Public Health Surveillance on Physical Activity via Twitter

Winter Activities, Related Injuries, and Sentiment Analysis of Tweets Across Canada

Minki Lee, Christy Sarmiento

Data Science 624-Winter 2021

April 15, 2021

Key Findings of the Project

Key Findings

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

1. The growing interest on subjects related to "Falling" for the past four years were driven by fall prevention initiatives and personal injury claims.
2. Falls on ice is the leading cause of sport and winter injury hospitalizations in Canada. In general, volume of injuries by province is associated to the proportion of volume of tweets by province.
3. Among the self-reported tweets, Ontario has 803 tweets, which has the highest number of tweets across Canada. Over 42% of these physical activity tweets are from the City of Toronto and Ottawa.
4. Among tweets from Western and Eastern Canada, a word "hockey" forms common centers of nodes.
5. Sentiment analysis comparing Twitter with Reddit revealed that most words in a negative category were not negative.

1 Introduction

Physical activity is an essential part of a healthy lifestyle. It reduces the risk of developing chronic conditions such as heart disease conditions, type 2 diabetes, depression, and anxiety. The government of Canada created the Physical Activity, Sedentary Behaviour and Sleep (PASS) Indicator Framework to better understand Canadians' health and provide useful guidance for the public to make healthier choices. To further support the PASS Indicator Framework, we examined public health surveillance using a social platform such as Twitter.

Twitter is one of the most-used social media platforms where people share thoughts, ideas, and opinions, communicating in short messages. Since it provides a real-time source of public health information, Twitter has been popular in public health studies. We performed Natural Language Processing (NLP) to analyze and extract meaningful insight stored in physical activity-related tweets and fitness subreddit posts. To enhance the relevance of our research questions we collected supporting data from internet sources such as Google Trends, Reddit, and the Canadian Institute of Health Information statistical reports.

1.1 Research Questions

In this study, we will investigate a collection of physical activity-related tweets and fitness subreddit posts to answer the following questions:

1. How physically active are Tweeters across Canada during the time of data collection? What are the health risks and benefits of staying active at this time of the year?
2. Do volume of Tweeter followers influence Tweeters to stay active?
3. Among the self-reported tweets, which provinces tweet the most about physical activity? Further identify which cities are more physically active.
4. What are the most popular tweeted bigrams (pairs of words) related to physical activity in Western, Eastern, and Northern Canada?
5. What is the overall social sentiment in the physical activity tweets and fitness subreddit posts?

1.2 Project Motivation

This study aimed to analyze the Twitter dataset as part of the Physical Activity, Sedentary Behaviour and Sleep (PASS) Indicators. This report focused on Physical Activity to provide actionable data on the health benefits and prevalence of physical activity-related injuries. Furthermore, we would like to provide information to motivate individuals and communities to adopt safe practices and develop healthy behaviours. We plan to draw insights from the study that would support the development of indicators under the Family/social environment group, which includes Community Norms and Presence, and Type of Barriers for Physical Activity[1].

Health care providers, educators, researchers, and policy makers are utilizing digital tools, specifically social media, in conducting public health surveillance projects. Monitoring the trends in the prevalence of physical activity is important for understanding population health risk and planning and evaluating policies and programs to promote physical activity [2]. We generally assume that active Twitter users would inform their followers of their almost real-time activities and that follower count is often used as an indicator of a person's popularity but is this equally valid when measuring the influence of these follower counts on a person's behaviour to stay active.

[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.

[2] Liesure-Time Physical Activity, Statistics Canada, https://www150.statcan.gc.ca/n1/pub/82-229-x/2009001/deter/lpa-eng.htmim

Keeping an active lifestyle provides health benefits but also confers health-related injuries. Unintentional injuries are the leading causes of Canadians aged 1 to 44 years and the third leading cause of death among all ages combined in 2019 [3]. Alongside this, there had been increased interest over the past four years on topics about "falling" on Google Trends. An interesting topic related to this was lawsuits. Law firms are increasingly investing in research on the incidence of slips and falls.

1.3 Problem Definition

A Twitter data set for physical activity is provided in a CSV format, with approximately 4600 rows and only the user IDs, text, date created, the name of the city and province, and the number of followers. The provided dataset on D2L was labelled.

To address the project's goal, information on Twitter user's demographics and geographical location is vital. The frequency of related tweets within a specific location and time will also help in drawing insights. Handling missing information, incorrect locations of the tweets and wrong data format would require a good amount of data cleaning and manipulation. The span of time covered in this dataset will also affect the relevance of the project's outcome. Most importantly, identifying the integrity of the tweets shared would be a challenge.

2 Methodology

2.1 Data Source Cleaning and Transformation

Datasets

Physical Activity.csv: The file has 110 columns and 4,696 rows, including column headers. It contains a collection of tweets in Canada from October 12, 2019, to December 1, 2019, provided by Data Intelligence for Health (DIH).

DATA 624 (W21)-COURSE PROJECT-DATASET - Physical Acticity.csv: This file is a reduced version of the Physical Activity.csv file. Contains a unique user id, message, and a label that classifies whether the tweet was Self-Reported, Physical Activity, or Not Clear.

The datasets were uploaded to Tableau Prep Builder to perform initial inspection and cleaning. We first selected columns relevant to the study such as user id, messages, hashtags, geographical data, and dates from the Physical Activity file. We also included in our selection 22 fields that stored hash-tagged text only. Another field that is relevant to the study is geographical location fields. Aside from geographical coordinates, we also selected fields containing location names. There were no standard string formats, but we were able to split the names to extract cities. The output from this dataset was joined to records from the original twitter dataset file where we only selected records that were labelled "Self-report: Yes, Physical Activity: Yes" and "Self-report: No, Physical Activity: Yes". There were 1,381 records that were selected by our criteria which is 29 percent of the file. We exported the output file into CSV and extract file types. These were then further manipulated in R, Python, and Tableau accordingly based on the topic of interest.

Followers_Tweets_Count.xlsx: This is extracted from the Tweeter dataset. It contains aggregate data of follower counts, total tweets per user, tweet message, record id, user id, label, province, and a new column that indicates if the tweet message was a physical activity or not.

WordFreqTweet.csv: This output CSV file from python contains the frequency of words from the processed Tweeter messages.

[3] Statistics Canada. Table 102-0561. Leading causes of death, total population, by age group and sex, Canada, annual [Internet]. Ottawa (ON): Statistics Canada; [cited 2016 Apr 20]. Available from: http://www5.statcan.gc.ca/cansim/a26?lang=engid=1020561

Reddit_Fitness_Posts.csv: This CSV file is from the top 1000 posts from the fitness subreddit and exported from the Google Colaboratory which contains features including the post title, id, url and body. There are 992 rows and 8 columns.

Additional data sources: MultiTimeline.csv and geoMap.csv were downloaded from Google Trends that contained the interest over time and sub-region data from March 20, 2017, to March 7, 2021. No further cleaning or transformation was performed. A 2018-2019 summary report from Canadian Institute of Health Information on the Injury and Trauma Emergency Department and Hospitalization Statistics was obtained from the CIHI website. There was no further cleaning or transformation performed on these datasets as well.

2.2 Data Analysis

Exploratory Data Analysis

The dataset Followers_Tweets_Count.xlsx was uploaded to Python for further data exploration. This dataset excludes records that have a "Not Clear" value in the Label column. The records were grouped to find out the proportion of messages by Province and labels. The length of messages was also explored and visualized in a histogram to find out the distribution. A scatterplot was generated to visualize the relationship between the number of followers and tweets. The average tweets per user in each Province was also visualized in a histogram.

Text pre-processing and classification

The records were labelled as to whether it is a physical activity message or not. The machine learning algorithm that we've decided to implement is Naive Bayes and Random Forest. First, we convert the raw messages into vectors. The messages were split into individual words and returned a list. Using NLTK library, common words were removed by "stopwords". The string.punctuation was passed through the list performed, then joining the words divided with space. The classification models used were Naive Bayes and Random Forest algorithms, and performance on both was also evaluated on test data.

Comparison Between Provinces

The reduced version of the Physical Activity.csv file was imported into Python for further cleaning. We focused on records labelled "Self-report: Yes", and dropped any duplicated user IDs in the dataset. We then split string by comma and extract the province from "Place.Full Name" column. We found some missing/incorrect data in "Place.Name" column; therefore, the errors were manually corrected. This CSV file was imported into Tableau for visualization and the examination of which provinces tweet the most about physical activity.

Text Mining

Bigram Modelling

A total of three bigrams were produced to each represent Western, Eastern, and Northern Canada. The original data was first filtered using Python to include the columns of interest for this study; unnecessary columns were dropped. The unique Twitter IDs were extracted, and an extra column consisting of provinces of tweets was added to the cleaned DataFrames so that texts were grouped based on the provinces into Western, Eastern, and Northern Canada. Three Pandas DataFrames were exported to a CSV file for language modeling in R Studio. After importing into R Studio, we first created a collection of documents. We performed the following data cleansing steps: 1. Used a content transformer to eliminate colons and hyphens. 2. Removed punctuation marks. 3. Transformed the corpus to lower case. 4. Removed stopwords using the standard list in the 'tm' package. 5. Removed all numbers and extra whitespaces that have been generated during data cleaning. Furthermore, uninteresting and irrelevant words, such as 'of a' and 'i am', were removed to make the visualization interpretable. The relationships between words were visualized and imported into Tableau. Furthermore, uninteresting and irrelevant words, such as 'of a' and 'i am', were removed to make the visualization interpretable. The relationships between words were visualized and imported into Tableau.

Google Trends and Canadian Institute for Health Information data

To provide context on the Tweeter data, we looked at additional internet data sources. We selected the word "Falling" for the period Mar 2017 to March 2021[4]. The suggested related topics were "prevention", sidewalk, lawsuit, and hematoma. These terms are significant to the topic that we are looking for. A meaningful triangulation among these additional datasets is plausible given the feature term is related to physical activity, geography, and period of data coverage are closely related.

Sentiment Analysis on Twitter and Reddit data

We used a natural language processing technique, called sentiment analysis, to determine whether Twitter and Reddit data are positive or negative. For Twitter data, we filtered the reduced version of the Physical Activity.csv file using Python and focused on records labelled "Self-report: Yes". We then dropped any duplicated tweets, and extracted a text column from the dataset. The CSV file was imported into R Studio for sentiment analysis. We followed the same cleansing steps taken as in bigram modelling. The "tidytext" package contains several sentiment lexicons, but we used Bing Liu's lexicon to evaluate the opinion or emotion in texts; the bing lexicon categorized words into positive and negative categories.

For Reddit data, we scraped it from the Reddit platform using the Python Reddit API Wrapper (PRAW). We first created a Reddit app and filled in a name, description and redirect URI. After we found the authentication information needed to create the praw. Reddit instance, we retrieved the top 1000 posts from the fitness subreddit by specifying "fitness" as the subreddit name using the Google Colaboratory and saved them into a .csv file. The reason for choosing fitness subreddit was because it was active with 8.1 million members and it was one of the biggest health and fitness community. The CSV file was imported into R Studio for sentiment analysis, and the same cleansing steps were performed as above. The sentiment analysis on the physical activity and fitness data were imported into Tableau for data visualization for comparison.

3 Performance Measurements

To evaluate the performance of the classification algorithms, we did a train_test_split technique. The split was 80% training data and 20%test data.

In the Random Forest classification, the parameters used were n_estimators = 100, random_state = 42, max_depth = 1,000, min_sample_split = 0.001 and bootstrap was set to 'True'.

[4] https://trends.google.com/trends/explore?q=%2Fm%2F04f32c6date=today%205-ygeo=CA

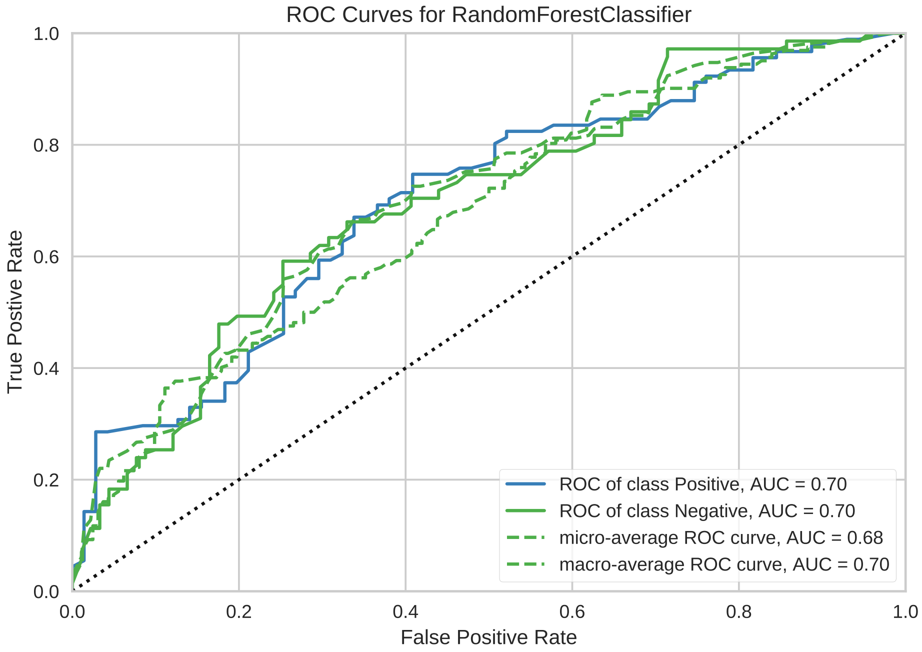


Figure 1: ROC Curves for RandomForest Classifier

After fitting the training data in MultinomialNB(), we tested the model and achieved an AUC of 0.70 on both classes.

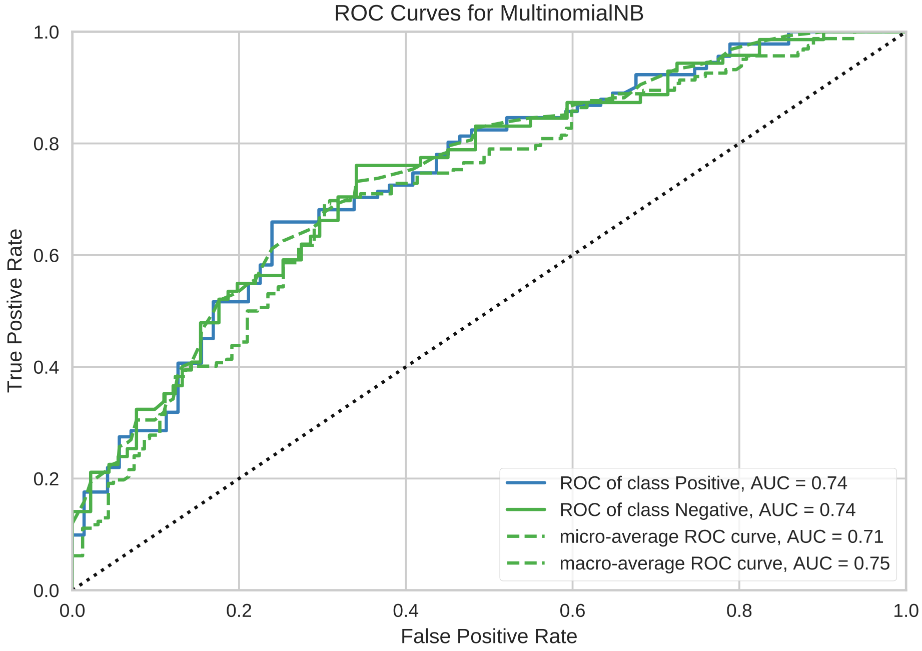


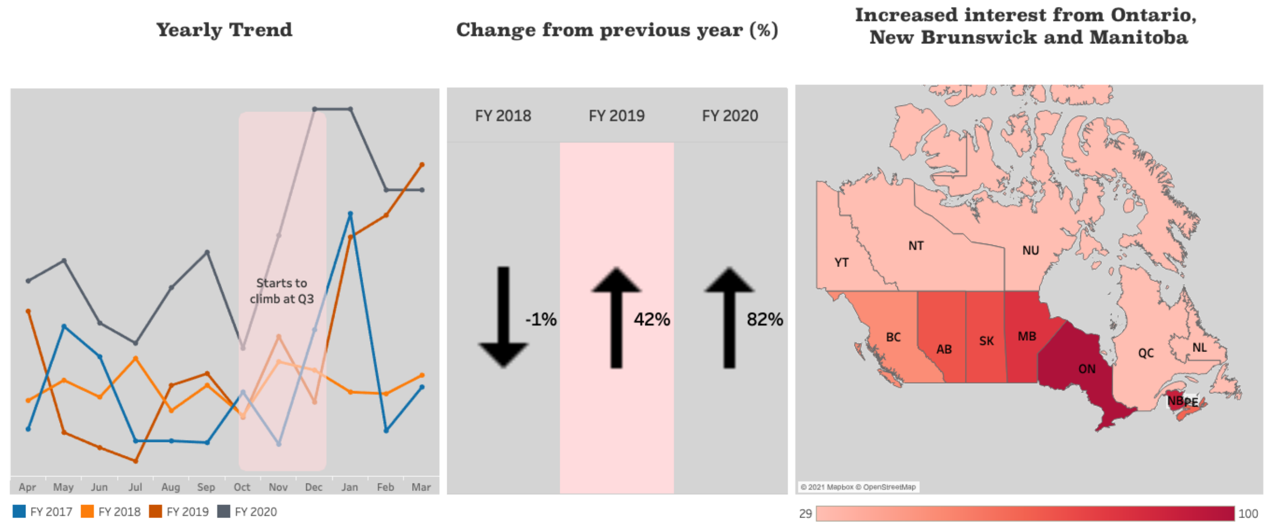
Figure 2: ROC Curves for MultinomialNB

Both algorithms achieved an accuracy score of 0.6049 and f-core of 0.6049. From the results of the evaluation, we used the variable "PHYSICALACTIVITY" to identify records that are included in the analysis.

4 Results

Physical Activity and Related Injuries

Google Trends data shows an increase of interest over the past four years with trends starting to climb in the fall seasons and early winter, from September to December, with highest attention from Ontario, New Brunswick and Manitoba. This was driven by the health and legal services sector with specific topics on fall prevention and personal injury claims.



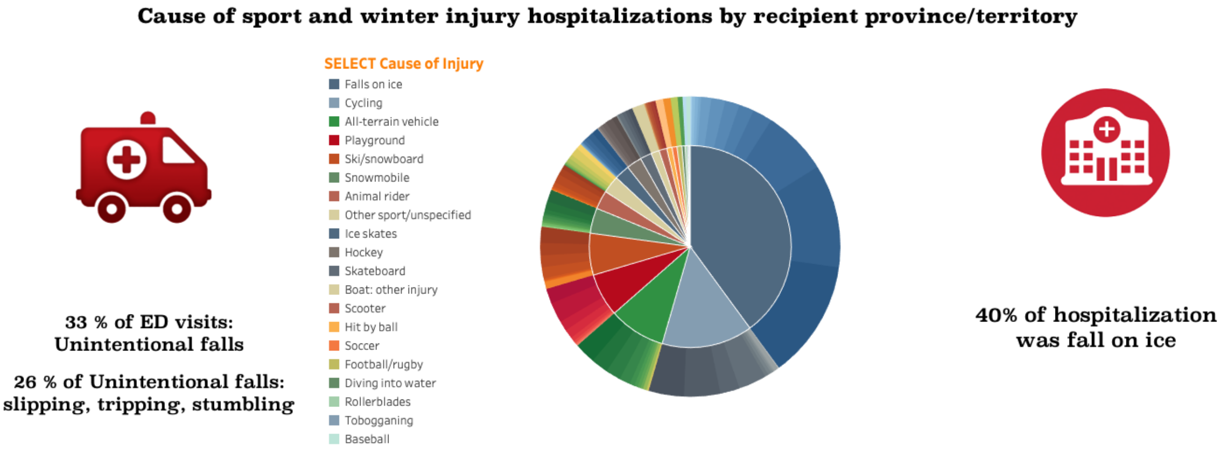
Yearly Trend

Change from previous year (%)

Increased interest from Ontario, New Brunswick and Manitoba

Figure 3: Google Trends: Keyword search on "Falling", March 2017 to March 2021

Approximately 33% of reported emergency department visits from Prince Edward Island, Nova Scotia, Ontario, Saskatchewan, Alberta and Yukon were "Unintentional Falls" from the FY2018-2019 summary report[5]. That is 686,107 ED visits where slipping, tripping and stumbling was 26% of the Unintentional Falls group.



Cause of sport and winter injury hospitalizations by recipient province/territory

33 % of ED visits: Unintentional falls

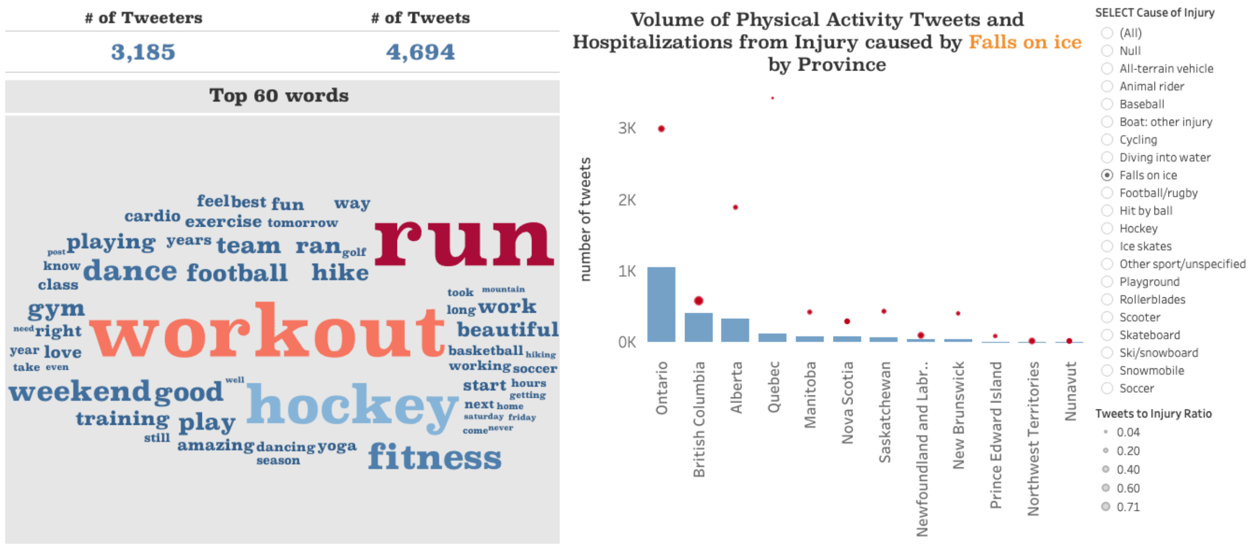
26 % of Unintentional falls: slipping, tripping, stumbling

40% of hospitalization was fall on ice

Figure 4: CIHI: Injury and Trauma Emergency Department and Hospitalization Statistics, 2018-2019

The most relevant words were "workout", "run", "hockey", "fitness", and "weekend". Therefore, we assumed that the number of tweets could represent how active a Twitter user. There is a close relationship between the volume of hospitalizations from injury and how active the population is. Except for the Province of Quebec, the visual representations generally showed linearity of the two measures

[5] Canadian Institute for Health Information. Injury and Trauma Emergency Department and Hospitalization Statistics, 2018–2019. Ottawa, ON: CIHI; 2020.



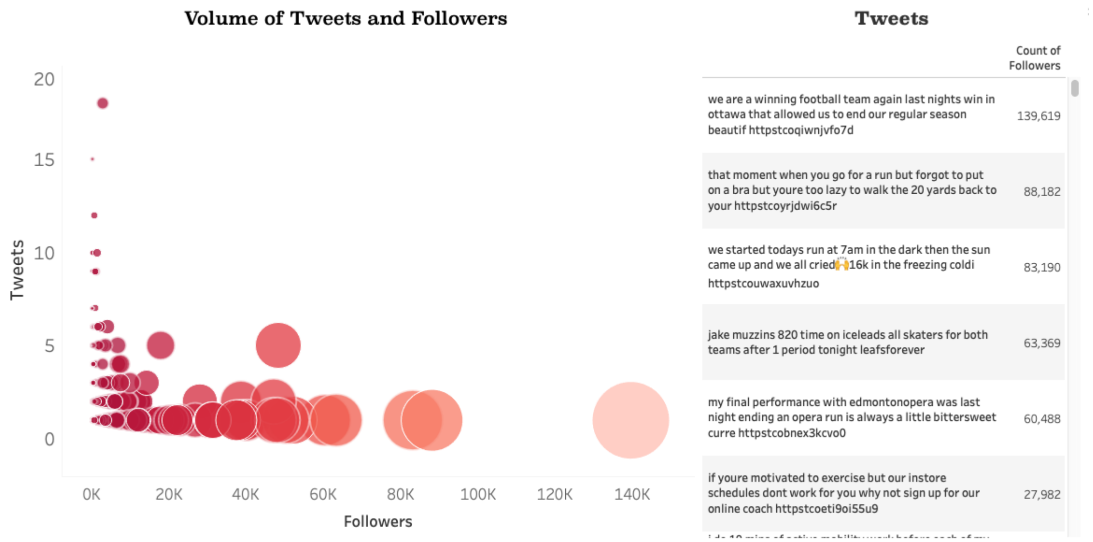
# of Tweeters	# of Tweets
3,185	4,694

SELECT Cause of Injury
○ (All)
○ Null
○ All-terrain vehicle
○ Animal rider
○ Baseball
○ Boat: other injury
○ Cycling
○ Diving into water
● Falls on ice
○ Football/rugby
○ Hit by ball
○ Hockey
○ Ice skates
○ Other sport/unspecified
○ Playground
○ Rollerblades
○ Scooter
○ Skateboard
○ Ski/snowboard
○ Snowmobile
○ Soccer

Figure 5: Wordcloud of Frequently Used Word and Combo-chart of Volume of Tweets and Hospitalization Cause

Volume of Followers to Frequency of Tweets

There is an inverse relationship between the number of followers and tweets of an active Twitter user. The amount of followers doesn't influence a person to be more active. The users with more tweets indicate more active participation in a sport, fitness routine, or workout activity. In contrast, the tweets of a user with a high number of followers were more passive participation.



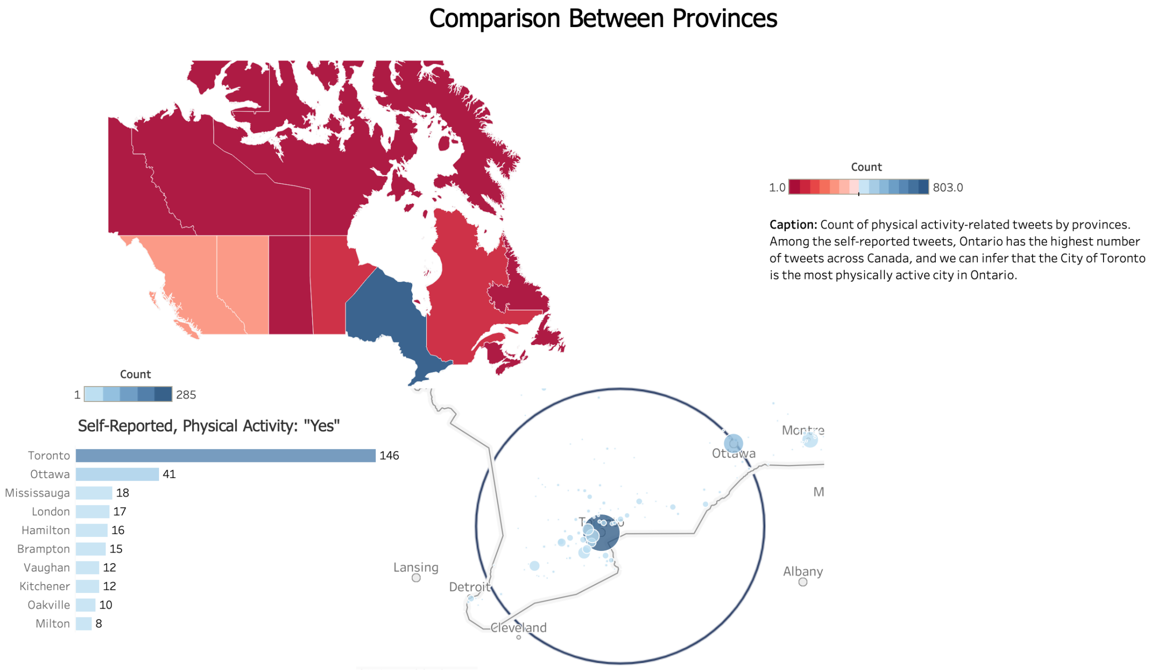
Tweets

	Count of Followers
we are a winning football team again last nights win in ottawa that allowed us to end our regular season beautif httpstcoqiwnjvfo7d	139,619
that moment when you go for a run but forgot to put on a bra but youre too lazy to walk the 20 yards back to your httpstcoyrjdwi6c5r	88,182
we started todays run at 7am in the dark then the sun came up and we all cried🙌16k in the freezing coldi httpstcouwaxuvhzuo	83,190
jake muzzins 820 time on iceleads all skaters for both teams after 1 period tonight leafsforever	63,369
my final performance with edmontonopera was last night ending an opera run is always a little bittersweet curre httpstcobnex3kcvo0	60,488
if youre motivated to exercise but our instore schedules dont work for you why not sign up for our online coach httpstcoeti9oi55u9	27,982

Figure 6: Volume of Tweets and Followers

Comparison Between Provinces

Among the self-reported tweets, Ontario has 803 tweets out of 1684, with the highest number of tweets across Canada, whereas Yukon has the least number of tweets. British Columbia ranked second with 277 tweets and followed by Alberta with 253 tweets. Provinces with bigger population counts are expected to have more tweets, but it wasn't for Quebec, which is the second-most populous province in Canada. According to Statista, Twitter was one of the least used social networks in Quebec[6]. Also, it is important to keep in mind that our data only contains English tweets. We observed that over 42% of physical activity tweets were from the City of Toronto and Ottawa. We can infer that people in Toronto are more physically active than in other cities in Ontario. This can be explained by the city having many opportunities for citizens to be physically active, including well-maintained bike lanes, pedestrian lanes, parks throughout the city, and gyms. Furthermore, watching others exercise can also help motivate people to be physically active.



Caption: Count of physical activity-related tweets by provinces. Among the self-reported tweets, Ontario has the highest number of tweets across Canada, and we can infer that the City of Toronto is the most physically active city in Ontario.

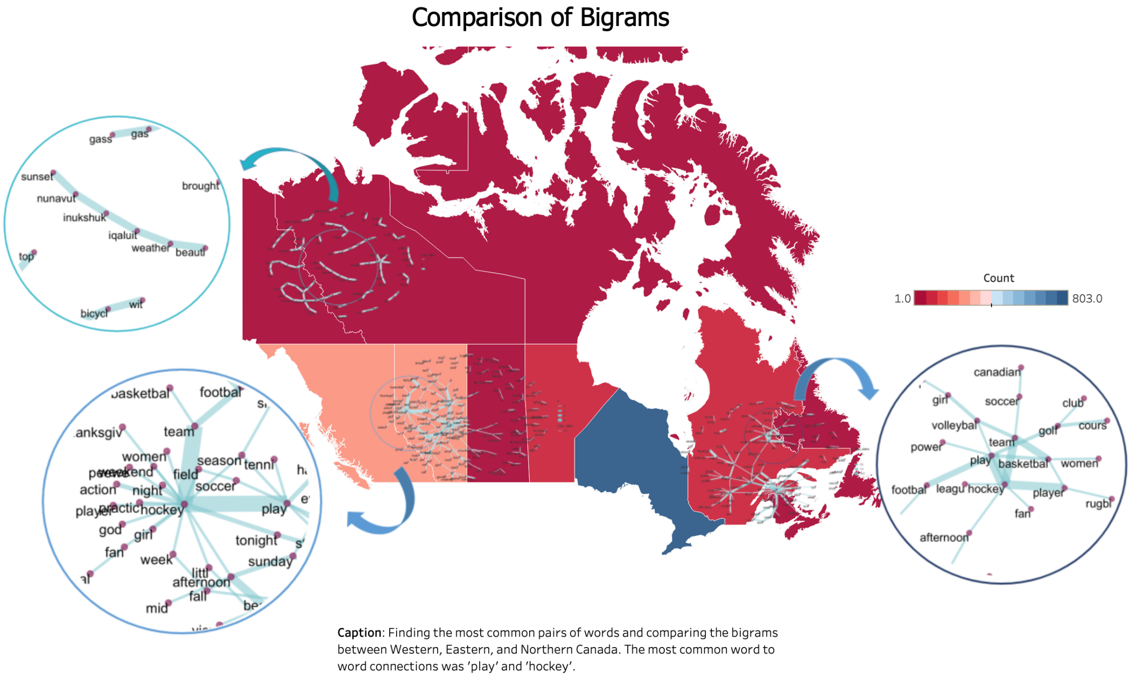
Figure 7: Count of physical activity-related tweets by provinces and cities

4.1 Unsupervised Learning

The Most Common Bigrams

Bigram of Western, Eastern, Northern Canada provided insight into multiple relationships that exist among words. The bigrams about physical activity included a variety of sports such as hockey, running, basketball, dance, and yoga. Popular bigrams include “spend time", ‘afternoon hike’, and ‘workout tonight’, and the most common word to word connections was ‘play’ and ‘hockey’. As we expected, there were more bigrams about nature in Western than in other parts of Canada. The bigrams showed that many Tweeters exercise or do outdoor activities in the morning, afternoon, and weekends. Due to a limited number of texts, we could not generate many meaningful bigrams; therefore, we looked at bigrams that have at least one occurrence.

[6] Tankovska, H. (2021, Jan 25). Canada: social network usage reach 2020, by province. Statista. https://www.statista.com/statistics/262804/social-networks-used-by-internet-users-in-canada-provinces/

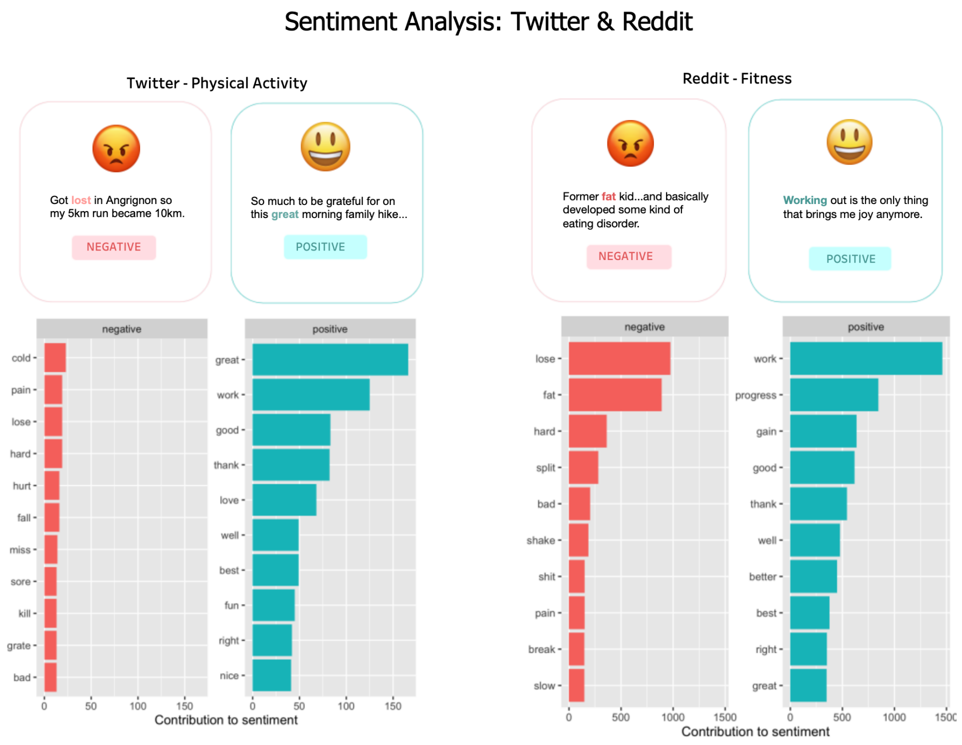

Comparison of Bigrams

Caption: Finding the most common pairs of words and comparing the bigrams between Western, Eastern, and Northern Canada. The most common word to word connections was 'play' and 'hockey'.

Figure 8: Relationships between words (bigrams)

The Overall Sentiment of Physical Activity and Fitness

We analyzed a body of text to understand the social sentiment for physical activity and fitness using sentiment analysis. We looked at the top 10 words contributing to positive and negative sentiment. For Twitter data, the most common positive words were "great", "work" and "good", whereas the most common negative words were "cold", "pain" and "lose". For Reddit data, the most common positive words were "work", "progress" and "gain", whereas the most common negative words were "lose", "fat" and "hard". Interestingly, we discovered that most words in a negative category were not negative. For example, the word "lose" from both data indicates weight loss from exercise, and the word "hard" explains a challenging but good workout. We observed similar words between the two platforms; however, more fitness-related words in Reddit data, whereas positive adjectives in Twitter for describing nature, time with family, and various activities. The overall sentiment about physical activity and fitness was positive.



Got lost in Angrignon so my 5km run became 10km.

So much to be grateful for on this great morning family hike...

Former fat kid...and basically developed some kind of eating disorder.

Working out is the only thing that brings me joy anymore.

Caption: Sentiment analysis on tweets about physical activity and Reddit posts about fitness. Most words in a negative category were not negative.

Figure 9: Most common positive and negative words

5 Limitation

When comparing the number of tweets between provinces, we could have normalized the data by population so that they may be compared in a meaningful way since provinces with bigger population counts are expected to have more tweets. This would have minimized the influence of population. It is important to understand that it is challenging to conclude Canadians' physical activity levels based on a few social media platforms, but instead, they can be beneficial for monitoring physical activity levels. Although sentiment analysis is extremely useful in social media monitoring, it cannot handle details such as sarcasm or irony, therefore, cannot accurately analyze people's emotional states or moods. We need to apply advanced analysis approaches to overcome these limitations of social media data.

6 Conclusion

Twitter remains one of the most popular platforms in health research, and it has become an important health resource for assessing and monitoring public health surveillance. In this study, we examined public health surveillance using Twitter data and used other multiple data sources such as Google Trends, Reddit, and the CIHI to support our observation's reliability.

We discovered that the number of injuries was associated with the proportion of tweets by each province, and the number-one cause of sport or winter injuries was due to falls on ice. The analysis showed that the number of followers does not influence an active Twitter account to engage in physical activity. As we expected, among the self-reported tweets, Ontario had the highest number of physical activity-related tweets and hockey was considered to be the most popular sport in Western and Eastern Canada. We performed sentimental analysis on Twitter and Reddit data to determine whether the overall attitude or perception toward a physical activity was positive or negative, and we observed that the overall sentiment was positive. There were not enough tweets from Northern Canada; therefore, we were not able to extract meaningful information from bigrams and make a comparison with Western and Eastern Canada. Furthermore, we noticed that there were not many places for people to discuss the physical activity. In fact, there was only one subreddit on the topic of physical activity, but not a single post on the page, and thus we studied the fitness subreddit for our analysis instead.

Health care professionals and policy makers can use our findings to gain valuable insights and have a better in-depth understanding of how Canadians consider physical activity and behave. Moreover, this study can be used to target interventions and health promotion efforts to improve Canadians physical activity levels.