare three keywords at once. The particular key words chosen were: “gym” (corresponding to physical activity in the Twitter data), “insomnia” (corresponding to sleep disorders in the Twitter Data) and lastly “lazy” (corresponding to sedentary behaviour in the Twitter Data). The key words were chosen with careful experimentation and were picked because together they yielded interesting trends. These keywords were also chosen keeping in mind certain information such as that the time period during which data was collected (12th October to 1st December) is typically winter season in North America and hence most people’s searches would pertain to indoor activity such as going to the gym. Also, the keyword pertaining to sleep disorders that generated a dearth of information was “insomnia”. Lastly, for sedentary behaviour “lazy” yielded the most quantitative information. Others words we experimented with were “free-time”, “idle”, “sports”, “sleep disorder”, “physical activity’. However the best results were from the chosen keywords. *Refer figure 10*

A Canadian map with pie-charts representing all three PASS indicators was used for this as it clearly depicted all three PASS indicators for each province (both a comparison between and within provinces could be made in regards to the PASS indicators).

Line graphs

In order to show a quantitative trend of the most current PASS indicators from Google Trends, we again extracted a compare functionality (in Google Trends) of the same three keywords used for the location pie chart visualization (insomnia, lazy and gym). We subsequently plotted a line graph of all three indicators in one graph to see the general trend of these PASS indicators over the last year.

A line chart was used for this visual as it enabled a general trend of PASS indicators to be observed during the past year.

Radar charts

In order to check the correlation between our two sources of data (Google Trend and Twitter), we used the Pearson correlation. Pearson’s correlation is a measure of the linear correlation between two variables X and Y. It has a value between +1 and -1, where 1 is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation. In order to do an appropriate correlation, the data from the two sources needed to be normalized first as there was a disparity in their scales. For this we used Sci-kit learn’s Data normalization. This allowed us to rescale real valued numeric attributes into the range 0 and 1. In order to show a comparison of PASS indicators between provinces and between the two data sources (Twitter and Google Trends), three separate radar charts were used to depict these comparisons. In order to calculate the PASS indicators per province, we calculated the total percentage of “Yes physical activity, yes self report”, “Yes sleep disorder, yes self report’ , as indicators of physical activity and sleep disorders. For sedentary behaviour however, we included both “yes sedentary behaviour, yes self report” as well as “yes sedentary behaviour, no self-report” as indicators of sedentary behaviour. This is because most non-self reported sedentary behaviour was also pertaining to sedentary behaviour but mentioned relatives or other people being sedentary as opposed to these entries being advertisements.

Radar charts were used for this question as they enabled us to make quick comparisons simultaneously, not only between provinces but also between the two data sets (Google Trends and Twitter data) as well.

2.3.9 Daisy Chart and Bar Plots for Data Comparison

For us to compare the Twitter data PASS indicator performance across a national database, we employed a dataset from open Canada on PASS indicator dataset which looked at aggregate results from a survey conducted by Public Health Agency of Canada in 2017 and released for access in 2018. We input this data directly in Tableau and used the aggregate results to develop daisy charts that were binned in a format to capture a twenty-four hour clock range.

Daisy charts are great for visual mediums to compare and analyse any type of survey data, hence we chose to use it for our comparison of a national survey with Twitter data.

We generated some visuals in R and Python, using libraries such as Seaborn, Plotly and Matplotlib. We developed additional visuals for more meaningful insights in Tableau[3].

3 Performance Measurements

One check we wanted to perform on our Twitter data was to see its accuracy to predict if a tweet or corpus was self-reported or not. This can be altered to check any other classification we had performed. We wanted to test the accuracy of the data on the model we created in the supervised learning method. The data being manually labelled, after splitting it for training and testing, predicted with almost 95%+ accuracy that the tweets were self-reported. This is relatively high accuracy and we did a little deeper dive into this. The classification was quite imbalanced and had more proportion of the tweets being labelled as self-reported even if the PASS indicator was a negative value, which wasn't of our interest in this case.

We tried fitting a model by oversampling the minority class but this did not yield us any great results and decided to stay put with our analysis performed above. Upon applying final model with random forest, we got an accuracy of almost 99% (refer to the Figure 1., for the ROC curves. This just tells us if a social media data, has a classifier variable in binary format and should that variable have higher proportion of a certain indicator, when we split the data for applying supervised learning on train set of the data, we would see a higher accuracy on predicting that result on the test split of the data. A cautionary exercise would be to perform more rigorous text processing to be able to get a good model to fit for prediction.

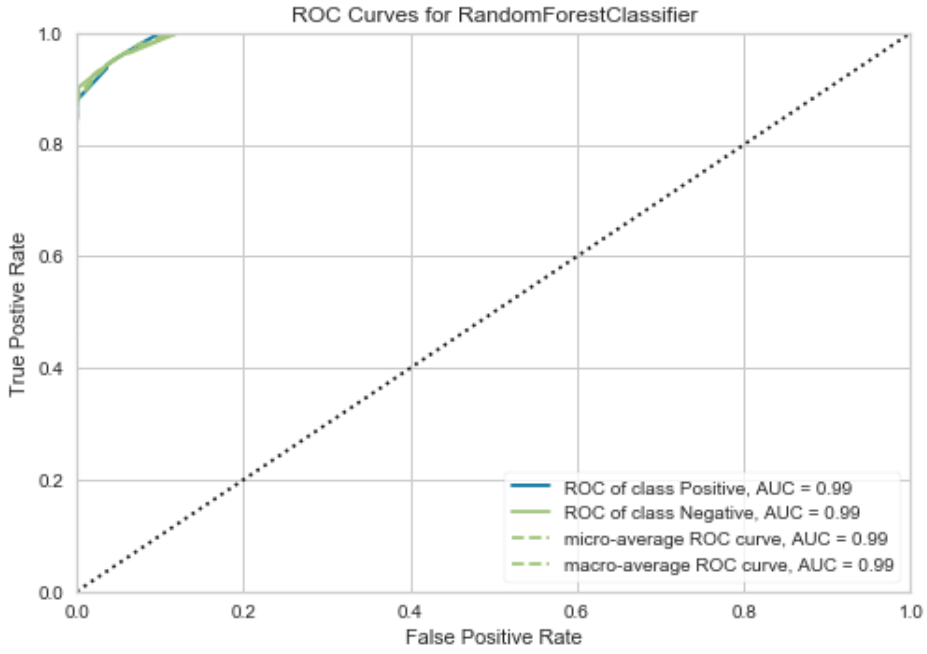

Figure 1: *AUC graph*.

[3]Licensed copy acquired from UCalgary account.

4 Results

In this section, we will now explore the actual visuals we were able to develop for our analysis and bringing back the questions we wanted to answer on grand scale of things, we will describe our results.

4.1 Exploratory Check

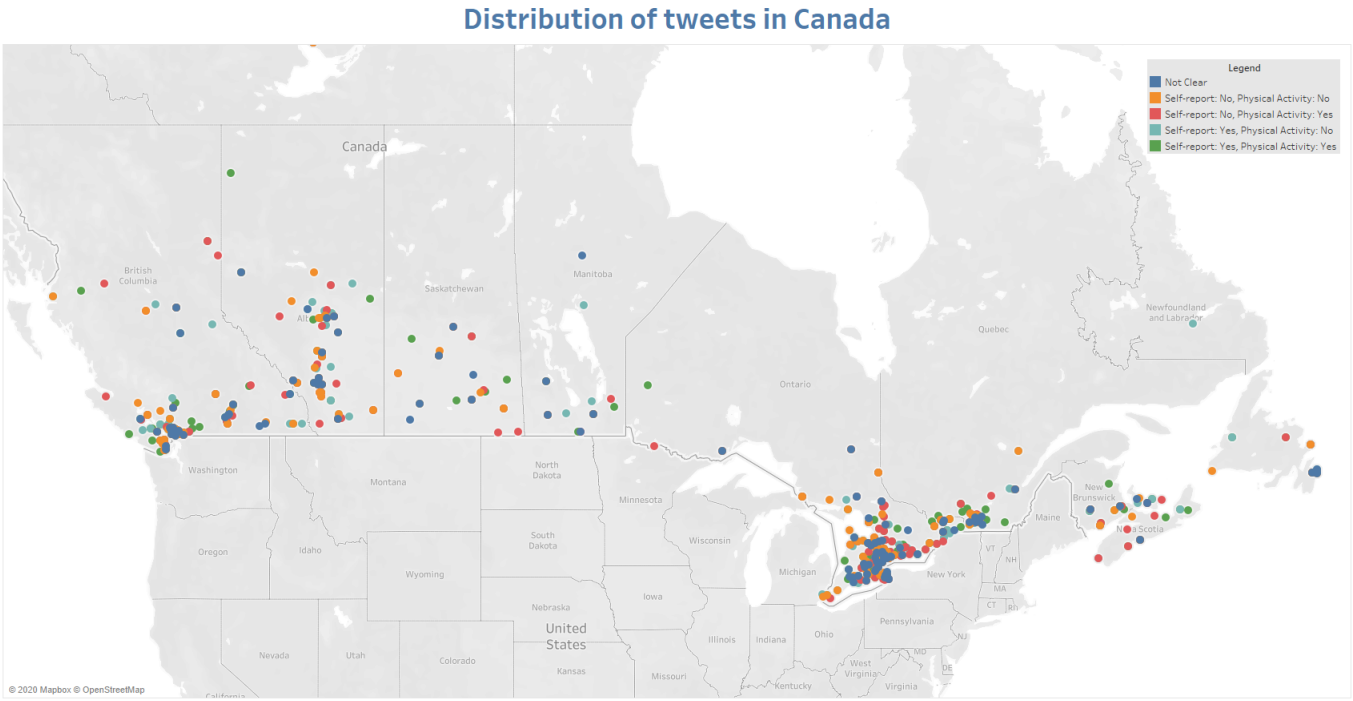

Figure 2: *Physical Activity Tweets Distribution in Canada*.

Using the map of Canada, we can clearly see the distribution of tweets in regards to the physical activity of the PASS indicators. Most tweets come from the densely populated regions in Canada[4]. Similar results were seen for the other two PASS indicators. Since, most of Canada's population is concentrated in the southern Ontario region, Greater Vancouver area in British Columbia and the Greater Montreal region in Quebec, we see more activity on Twitter in these regions. For the Canadian Prairies, we see the most significant area to be in Calgary that is sprawling with Twitter activity.

Capturing such tweets and analyzing population distribution such as sex, age ranges, etc., that is social media savvy and active on these platforms would need more data from personal user end which is generally restricted and private because of *The Privacy Act* in Canada. Nonetheless, if someone is on social media platform, it would seem they are going to have some digital footprint on the internet and have a reason for being there, such as seeking help and advise, participating in live chats, engaging on discussion forums, etc. This data would become extremely valuable, if used correctly, to make an impact in health surveillance

[4]Some towns and municipalities are not reflected since their coordinates do not have values in Tableau's Mapbox

research.

4.2 Cities We Must Focus On for Healthier Living

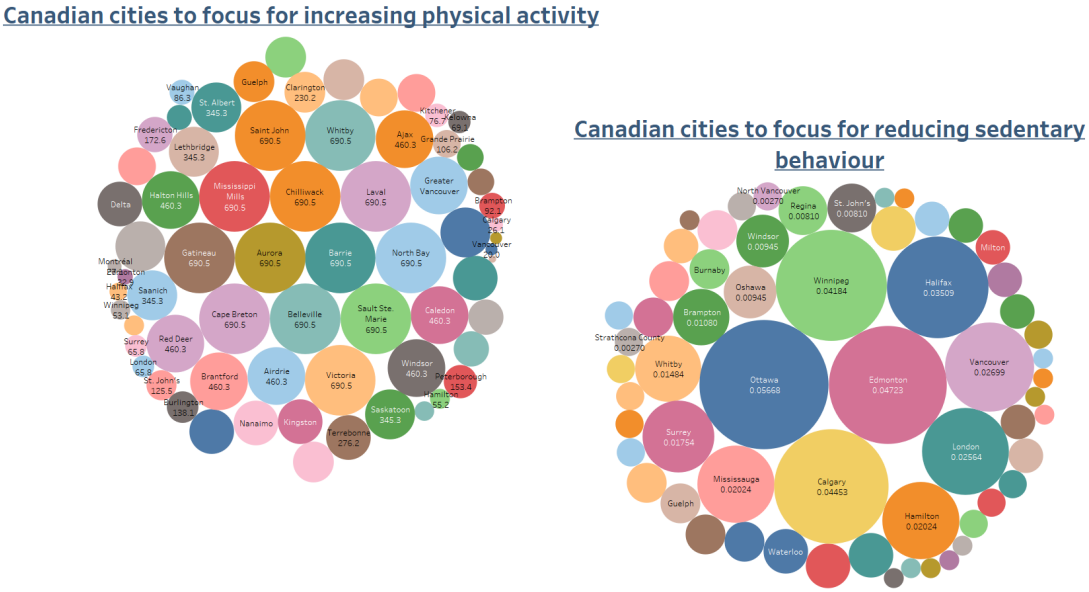

Figure 3: *Bubble Chart for cities to focus on*.

Using the quantity of physical activity tweets and sedentary behaviour tweets in our dataset, we were able to identify cities that had amongst the lowest number of physical activity tweets as well as the highest number of sedentary tweets. Armed with this knowledge, we can assess which cities require focusing on in the context of improving physical activity and decreasing sedentary behaviour.

From the visualizations, we can see that some of the cities that have the lowest proportion of physical activity tweets are Aurora, Barrie, Chilliwack, Laval and Cape Breton while some of the cities with the highest proportion of sedentary behaviour tweets are Ottawa, Edmonton, Winnipeg, Vancouver, Halifax (to mention a few). We also see that a lot of cities with a low proportion of physical activity tweets have the same proportion or similar proportion of physical activity tweets whereas there is a greater disparity and variation in Canadian cities with a high proportion of sedentary behaviour tweets (a lot of the cities show up as the clear winners, having a far greater number of sedentary tweets than others). This shows that there are some clear indications within the context of twitter data as to which cities have the highest proportion of sedentary tweets. From the visuals, it is also interesting to note that there is not a lot of overlap between cities with lowest proportion of physical activity tweets and cities with highest proportion of sedentary tweets.

This visualization therefore gives us an overall summary of physical activity tweets and sedentary tweets per Canadian city and ties in with the question of which Canadian cities to focus on to increase physical activity and decrease sedentary behaviour in the context of Twitter data.

4.3 Relationship between Physical and Sedentary Tweets

From looking at the packed bubble charts, we were interested in doing some exploratory data analysis in order to see if there was any emerging trend between the percentage of sedentary tweets per city and percentage of physical activity per city. From the scatterplot, we noticed smaller cities (with smaller populations) such as Regina, Lethbridge, Kingston had lower percentages of sedentary and physical activity tweets whereas larger cities (with higher populations) such as Edmonton, Calgary, Ottawa had a higher

number of sedentary and physical activity tweets. This led us to question whether having a smaller population amounted to fewer tweets in general from those cities irrespective of topic. From the scatter plot, we also noticed that a couple of cities such as North Vancouver do not follow in the same trend and had a low percentage of sedentary tweets and a high percentage of physical activity tweets despite being a bigger city with a higher population (an anomaly compared to other cities).

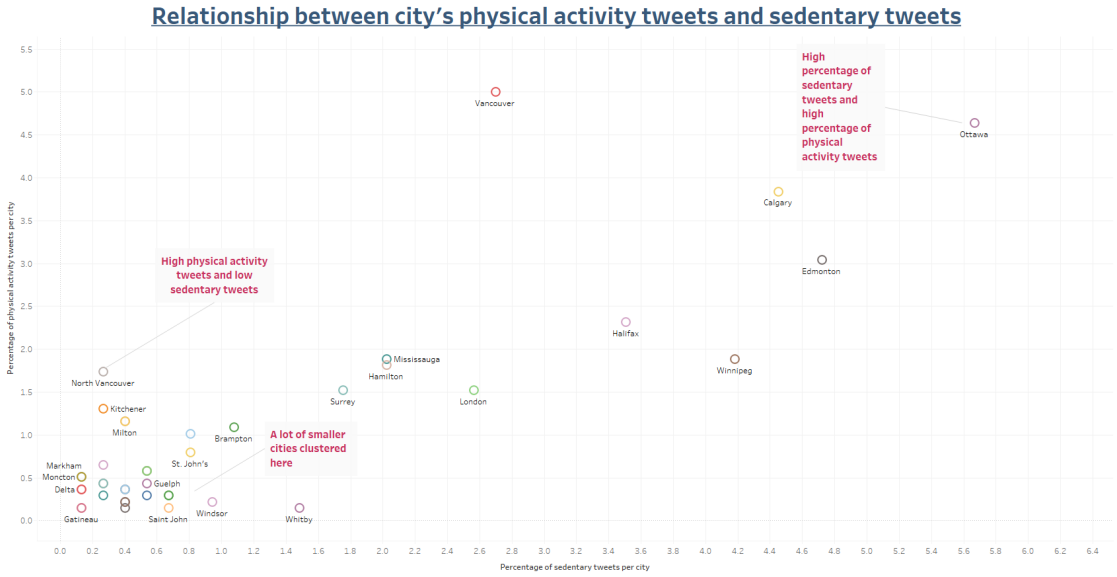

Figure 4: *Relationship scatter plot*.

These results indicated to us that population size may play a role in the percentage of physical activity and sedentary behaviour tweets from cities and this may be blurring and having an effect on determining which cities to focus on to promote physical activity and decrease sedentary behaviour. This also made us question whether individuals from smaller cities are less interested in tweeting in general as opposed to individuals from larger and more urban cities.

Our clients requested us to explore this in more detail and so we proceeded to calculate the percentage of physical activity and sedentary tweets as a function of the population for that particular city (as shown in the next figure).

4.4 Distribution of Tweets Per Capita

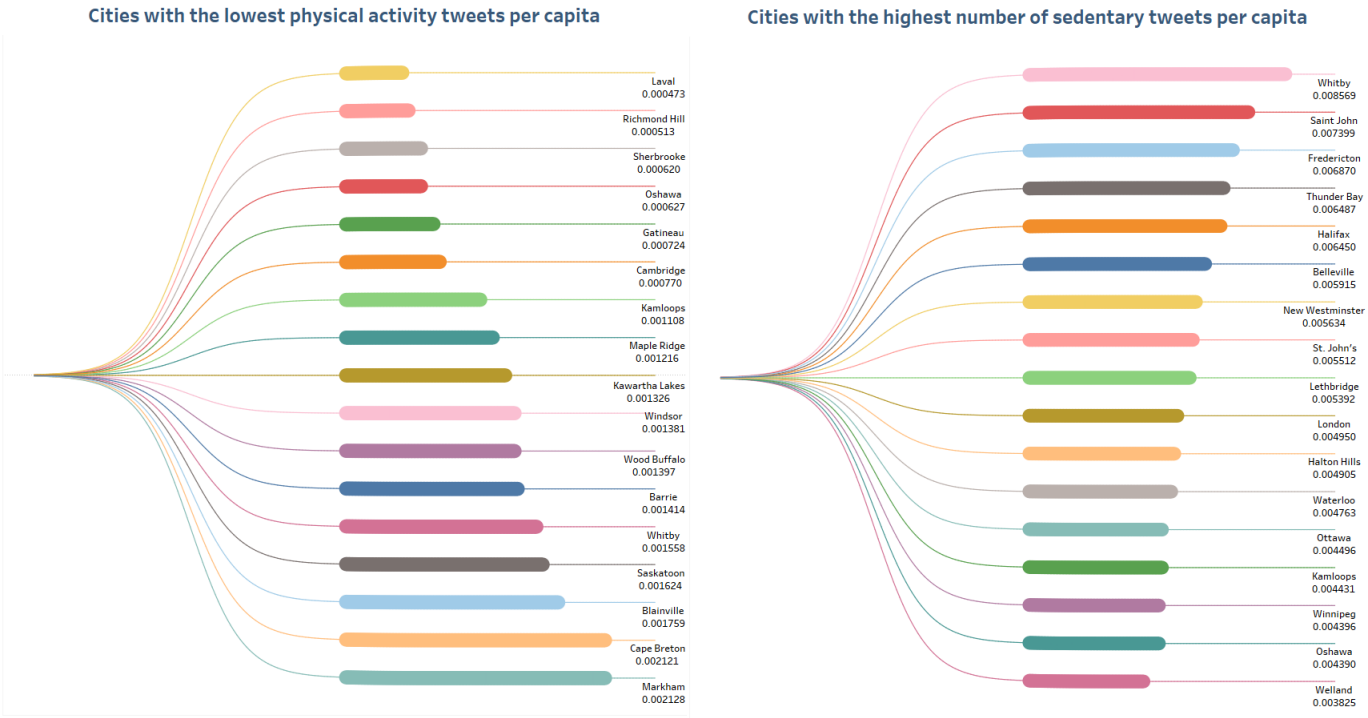

Figure 5: *Distribution between tweets per capita*.

The dendrogram shows us the cities to focus on to increase physical activity and decrease sedentary behaviour using tweets per capita. As we can see, the cities that show up in the dendrogram are slightly different from the cities in the bubble chart (that do not take into account each city's population). Cities such as Laval, Barrie and Windsor show up in both visuals (bubble chart and dendrogram from cities with the lowest physical activity tweets) whereas from the cities with the highest sedentary tweets Ottawa, St. John's, Halifax and Oshawa end up in both visuals too (dendrogram as well as bubble chart). From the few cities common in both per capita and per total number of tweets, the rest of the cities are quite different and we have the appearance of new cities such as Thunder Bay, Cambridge, Kawartha Lakes when per capita tweets are taken into consideration. We also see that cities such as Ottawa and Winnipeg no longer have the highest number of sedentary tweets whereas cities such as Whitby and Saint John become the cities with the highest percentage of sedentary tweets per capita. Interestingly, we see Whitby appear in both the cities with the lowest physical activity tweets per capita as well as the city with the highest sedentary tweets per capita thus indicating that it should definitely be one of the Canadian cities we should be focusing on to improve healthier living. Therefore, these visuals show that we get slightly different results when population size is taken into account but we also see that there is quite some overlap too between the results when population size is not taken into account.

4.5 Provincial Gradient of PASS Indicators

From our results, for the majority of provinces (British Columbia, Alberta, Ontario, Manitoba and Quebec) we see a clear gradient between the capital cities in each province and corresponding smaller cities in the respective province. From the shape of the funnel chart we observe that the capital cities of the aforementioned provinces have the highest PASS indicators and as we move away from the city, we develop a gradient in the prevalence of PASS indicators. We notice that British Columbia, Manitoba and Alberta have quite a sharp decrease in prevalence of PASS indicators further from the capital city and the gradient

decrease is gentler for Ontario and Quebec. From the shape and color of the funnel chart we also see that the smaller cities have varying percentages of PASS indicators, so for example, while Gatineau in Quebec has a high proportion of physical activity indicators, it also has among the lowest sedentary behaviour. We also see that cities such as Sherbrooke (Quebec), have a very small percentage of sedentary behaviour tweets but there is no physical activity or sleep disorder tweets from that city.

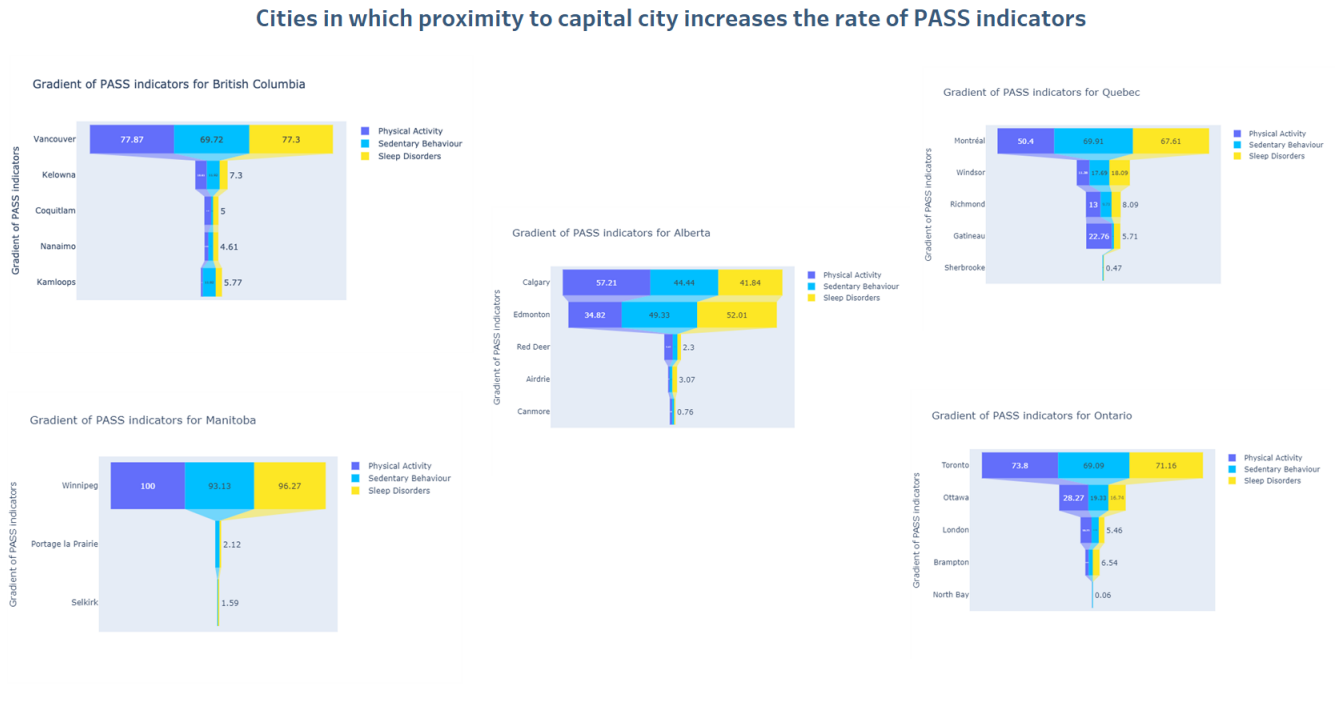

Figure 6: *Increase in rate for cities closer to provincial capital*.

Lastly we see that the provinces that do not follow in this same gradient of PASS indicators is New Brunswick and Saskatchewan. For example, we see that Saint John in New Brunswick has a higher percentage of sedentary and sleep disorder tweets in comparison to Fredericton (the capital). We can quickly distinguish these gradients by the shape of the funnel chart. Sleep disorders tweets in Saskatoon are also higher than in Regina thus showing that the province of Saskatchewan also does not follow the same gradient.

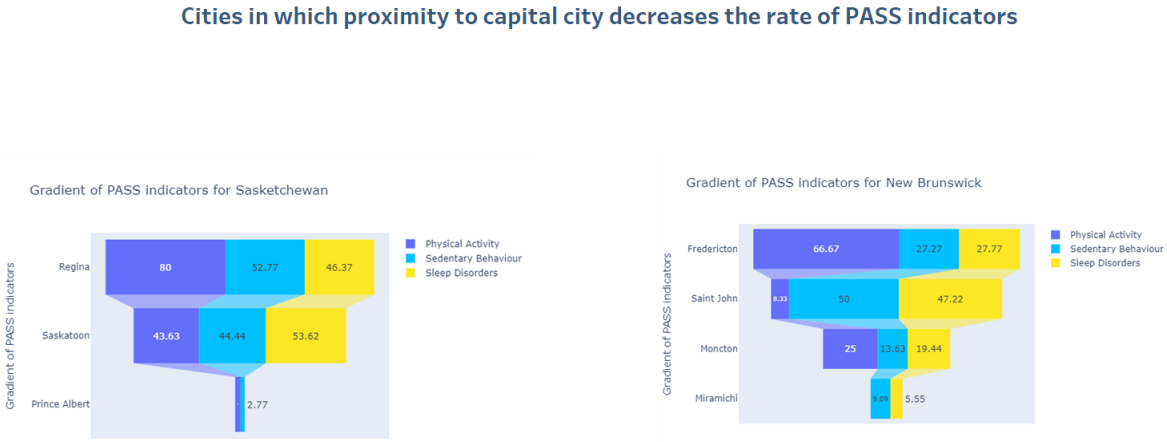

Figure 7: *Decrease in rate for cities closer to provincial capital*.

4.6 Relationship Between PASS Indicators and Obesity Rates

It was very important for our clients to know whether or not PASS indicators specifically sedentary behaviour tweets and physical activity tweets obtained from Twitter data could be comparable to data obtained from government sources such as the Canadian Community Health Survey (CCHS). This is important to know as similarities in the data between the two sources may allow us to draw inferences about the Canadian population directly from Twitter data.

Interestingly, we see some meaningful trends in our data. We see that as the obesity rate (calculated from CCHS data) for a province increases, the rate of physical activity tweets decreases and there is a weak negative correlation between the two measures (Pearson's coefficient of -0.21). Provinces such as Saskatchewan, Manitoba, Alberta, Ontario, Québec and British Columbia all exhibit and decrease in the rate of physical activity tweets as obesity rates increase and vice versa. Provinces that do not exactly follow in this trend are New Brunswick, Newfoundland and Labrador and Nova Scotia.

We also see a strong positive correlation (Pearson's correlation of 0.77) between obesity rate (calculated from CCHS data) and sedentary tweet rate. Thus as the sedentary tweet rate increases, the obesity rate also increases and vice versa. This is especially true for provinces such as Alberta, Nova Scotia, New Brunswick, Saskatchewan and Newfoundland and Labrador.

This is an important finding as we can see that there is quite a bit of overlap between PASS indicators from Twitter data and CCHS. Despite the two sources of data not being exactly identical, there is indication that Twitter data may potentially be a rich source of real-time, PASS indicator information for Canadian cities. It is important to note that the latest government data was only available for the year 2016 and this may result in some discrepancies observed between the two sources of data. This results also shows that whereas it may be too early to say that Twitter data may one day supplant government survey data, we can say that Twitter data can be used in conjunction with government data such as CCHS.

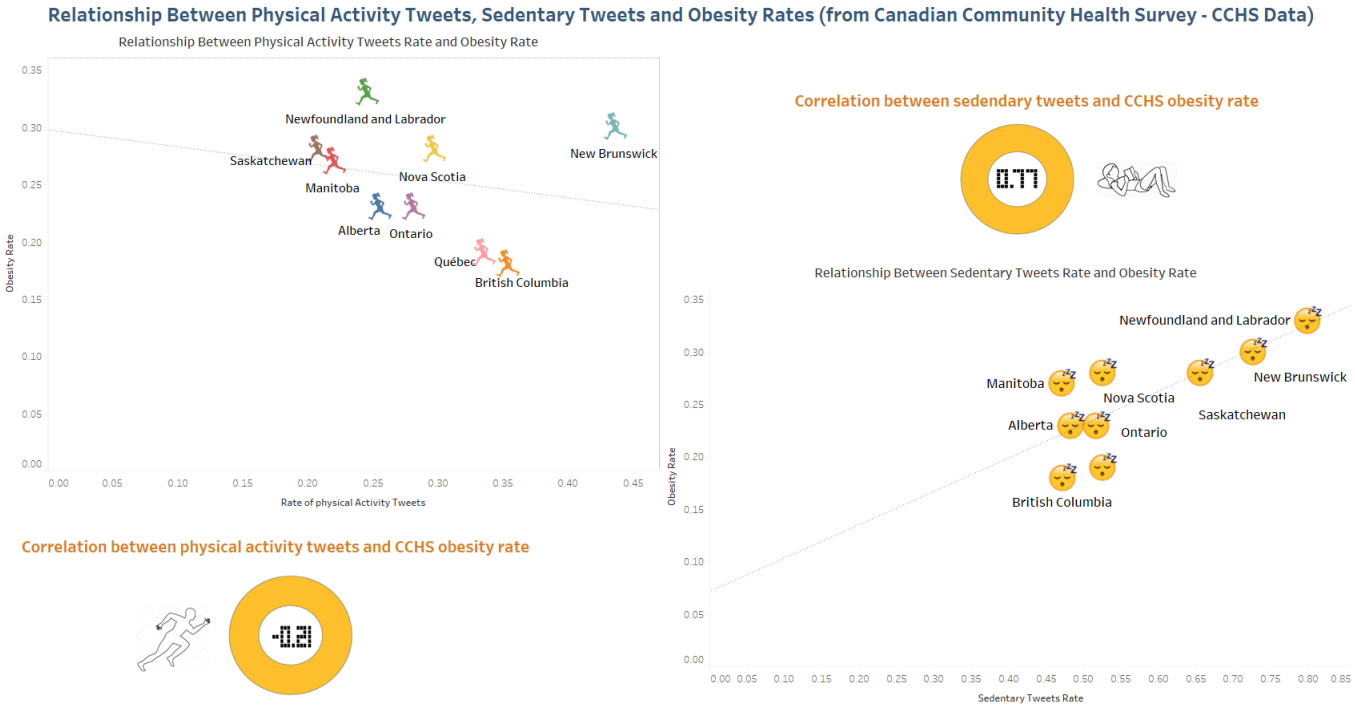

Figure 8: *Correlation of PASS with obesity rates in CCHS.*

4.7 Other Web Data Comparison

Other data sources we tried collecting data from were Google Trends and Reddit and compared it back to our Twitter data.

4.7.1 Twitter

As shown in our visualizations, using topic modeling, we found that some of the most common terms were: 'workout', 'night', 'dance', 'great', 'last'. The 5 most common terms with the other topic were found to be: 'play', 'great', 'first', 'workout', 'hockey'. Some of the interesting findings were that as we know Canada is a country where people love to play hockey and this is depicted from our topic modeling as the physical tweets data set has hockey mentioned a great number of times. This could be because of people playing hockey as well as enjoying watching it as sports entertainment. Workout was another term that commonly appeared in the topic and this shows that the tweets have a lot to do with working out. With the topic modeling we found the topics to be in sync with what we expect from the tweets in the physical data set.

From our sentiment analysis and wordcloud, we found both negative sentiment associated words such as 'hate', 'complain' as well as positive sentiment associated words such as "thank" and 'love". We also found some slang and informal words such as 'suck' as well as some swear words. This shows us that tweets from people are generally tweets from people are generally geared toward an unrestricted and informal audience a lot of the time and the nature of Twitter as a platform may facilitate people to be more open and exhibit their emotions through swearing. This is especially true for people tweeting about sedentary behaviour and we see this from the wordcloud.

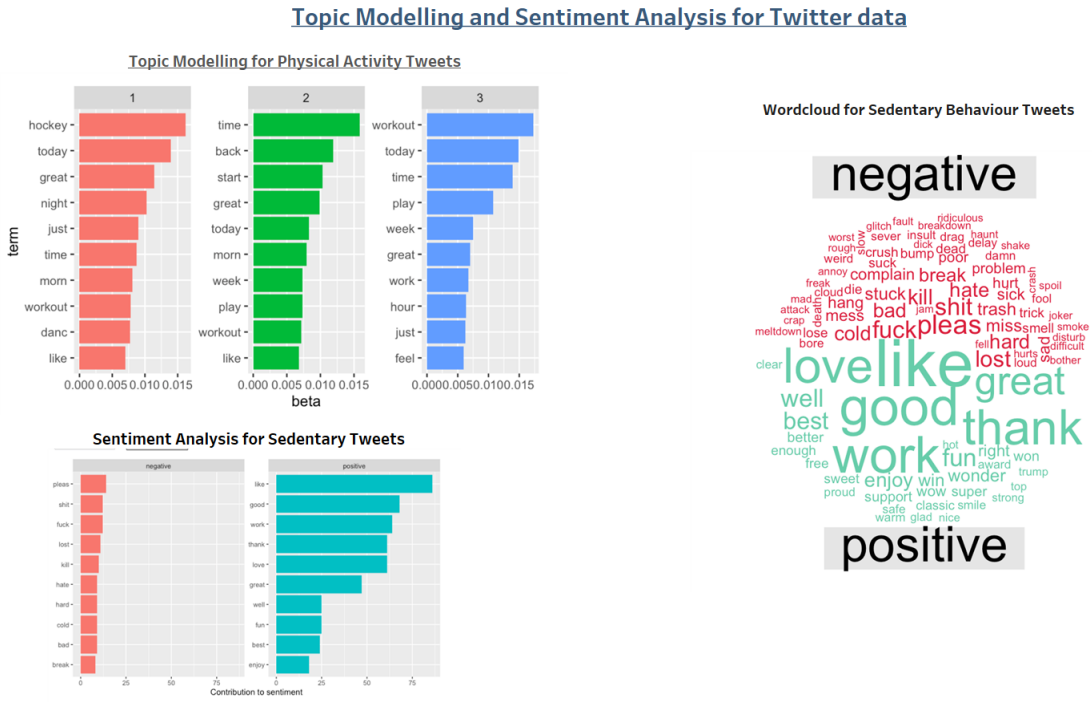

Figure 9: *Twitter Data*.

4.7.2 Google Trends

From the geographical map with pie charts as PASS indicators (from Google Trends) we see that the majority of Canadian provinces have a higher proportion of physical activity as compared to other PASS indicators (sedentary behaviour and sleep disorders). We also see that Ontario has the least amount of physical activity (approximately 50%) while sedentary behaviour and sleep disorders make up the rest 50%. From a visual inspection, we see that Saskatchewan ranks highest in sedentary behaviour. From Google Trends data we also see that sleep disorder takes up only a sliver of each pie chart thus suggestive of a lower proportion of sleep disorders such as insomnia plaguing Canadians.

From the general trend of PASS indicators from the Google Trends data for the past year, we see a sharp rise in physical activity since the second quarter of 2019, however, unfortunately we also see an increase in sedentary and sleep disorders during that time period. On a positive note, although sleep disorder and sedentary behaviour increased during that same period, the rise was not as drastic as physical activity. We also see a tapering effect for sleep disorder during the 3rd quarter of 2019 and continuing while sedentary behaviour declined beginning in the 4th quarter of 2019.

From the results of the correlations, interestingly we see that the correlations for all the PASS indicators have a negative correlation very close to negative 1. All very similar in value -0.962, -0.957, -0.940). This is indicative of a negative correlation between the Twitter data and the Google Trends data - an interesting pattern. This means that as a value increases in one data set, the value decreases in the corresponding data set. Since this correlation is close to -1, this is indicative of a strong negative correlation. The correlation helps us make meaningful comparisons between the two sets of data as they are based on different scales. Calculating the correlation also allows us to infer the relationship between the two data sets and to determine if there is a pattern in the trends of PASS indicators detected by each (negatively correlated).

The corresponding radar charts for each correlation, shows a visual of how this data from two different sources differs between provinces and how PASS indicators differ between provinces. We see a very similar trend for both sleep disorders and physical activity however sedentary behaviour follows a different trend. The measurement scale for google trends is generally higher than calculated Twitter percentages

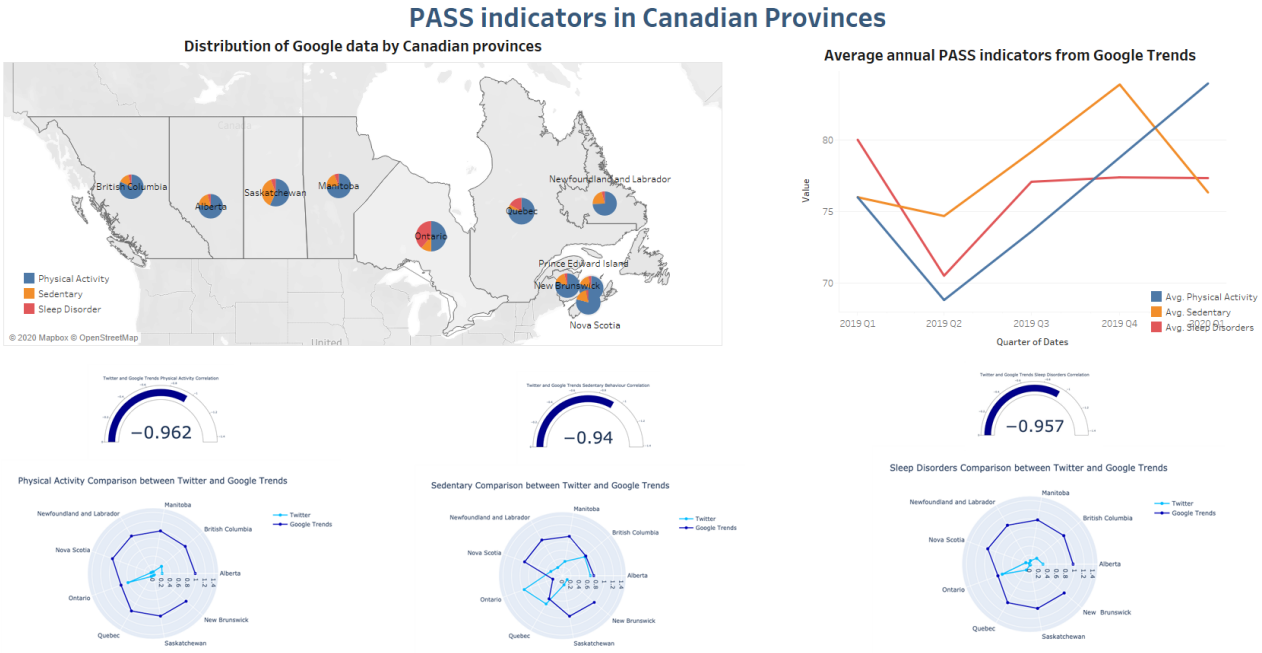

Figure 10: *Reddit*.

and we clearly see this in the radar graphs as Twitter data trend line is closer to the nucleus whereas the Google Trends data is further out. We also see very different patterns between the provinces when we compare the two data - they do not show the same trend for PASS indicators as the trend lines differ quite a bit. This begs the question of how accurate crowd sourced data is from both these sources and further investigation would be to gauge whether one is more accurate than the other. It is important to keep in mind however that both types of data are measuring slightly different things (depending on keywords chosen while scraping data from both sources).

Therefore, from this visual we see that the trends are quite different between the data sources. We also see some contradictions such as from the Twitter data, we notice that Ontario has the highest physical activity indicators however Google Trends for this same period points to Ontario having one of the lowest. These contradictions emphasize the differences in the data from these two sources. Interestingly, using Twitter data, again Ontario ranks highest in terms of sedentary behaviour however using Google Trends data, we see that it ranks the lowest in sedentary behaviour. Therefore, the radar charts allow us to quickly infer how these data sources are different and therefore may not necessarily complement each other.

4.7.3 Reddit

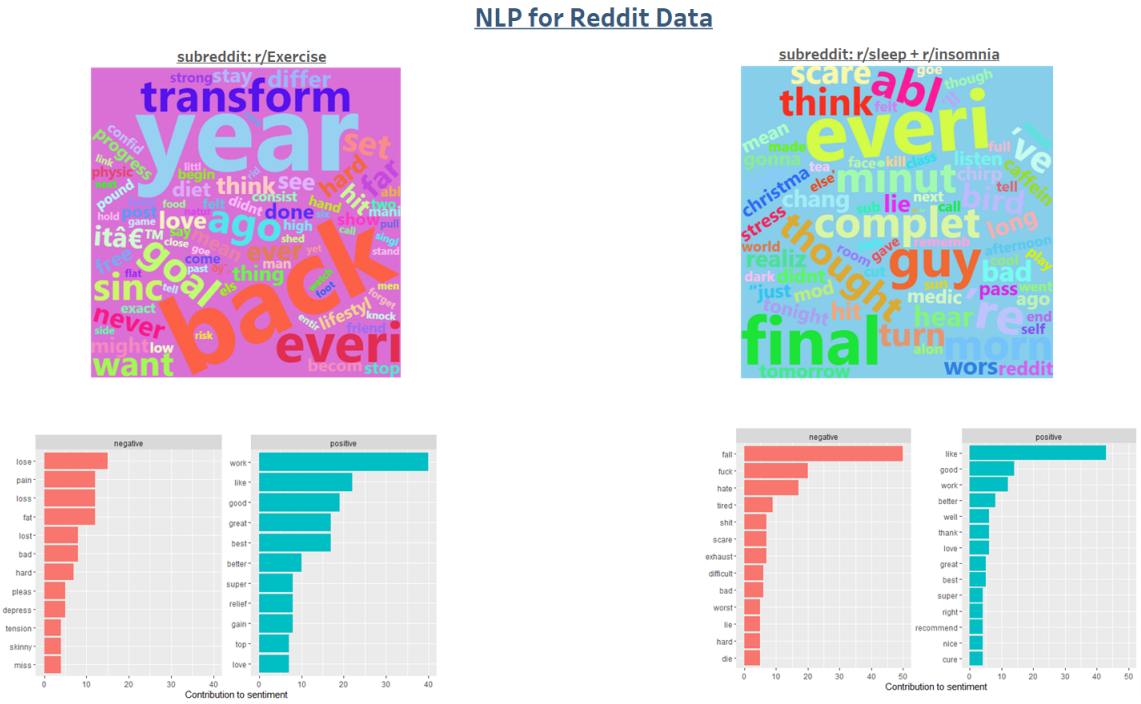

Figure 11: *Reddit*.

The Reddit data that was extracted for the purpose of our analysis ranged from November 2019 to March 2020. This time period encompasses the holiday season as well new year. Scraping through the different discussion titles for the aforementioned subreddits, we generated the below figure (*figure 11)*. The wordcloud in the figure tells us that the common subjects or topics were around fitness, exercising, goal setting, progress, lifestyle, etc. for the discussion on exercise and physical fitness. When compared to the topics from topic modeling of Twitter data (*figure 9*) on physical activity, we notice that some of the topics related to exercise and fitness are also discussed. Although the time period was different and the platform itself is not the same, we can generally find few common subjects being discussed across different social channels.

From Reddit, for the purpose of sleep disorder, we see the topics of discussion are spanning across potentially multiple topics. The prominent terms related to caffeine beverages, long day, stress, Christmas, medical issues, etc., are commonly causing people to have irregular sleep patterns which is a harmful element to human health. Also, the sentiments noted in the wordcloud about fear, scare, remembering someone are also raised in the Reddit forums. This could tell us that general population seems to have a range of reasons for why they do not seem to have a healthy sleep pattern.

A note to be made here, Reddit is an open discussion and new aggregation platform. Contributors on this platform come from different geographical regions, so we might not be always accurately seeing the sentiments of Canadians from all sources of web content scrapping.

4.8 Relation Between Twitter Data and National Survey

One of the query our client wanted us to research on was whether there some relation or source of information that allows us tie back some general findings from the Twitter data. The specific requirement to look at was if health policies and education curriculum would help improve the PASS indicators in Canadian population. For this we note from the bar plots in our figure below (*figure 12*), that the population in the Twitter data seems to be physically active while similar proportion also is experiencing or living sedentary lifestyle. As for sleep disorder, majority Canadians who reported by themselves, said they do no seem to have much sleep disorder problem. To attest to these Twitter findings, we searched for a government conducted survey which would provide us some information to make a relation with social media data analysis. Using the Open Canada's PASS indicator survey dataset, which encapsulates the summary of a detailed survey to study Canadian movement behaviour (physical activity, sedentary behaviour and sleep). This data source is maintained under the open government licence by Public Health Agency of Canada.

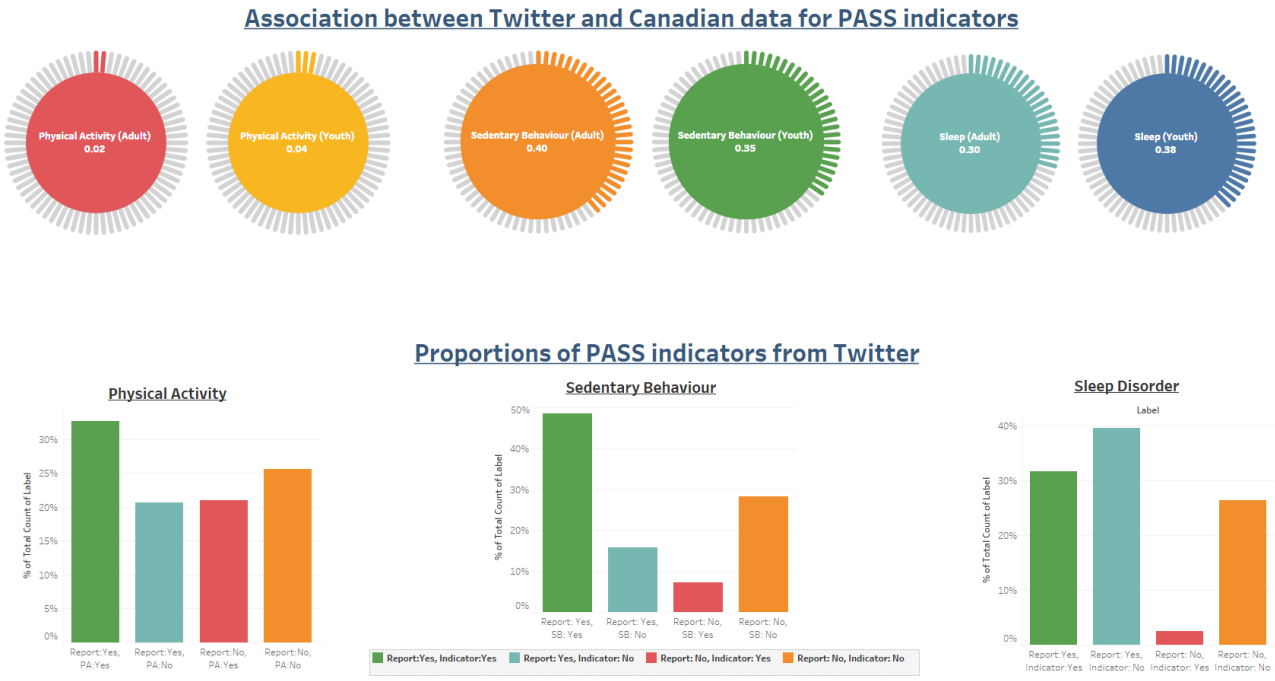

Figure 12: *PASS Survey (2018) vs. Twitter.*

Using this survey, we split it in two groups for each PASS indicator, one for adult (ages 18+) and the other for youth (ages 10-17). We found the following interesting facts which comparatively relates to our Twitter data:

- For physical activity, we that the adults are about twenty-five minutes active in a twenty-four hour period while the youth are active for about 60 minutes in a day and display a healthier lifestyle. There clearly is an effect of educational policy and curriculum by each jurisdiction which makes the Canadian youth lead an active life. We could potentially have a policy tailored towards adults where being active and leading healthy lives would be encouraged and potentially rewarded, perhaps in forms of tax credits. Twitter data highlights Canadians being active but we are missing the attribute of age groups which are actually staying active.

- For sedentary behaviour, we again see adults living more sedentary lifestyle spending about ten hours with no mobility compared to the youth that spend about eight hours in stationary position. The pro-

fessional life of an adult perhaps makes them spend more time at their desk, against youth who might be exuding more mobile and active life. This could be due to activities revolving outside classroom hours. Perhaps policies at grassroots level at work place or related institutions to promote more movement will aid in less posture related ailments and decrease in obesity rates as seen in *figure 8*. Comparing to the Twitter data, we do know that Canadians seem to have quite a bit sedentary lifestyle.

- As for the sleep disorder results from the survey, with no surprise adults do sleep less (seven hours a day approximately) compared to the youth (ten hours a day approximately). This again could be a less strict and disciplined life of an adult compared to the youth who tend to follow a certain norm in their lifestyle. Twitter data says that the Canadians do not seem to have sleep-disorder in general, which might be true but we do miss a larger population in both our data which could reflect the results otherwise. Sampling on a percentage of Canadians, still tells us that the adults in Canada do not seem to be getting a complete set of eight hours of sleep recommended by medical experts. For policy-makers, this would be an interesting arena to explore on how to work with medical experts and implement healthy and sustainable sleeping patterns in the population.

---

5 Timeline

Throughout the semester, there were few preset guidelines in terms of timing for the project work to continue at a smooth sailing pace. Some of these deadlines were defined by Dr.Abad within the scope of application of new concepts as we kept learning them through the term. The other set of deadlines in terms of timing came from the working aspect with our clients. We tried to enforce a very dedicated structure in completing our work in different bits and seek feedback from our clients and the supervising staff.

Synopsis of our timeline:

- We met with clients, at times separately or virtually to share our preliminary analysis and its results on a two to three week intervals. This was to ensure we are meeting their expectations from the queries they had asked us to research about.
- Preliminary results were shown a week in advance of the final presentation with few insights shown a couple of days before the final delivery.
- A virtual meeting in the last week with one representative of the to blend a couple of question into one as they were essentially ask to produce the same thing.
- From the supervising staff, we had a couple of reviews to be submitted for further develop our analysis into more complete and concrete research. These were due for submission in early and late March, 2020.
- Our visual presentation was delivered on April 8th and the narrative report submitted on April 18th,2020.

---

6 Collaboration

Collaborative work was constantly happening between the authors through the entire phase of the project. Additional collaboration was done with the clients on learning their expectations, setting immediate goals to ensure deliverable per at par in terms of quality of results and value from using various credible sources.

Ad-hoc guidance was provided and received from the supervising staff throughout the term. There two fundamental outcomes which were expected from this entire exercise of research for people who were involved.

- With various stakeholders involved in this, each one involved had a significant role to play in guiding this research in the right direction. Clients provided us with a research framework and check if their queries were feasible to work with and whether substantial insight can be gained from the social media data. Supervising staff were critical to our learning and providing feedback in application of various techniques to generate value from the data.
- The second key piece in this collaboration was to ensure a fundamental understanding and learning from the course work. Drawing from these learning, we wanted to apply different techniques to enhance our results and harness all the value from the Twitter dataset and articulate more variants from other data sources.

---

7 Conclusion

- From Twitter data we find cities with the lowest amount of physical activity tweets (Aurora, Ajax, Gatineau) and sedentary tweets (Ottawa, Edmonton, London).
- People from smaller cities generally just tweet less irrespective of physical, sedentary or sleep disorder tweets.
- Tweets per-capita for physical activity and sedentary behaviour shows very little overlap with the percentage of tweets per city therefore, per-capita may be better to use.
- There is a general lessening gradient in the prevalence of PASS indicators when you get further from a city except for Saskatchewan and New Brunswick .
- There is a negative correlation (Pearson's coefficient of -0.21) between rate of physical activity tweets and Obesity Rate (from CCHS data)
- There is a positive correlation (Pearson's coefficient of 0.77) between sedentary tweets and obesity rate (as sedentary tweets increase, obesity rate increases).
- Inferring from google trends and Reddit, we see that very few Canadians are inflicted with sleep disorders, physical activity such as exercising at gyms is prevalent and there is not a lot of overlap between information from google trends and Twitter data.
- From natural language processing(NLP), we see that Canadians are prone to swearing in their tweets but generally there is more positive sentiment rather than negative sentiment and topic modelling includes love for hockey, working out and making time for these activities.

8 Limitations and Future Recommendations

Throughout the project and research, we came across a series of limitations that we thought could be brought forward as some recommendation from a consulting view for our clients to think about exploring. Perhaps we could render more help applying more technical mechanics to more queries they might have later and work with public institutions to develop strategies dealing with population health and developing a systematic surveillance methods.

- When analyzing our data, we did not find any similar timeline survey data available nationally that would be extremely useful to review results with an external source for the Twitter data. Conducting the research during the same time when national surveys are conducted, we could produce models to check accuracy from the national data collected and vice versa. Thus for our project, using CCHS data for which the most recent data available was 2016 would not yield the most accurate results for comparison purposes. However, the lack of availability of the the most recent and relevant national data also sheds light into the need for more recent and real-time alternative sources of data such as Twitter PASS indicators data used for this project.

- Other limitation we found is, just because we have various sources of data from different social platforms, not all capture the same topics or state of data in general. This means, we cannot capture the same information from Google Trends and Reddit and directly compare this to data from Twitter. Albeit sophistication in machine learning and deep learning happening daily, we still cannot get the information in a similar subject form or quantitative scale. This could make the reliance on social media data a little ambiguous and might not allow for a very accurate findings. Our recommendation is that, sole dependency on the social media should be possibly avoided to ensure good and informative surveillance measures. Also, certain social platforms will include global contributors, making it extremly challenging if the content reflects Canadian population or just general global population.

- Geographical location of users on social media is a great asset to have in the data. This invaluable variable must be included if possible in all data collected. This allows us to check with an increased level of granularity which regions are going to be needing help in health context in the near future[BRC+11].
- As per our clients requirements, their preference was to have PASS indicators at the granularity level of cities. We were able to provide this for all questions except whether Twitter PASS surveillance results match up with the reported rates of obesity (and related health concerns) for those areas. We were unable to provide city-level granularity for this question because the CCHS[5] data was very sparse in regards to containing data for cities. Instead the CCHS data only contained data at the province level and hence this was a limitation. There was also no other data available with this information.

Social media data is definitely a valuable potentially useful data source in health surveillance and can be widely used for digital health optimization, but it should not be treated as the only sophisticated source and replace some of the more fundamental and traditional data collection methods.

References

[BRC+11] Maged N Kamel Boulos, Bernd Resch, David N Crowley, John G Breslin, Gunho Sohn, Russ Burtner, William A Pike, Eduardo Jezierski, and Kuo-Yu Slayer Chuang. Crowdsourcing, citizen

[5] CCHS - Canadian Community Health Survey

sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. *International journal of health geographics*, 10(1):67, 2011.

[Eke11] Paul I Eke. Using social media for research and public health surveillance. *Journal of dental research*, 90(9):1045, 2011.

[MBD+12] Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data: the management revolution. *Harvard business review*, 90(10):60–68, 2012.

[TTSR14] Suppawong Tuarob, Conrad S Tucker, Marcel Salathe, and Nilam Ram. An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages. *Journal of biomedical informatics*, 49:255–268, 2014.