Population-level Indicators of Physical Activity, Sedentary Behaviour and Sleep in Canada based on Twitter

by

Olivier Nguyen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Sciences
in
Electrical & Computer Engineering

Waterloo, Ontario, Canada, 2018

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Social media platforms contain large amounts of freely and publicly available data that could be used to measure population characteristics across different geographical regions. Analyzing public data sources such as social media data has shown promising results for public health measures and monitoring. This thesis addresses challenges in building systems that collect high-volumes of data from social media platforms. More specifically, we look at Twitter data processing, filtering, and aggregation to provide population-level indicators of physical activity, sedentary behavior, and sleep (PASS). In the first part of the thesis, we go over the whole machine learning pipeline built: (i) Twitter data collection from November 2017 to May 2018; (ii) data preparation through manual annotation, keyword filtering, and an active learning technique for the labelling of 10,283 tweets; and (iii) training a classifier to identify PASS related tweets. Training the model involves building an initial classifier to efficiently find relevant tweets in subsequent annotation iterations. Our classifiers include an ensemble model consisting of several shallow machine learning algorithms, along with deep learning algorithms. In the second part of the thesis, we look at the performance of different solutions. We provide benchmark results for the task of classifying PASS related tweets for the various algorithms considered. We also derive health indicators by aggregating and computing the proportion of classified tweets by province and compare our metrics with the prevalence of obesity, diabetes and mood disorders from the Canadian Community Health Survey. Our work shows how machine learning can be used to complement public health data and better inform health policy makers to improve the lives of Canadians.

Acknowledgements

I first would like to thank Dr. Joon Lee and Dr. Mark Crowley for their supervision and continuous support throughout my master's study and research. I am incredibly grateful for the opportunity to have learnt from such great researchers.

I would also like to thank Amine, Emma, Laura, Côme, Ben, and the rest of my friends and labmates who have helped me over the past two years. In my time here, I was fortunate enough to have been surrounded by wonderful people who inspired me, helped me grow, and with whom I had many great moments. I thank you all for the fun and laughter; I've sincerely enjoyed my time here.

I am also thankful to Dr. Jesse Hoey and Dr. Allaa Hilal for taking the time to read my thesis.

Last but not least, I would like to thank my parents for their unparalleled love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AMT Amazon Mechanical Turk 13, 57

AUROC Area Under the Receiver Operating Characteristic 33

BMI Body Mass Index 32

BRFSS Behavioral Risk Factor Surveillance System 9

CCHS Canadian Community Health Survey 8, 32, 52, 53

CHMS Canadian Health Measures Survey 53

CNN Convolutional Neural Networks 15, 16, 24, 28, 30, 40, 44, 52

LDA Latent Dirichlet Allocation 14, 15

LIWC Linguistic Inquiry and Word Count 15

LR Logistic Regression 26

LSTM Long Short Term Memory 16, 29, 40, 45, 48, 52

ML machine learning 4

NB Naive Bayes 26

NLP Natural Language Processing 15, 16

NN neural network 15

PASS Physical Activity, Sedentary Behaviour and Sleep 1, 4–6, 9, 11, 17, 19–22, 25, 27, 31, 32, 34, 38–40, 42, 44, 51–57

PHAC Public Health Agency of Canada 1, 8, 53, 54

RF Random Forest 26

RNN Recurrent Neural Network 16, 24, 29

SVM Support Vector Machine 13, 26

TF-IDF Term Frequency-Inverse Document Frequency 14, 25, 47, 48, 50

U.S. United States 9, 14

Word2vec Word to vector 16

XGB XGBoost 26

Chapter 1

Introduction

1.1 Motivation

In today's digital world, people are getting less physical activity, and are spending increasingly more time in front of computers and electronic devices. A less active society has in part led to a rise in obesity, diabetes, heart disease and mental health issues in Canada [8]. Yet, physical activity is only one variable of many that affect health, thus developing guidelines and a framework to identify what makes a healthy lifestyle is a difficult task.

In the past, routine reporting in Canada focused on moderate to vigorous levels of physical activity. However, recent studies have found that activities at all levels affect health [21, 58]. The Public Health Agency of Canada (PHAC) contends that reporting on a range of indicators permits a clearer picture of Canadians' health and well-being. For instance, light Physical Activity, Sedentary Behaviour and Sleep (PASS) are now also included in health monitoring in Canada [41]. PASS information can be used to develop effective policies and programs aimed at improving the population's engagement in physical

activity and overall health. Most public health data comes from Canadian health surveys, but traditional methods of collecting public health data are costly and slow. Hence, there is a need to modernize the monitoring of Canadians' physical activity levels.

1.2 Why Study PASS?

The average person today spends most of their time stationary, whether it is work hours at a desk or relaxed time in front of a screen (e.g., a television, a computer, a smart-phone). While our daily lives involve varying intervals of physical activity and sedentary behaviour (i.e., sitting and sleeping), the amount of time spent in each of these domains has a direct impact on health, well-being, and quality of life. Recent studies show that the average amount of time spent sedentary among Canadian adults and children is 10 and 8.5 hours per day, respectively [11, 10]. These statistics are alarming given that the recent Canadian 24-Hour Movement Guidelines recommend no more than 2 hours of screen time for children and youth [54].

Similarly, surveys in past years suggest that 1 in 4 Canadian children are not getting enough sleep i.e., less than the recommended 8-11 of sleep per night [54]. Good, quality sleep is a crucial part of a healthy lifestyle, and the amount of sleep an individual gets has been associated with various health outcomes, including chronic diseases and increased chance of death [7].

More effort is needed in order to promote healthy living and behaviours. Indeed, it is important for public health organizations to effectively monitor the health of a nation. The development of new technologies and new methodologies provides an opportunity to improve the prevention of chronic diseases and injuries. Non-traditional data sources like

social media could help in understanding population trends in physical activity, sedentary behaviour and sleep, and identify populations at risk of developing diseases such as diabetes or obesity. Users on social media often share their thoughts, feelings, and experiences, which can be used to track attitudes and behaviours. A more advanced public health surveillance could better inform those responsible for creating programs and policies impacting the health of Canadians so that citizens can spend more years living productively and happily.

1.3 Problem Definition

Advances in natural language processing and machine learning today allow for novel methods of collecting health data and analyzing it. Social media platforms represent a rich pool of information where people publicly share content. Platforms such as Twitter offer real-time and geotagged data collection. The constant stream of data about user's daily lives can greatly enhance and improve traditional public health surveillance techniques.

This thesis explores the challenges of building systems that collect high-volumes of data from Twitter for public health surveillance. We investigate the necessary steps for the collection and preparation of data for building a robust machine learning classifier to identify tweets related to physical activity, sedentary behaviour and sleep. Once relevant tweets are obtained, our interests lie in discovering insightful trends across different regions and time periods in Canada.

1.3.1 Research Questions

The main focus of this thesis is to extract insightful trends from social media data in relation to PASS. Another important component involves the use of machine learning algorithms for the collection of a labeled dataset and the classification of tweets. Developing a system that accomplishes such tasks requires answering the following research questions:

1. What do Canadians share about PASS on social media?
2. Can social media be used to measure PASS health indicators?
3. How are machine learning classifiers affected by the period of the year data is collected?
4. What machine learning classifiers perform best for identifying PASS-related tweets?

1.4 Thesis Contributions

In this thesis, we present a system that collects Twitter data for public health monitoring. The main contributions of this thesis are threefold:

1. **Twitter data collection and annotation**: We labelled a total of 10,283 tweets that are related to the PASS domain through manual annotation, keyword filtering and an active learning technique. This method helps increase the performance of our machine learning (ML) algorithm by finding relevant tweets to label.
2. **Machine learning models for the classification of tweets**: We built machine learning classifiers to determine whether a tweet was related to physical activity,

sedentary behaviour and/or sleep. Our models included an ensemble model to classify relevant tweets. The ensemble model consists of various machine learning algorithms that were combined for better performance. We also build deep learning models to obtain the highest performance for the classification of tweets in the PASS domain.

3. **PASS Health Indicators**: We derived health indicators for every province/territory in Canada by predicting PASS labels on our Twitter dataset of 8.4 million tweets. The health metric was obtained by aggregating the counts for every region in Canada and computing the proportion of labelled tweets. We compared our results with Canadian health surveys using Pearson's correlation.

1.5 Outline

The remainder of the thesis is structured as follows:

In **Chapter 2**, we provide relevant background information for public health surveillance systems, and shows some common approaches used for the problem of text classification.

In **Chapter 3**, we describe the techniques used for the preparation of our data. In particular, we discuss methods for the collection of Twitter data for PASS indicators, and the preprocessing and the cleaning of the tweets.

In **Chapter 4**, we present the various machine learning algorithms used for the classification of PASS-related tweets. We provide the details for the ensemble model that makes predictions based on shallow base learners. We also show the model architectures of our deep learning models for the same problem.

In **Chapter 5**, we provide the results from various experiments, including benchmark results for the three PASS indicators for every model used. We also run experiments to show the effect of seasonality on our models.

Finally, in **Chapter 6**, we discuss the findings, limitations and implications of our study. We discuss potential areas for further exploration and improvement.

Chapter 2

Background

2.1 Social Data and Public Health Monitoring

In this work, we leverage the massive, freely, and publicly available data coming from social media, commonly referred to as *social data*. Social data refers to data that is created by people with the goal of sharing with others [44]. In general, social data is generated by a large percentage of the Canadian population, approximately 22 million users, which makes it appropriate for public health monitoring [53]. In this section, we first present traditional approaches for public health monitoring and then explore alternative sources of health data.

2.1.1 Current Approach to Public Health Monitoring

Public health information is crucial for policy makers and health care professionals to develop health care programs and anticipate health care services [4]. Public health organi-

zations like the PHAC rely on Statistics Canada through the Canadian Community Health Survey (CCHS). The CCHS collects health information from Canadians through surveys that are performed biennially. The survey is designed to produce cross-sectional estimates to address priority health data gaps at national, provincial and regional levels [25].

2.1.2 Social Media Platforms

Social media platforms has recently been a popular alternative source to explore public health data and applications. The Google Flu Trends system [14] famously used Internet search activity to provide estimates of influenza prevalence. Using Google's search engine, they inferred the public's interest from their searches for flu-related information. Moreover, blogs, such as Tumblr and Wordpress, are online platforms where users post messages and articles intended to be broadcast to a general, public audience. Blogs are common for extracting social data and social monitoring, as people tend to share their beliefs and the time of day during which they partake in various activities.

Finally, social media platforms like Facebook, Twitter, Instagram or Youtube have been particularly popular for public health monitoring, as they provide publicly accessible data from billions of daily users [52]. The data that can be publicly found in these platforms can reveal population attitudes and behaviors. Previous work assessed alcohol behavioural stages from tweets [29], revealed dietary choices through photos [18], and captured drug use in videos [34]. Since the work in this thesis focuses on population-level monitoring, social media networks are an appropriate source of data. However, it is worth noting that young adults are over-represented on Twitter [50].

2.1.3 Data-Driven Public Health Applications

Early work in social monitoring for health-related applications investigated what public health information could be learned from Twitter by building a binary classifier to identify health-related messages from a corpus of tweets [42]. Their studies led to the construction of structured disease information from tweets for public health metrics.

Recently, Nguyen et al. [37, 38] built a US neighborhood dataset from Twitter data for indicators of happiness, diet, and physical activity. The aim was to obtain area-level indicators of well-being and health behaviours from geo-tagged tweets. They evaluated their method by comparing predicted values for their health indicators to those generated by human labelers. Additionally, geotagged tweets were spatially mapped to census data which allowed them to assess the associations between tweets and their neighborhood variables with demographic, economic and health characteristics [38].

Other similar work attempted to predict general population health behavior and derive mental health indices from Twitter data [39, 40]. They approached their study as a regression problem where they predicted the health index for each US state. For instance, their health index for sleep represents the proportion of people who sleep less than 7 hours daily. They evaluated their results with correlations between actual health rankings from the United States (U.S.)' annually conducted Behavioral Risk Factor Surveillance System (BRFSS) surveys and the estimated health ranking from their models.

2.1.4 PASS Indicators

More specific research on social monitoring focuses on the PASS domain in relation to behavioural medicine. Researchers study how people make choices about their health and how it affects their personal health and well-being.

Researchers have increasingly studied exercise and physical activities using Twitter [27, 61]. Vickey et al. [57] and Kiciman et al. [27] compared self-reported estimates of daily physical activity data provided from surveys to a mobile fitness app where users share their physical activity over Twitter. These studies found that participants' actual physical activity levels were lower than those reported in the self-reported surveys. Akbari et al. [2] built a classifier to detect tweets that mention actions related to wellness, including exercise, diet, and healthcare utilization [44]. Similarly, Dos Reis et al. [19] were interested in whether social media data could be used to make causal inferences for health surveillance. They trained a text classifier to estimate the volume of a user's tweets that express anxiety and depression, then compared it to groups that exercise regularly and a matched control group.

The public monitoring of sleep has been studied less extensively. Nonetheless, research has gone into examining whether social media could be used to infer sleep issues and investigating common sleep problems and patterns [1, 32].

2.2 Text Classification

Text classification is the task of assigning one or more categories to a text document based on its content. In our work, we classify tweets in order to filter for relevant data for analysis. Figure 2.1 shows a standard pipeline for the analysis of social media for public health surveillance, with data being the most important component. The pipeline is made of two parts: filtering and inference. For the filtering task in the pipeline, we consider *machine learning classifiers*. A classifier is an algorithm that assigns labels to the input tweet messages. In the next subsections, we formally define the problem of text classification and provide a context within our system. We also cover different techniques for the data

collection of relevant PASS data. We cover various techniques considered for filtering relevant data, such as keyword filtering, supervised, semi-supervised, and unsupervised machine learning algorithms.

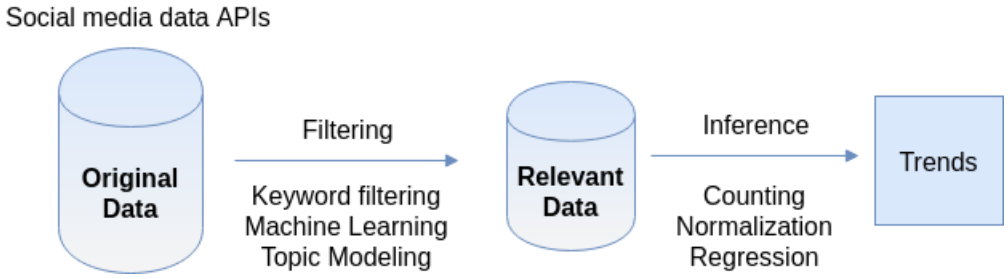

Figure 2.1: Standard pipeline for monitoring of social data

2.2.1 Multi-label Classification

In traditional single-label classification, the task is defined as the concept of learning from of a set of examples that are associated with a single label l from a set of disjoint labels L, $|L| > 1$. If $|L| = 2$, then the learning problem is referred to as a binary classification, whereas if $|L| > 2$, then it would be called a multi-class classification problem [55].

In text classification, text documents can belong to more than one category. In this case, a tweet can be classified into one or more of the three indicators: physical activity, sedentary behaviour or sleep. Multi-label classification assigns each input sample to a set of binary target labels. Multi-label classification is a generalization of the multi-class classification, where in the multi-label problem, there exists no constraint on how many of the classes the input sample can be assigned to.

The existing methods for multi-label classification can be grouped into two separate categories: a) *problem transformation methods* and b) *algorithm adaptation methods*. Problem

transformation methods are the most common approach, where a multi-label classification task is transformed into a one or more single-label classification problem. Algorithm adaptation methods involve extending specific learning algorithms to handle multi-label data directly.

Binary Relevance

The simplest and most commonly used problem transformation method is called binary relevance. This method treats each label as a separate single classification problems, making the assumption that there is no correlation among the various labels in the problem. In other words, $|L|$ datasets are constructed from the original one that contain all examples of the original dataset, labeled as y if the labels of the original example contained y and as $\neg y$ otherwise. More formally, this technique learns $|L|$ binary classifiers $H_l : X \to \{l, \neg l\}$, one for each different label l in L.

Classifier Chains

In this technique, a classifier is first trained on the input data. Subsequent classifiers are then trained on the same input data, in addition to all previous classifiers' predictions in the chain [49]. This allows the model to learn from the signals from the correlation among preceding target variables.

2.2.2 Filtering Techniques

Any quantitative methods of analyzing social data requires the collection and processing of relevant data. Below, we describe numerous ways for identifying, filtering and classifying

pertinent data. We first cover a simple method, namely keyword filtering, and move to more sophisticated machine learning techniques.

Keyword Filtering

Keyword filtering searches for tweets that match a certain pattern or keyword for the identification of relevant information. Keyword filtering is usually the simplest and quickest way for collecting relevant data. In past social monitoring studies, Paul et al. [43, 42] used a list of 20,000 keyphrases related to illnesses, diseases, symptoms, and treatments to collect data for labelling on Amazon Mechanical Turk (AMT). They ensured the quality of their data by using a majority vote technique i.e. removing example points where the majority of annotators did not agree and were unsure of the best label. Similarly, McIver et al. [32] used a list of keywords from researchers with expertise in sleep-related fields of study and experimental queries to obtain their dataset of tweets for characterizing sleep issues. In certain cases, it is also possible to collect data using only hashtag keywords. Akbar et al. used sleep logs on Twitter from the #Sleep_as_Android mobile phone application [1]. Hashtags allow messages to be directly categorized by different topics. Even though keyword filtering is popular, the method has many disadvantages. Keyword filtering only looks at particular individual keywords and fails to obtain the full context of a document.

Supervised ML through SVMs and Logistic Regression

When context matters in text classification, the problem of classifying relevant tweets can be approached as a supervised machine learning problem, where the model is trained on many labeled examples. Support Vector Machine (SVM)s and logistic regression are the most commonly used classification models for identifying health-related tweets [29, 42, 59].

Classifiers rely on a set of predictors, or features, to represent information and signals in text messages. The *bag-of-words* method is a typical approach in natural language processing to represent the features of a sentence as a bag of words while keeping word count and ignoring order. An extension of the bag-of-words approach is to take order into considering and look beyond individual words in a tweet. Considering bigrams, a sequence of two words, trigrams, a sequence of three words, or n-grams of contiguous words adds context. Both bag-of-words approach and its extension treat all words with the same importance. A more powerful technique is to extract Term Frequency-Inverse Document Frequency (TF-IDF) [48] features which assigns a weight to different words depending on their uniqueness and importance rather than giving equal weight to all words.

In certain cases, extracting features from metadata can also be useful. For characterizing sleep issues, McIver et al. [32] used the number of tweets, friends count, followers count, tweet time and location as added features in their model.

Unsupervised ML through Clustering

An alternate approach to the problem of identifying relevant tweets is through a clustering technique known as topic modeling.

Similar to classification, this approach organizes tweets into categories without requiring labels, making it an unsupervised machine learning problem. The main idea behind topic modeling is to view text documents as a composition of many underlying topics, with each topic represented as a cluster of related words. Latent Dirichlet Allocation (LDA) is the most commonly used topic model and has been applied to social media through different applications. For predicting U.S. county-level indices, Nguyen et al. [39] extracted low-level features such as the topics and linguistics for tweets, in addition to statistical features.

They used two different textual features: the Linguistic Inquiry and Word Count (LIWC) package [46] for extracting language style of tweets and latent topics extracted using the LDA method.

Semi-supervised ML through active Learning

Active learning is a type of semi-supervised machine learning in which the learning algorithm interactively query for a user or teacher to label new data points. Liu et al. [29] employed this approach to find alcohol-related tweets efficiently to reduce the manual effort for labelling tweets by identifying tweets that were more useful to annotate. Their method iteratively requested the labelling of tweets that were close to the decision boundary of their supper vector machine classifier, which was initially trained on available labeled data.

Deep Neural Networks

Deep learning models have been used for numerous Natural Language Processing (NLP) applications and achieved success in many traditional NLP tasks [12, 62]. In text classification, these methods use a multilayer neural network (NN) architecture. The model takes an input sentence and automatically learns features from the input through training via backpropagation [13]. The input to these neural networks are raw words represented as a vector of indices taken from a finite dictionary of words. These word indices are then mapped to a feature vector through a look-up table to obtain *word embeddings*. Word embeddings are vector representation of words that are trained from very large unlabeled corpora containing billions of words [33].

While a 2-dimensional Convolutional Neural Networks (CNN) in computer vision involves sliding a window of filters over an image, there is a similar notion in NLP, where

a 1-dimensional CNN is a sliding window over a sequence of words. In other words, a 1-dimension convolution of size k can be thought of as an n-gram detector that learns to identify relevant k-grams in the input [24]. Kim et al. [28] trained a CNN with one layer of convolution on top of Word to vector (Word2vec), word embeddings trained on 100 billion words of Google News, and achieved excellent results on sentence classification tasks.

Similarly, text data from tweets can also be represented as temporal data, or a series of words. A Recurrent Neural Network (RNN) is a type of neural network that processes sequential data. In this case, tweets can be modeled as a sequence of words, where each word is encoded with an integer representing the word index in a dictionary. A Long Short Term Memory (LSTM) unit is a special type of RNN that is designed to overcome certain limitations, such as the vanishing and exploding gradient problems, when dealing with larger sequences and quantities of data. Modeling RNNs for text classification is similar to the deep learning architecture previously described. The input is a fixed-sized input vector of word indices that are mapped to word embeddings, and the model outputs labels of a specified length. For many NLP tasks, context from previous and future words are important and beneficial to a model. To capture the full context of a sentence, a bi-directional LSTM is employed, which presents a forward and backward sequence as input to a network such that it captures past and future information [22].

Chapter 3

Data Preparation

Twitter contains data about the daily lives of millions of Canadian users [52], thus making it an appropriate data source for a study of PASS health indicators. We first collected a large dataset of tweets, then built a labelled dataset for the three PASS indicators using keyword filtering to select relevant tweets for manual annotation. In this chapter, we describe the techniques used for collecting, labelling and cleaning our data before it is used in a machine learning classifier.

3.1 Data Aggregation

The Twitter Streaming API service provides a constant 1% random sample of all tweets in real-time. Developers and researchers can collect targeted datasets based on specific keywords, locations, or users. Twitter's developer platform offers several tools and APIs for extracting data from their social media, hence incoming samples were constrained to geo-tagged tweets from Canada using a bounding box, represented in latitude and longitude

measurements. Using the Twitter Streaming API, we stored incoming English language tweets from Canada to a MongoDB database from November 17, 2017 to May 24, 2018.

3.1.1 Exclusions

Since the Twitter platform is open to anyone, posts can sometimes originate from organizations, clubs, governments, public figures, etc. We excluded all re-tweets since these messages do not originate from the user posting the original tweet. In addition, all non-English tweets were removed to simplify our model learning the linguistic structures of a single language.

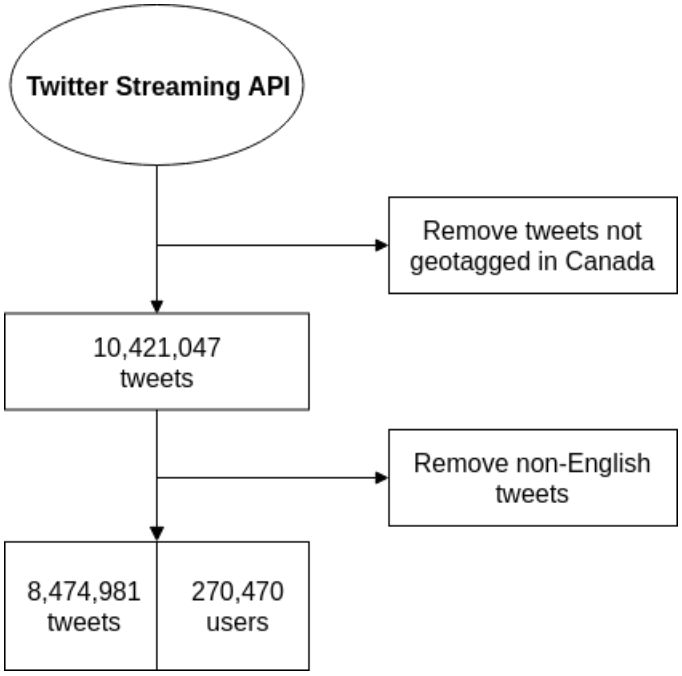

Figure 3.1: The number of tweets and users throughout the filtering stages

3.1.2 Keyword filtering

In order to quickly obtain relevant tweets for manual annotation, we used regular expressions to identify tweets that contained specific keywords appropriate for PASS. Tables 3.1, 3.2 and 3.3 show the lists of keywords used to identify relevant tweets to build the dataset. For sleep, we used the provided keywords from previous work that studied sleep issues using Twitter [32]. The search terms, identified through consultation with researchers with expertise in sleep-related fields of study, included names of sleep disorders, specific medications for sleep problems and popularly used hashtags related to sleep issues. Our list of keywords for physical activity were based on the ones used in similar work [37]. The list included activities gathered from physical activity questionnaires, compendium of physical activities and popular available fitness programs. Popular mountains and parks covering most provinces in Canada were also included. While the keywords do not consist of an exhaustive list of all provincial and national parks, most mountains and parks in Canada are intended to be captured from active learning. For sedentary behaviour, we focused on obtaining tweets that clearly indicated that the user was in a stationary state for a prolonged period. We first captured media that a person usually spends watching on television or online: keywords contained names from the most popular sports leagues, steams, players, movies and TV shows. Additionally, keywords for the most popular books were included.

Although this approach does not include the entire set of all PASS relevant tweets, it is hypothesized that it covers a good majority.

Table 3.1: Sedentary behaviour keywords

Sports athletes	Sports teams	Generic sports	Sports leagues	Movies	TV shows	Books	Verbs
stephen curry	toronto raptors	mma	ufc	black panther	stranger things	book	watch
lebron james	toronto maple leafs	basketball	nba	avengers	netflix	read	read
james harden	edmonton oilers	baseball	mlb	infinity wars	game of thrones	reading	reading
demar derozan	calgary flames	hockey	nhl	dunkirk	strangerthings		watch
kyle lowry	ottawa senators	soccer	mls	annihilation	arrested development		watching
	montreal canadiens	football	nfl	star wars	13 reasons why		
	toronto blue jays			ready player one	mindhunter		
	Toronto FC			red sparrow	house of cards		
				isle of dogs	walking dead		
				jurassic world	black mirror		
				handsmaid's tale	luke cage		

3.2 Data Labelling

The construction of our labelled dataset consists of two main parts: manual annotation from the keyword filtering, and active learning after training an initial classifier. Due to cost restrictions and to ensure consistency across labels, the tweets were annotated by a single person i.e., the main author of this thesis. Familiarity with Canadian cultural and geographical references was also an important factor, as mentions of popular shows, sports, athletes and locations was commonly found in tweets.

3.2.1 Annotation

The main idea behind the keyword filtering was to find tweets that were relevant to the PASS indicators and to narrow down the search for potential tweets to be labelled quickly. We also avoid tagging tweets that are clearly irrelevant to the PASS domain.

We separate the tweets from the keyword filtering into three groups each representing a health indicator independently. The manual annotation task had to determine whether

Table 3.2: Physical activity keywords

Activities	**Outdoors**
gym	hiking
workout	climb
basketball	mountain
baseball	mt
climb	algonquin
dance	mont tremblant
squat	bruce peninsula
skating	manitoulin island
weights	tobermory
gains	national park
volleyball	provincial park
yoga	bromont
park	chicopee
mountain	blue mountain
play	whistler
train	jasper
run	banff
hike	sutton
ski	orford
lift	

Table 3.3: Sleep keywords

Generic	**Medications**
bed	ambien
sack	melatonin
insomnia	zolpidem
dodo	lunesta
zzz	intermezzo
siesta	trazadone
tired	eszopiclone
nosleep	zaleplon
cantsleep	
rest	
zzz	
pass out	
get up	
wake up	
asleep	
slept	
power nap	

a tweet was related to the health indicator or not (positive or negative). Table 3.4 shows samples tweets and how they were annotated to build our PASS dataset.

Table 3.4: Sample tweet annotations

Tweet text	Indicator		
	Physical activity	Sedentary behaviour	Sleep
4 hours of sleep last night & im ready for bed like 2 hours ago	×	×	✓
i have to be up in 6 hours for a twelve hour shift and i can't sleep	×	×	✓
got up early to go mountainbiking before work and have been rewarded with sunshine	✓	×	×
i did yoga today!	✓	×	×
geralds game - stephen king. i am reading it over christmas	×	✓	×
i've watched avengersinfinitywar trailer about 7 times now. it gets better with each viewing	×	✓	×

Since the expression of how a person is involved in a physical activity, sedentary behaviour or poor sleep can vary in so many different ways, we followed some guidelines described below to determine when a tweet should be positively labeled. In general, we are interested in tweets that show that the user was involved with one of the PASS indicators recently, either in the past or upcoming week.

Physical Activity

A tweet should be positively labeled for this indicator if the post has sufficient information that the user was involved in some form of physical activity. These tweets incorporated gym-related exercises (e.g., weight lifting, working out), sports (e.g., basketball, soccer), recreation (e.g., hiking, skiing), and light activities (e.g., shoveling snow, walking the dog). We also identified some positive examples that were more difficult to label below.

Location: When a location is mentioned in the tweets that heavily implies that the user did physical activity, then that post was positively labeled.

- Sunday morning @ [YMCA/yoga club/centre]
- missed yesterday but putting in some time today! (@ goodlife fitness in oakville)

- good to be flying again @ batawa ski hill

Sedentary Behaviour

A tweet labeled for this indicator must show that the user was in a sedentary behaviour, which is defined as in a state of sitting or lying down for long periods of times. A person is usually sedentary at work, at school, at home, when travelling or during leisure time. These tweets comprised of watching a screen (e.g., sports games, movies, TV shows) or reading (e.g., book). Examples of more challenging tweets to identify are shown below.

Commentary while watching television: People will often tweet while watching something on TV or online. There must be sufficient information to suggest that the user was in a sedentary behaviour at the moment the tweet was made. Generic statements about a show, movie or sporting event do not qualify.

Sleep

Tweets that demonstrate that the user has poor sleep should be positively labeled for this health indicator. If the post expressed how many hours of sleep the person had, then the general rule to follow is that 7 hours or less represents inadequate sleep [40].

3.3 Data Preprocessing

3.3.1 Metadata

Twitter data comes in the form of text, but can contain attachments such as images and URLs. The data content is accompanied with metadata, such as the timestamps and

location which are used for the analysis in order to understand variation in populations. Location information is often provided by the social media platform. For instance, Twitter provides detailed information in the form of latitude and longitude coordinates when users participate with a GPS-enabled device. Otherwise, the platform can sometimes infer the location of the user through their network and tag a tweet with a nearby city or location name. Finally, Twitter messages are restricted to 280 characters to ensure that users only share short messages. Tweets were parsed to obtain the tweet ID, timestamp, location, and self-stated location.

3.3.2 Data Cleaning

All non-English tweets were excluded in the dataset, although tweets that use non-English number characters or notations were still accepted. We used a tokenizer [1] based on nltk [2] that was adapted for Twitter. The tokenization step allows us to appropriately distinguish different parts of a tweet such as hashtags, usernames, URLs, retweets, emoticons and emojis. The cleaning of the text data from tweets involved removing usernames, emoticons, emojis, and URLs from the content in preparation for feature extraction. Finally, the tweets were all lower-cased, special characters were stripped and generic stop words were also removed.

Because our experiments include the use of deep learning models such as RNNs and CNNs, we treat tweets as temporal data, or a sequence of words. After tokenization, we convert the raw text data to a vector of integers, representing word indices from the dictionary of all words in the dataset. The vectors are then left-padded with zeros in order to have a fixed-sized input to the deep learning models.

[1] https://github.com/erikavaris/tokenizer

[2] https://www.nltk.org/

Chapter 4

Methods

4.1 Feature Extraction

We used a set of various different features for the PASS classification of tweets that are summarized in Table 4.1.

TF-IDF. We extracted (1, 4)-gram TF-IDF features from the words and characters of a tweet. The most relevant terms from the TF-IDF features are displayed in Figure 4.1.

Hand-crafted features. Additionally, we extracted indirect text features from the tweets such as the number of unique words used, the character length of the tweet, the number of words per tweet, and the average word length of the tweet. We show the tweet data histograms for the average word length, word length, and character length in Figure 4.2, 4.4, and 4.3 respectively.

Word embeddings. We also used GloVe [1] word embeddings [47] weighted with TF-

[1] https://nlp.stanford.edu/projects/glove/

Table 4.1: Features used in shallow machine learning algorithms

Feature	**Dimensions**
Average word length	1
Number of words	1
Number of characters	1
Number of unique words	1
(1, 4)-gram word TF-IDF	5000
(1, 4)-gram character TF-IDF	5000
Glove word vectors	200
Total	10204

IDF. We used GloVe word vectors pre-trained on 2 billion tweets, which contained a vocabulary of 1.2 million words with dimension 200.

4.2 Classifier Details

We experiment with two different types of classifiers: an ensemble model consisting of several machine learning models, and deep learning classifiers.

4.2.1 Ensemble Model

We employ an ensemble of machine learning algorithms that combines base learners into one predictive model. The classification is accomplished by merging the predictions of 5 separate models: Naive Bayes (NB), Logistic Regression (LR), RBF-SVM, Random Forest (RF), and XGBoost (XGB). A meta learner then learns how to best combine the predictions

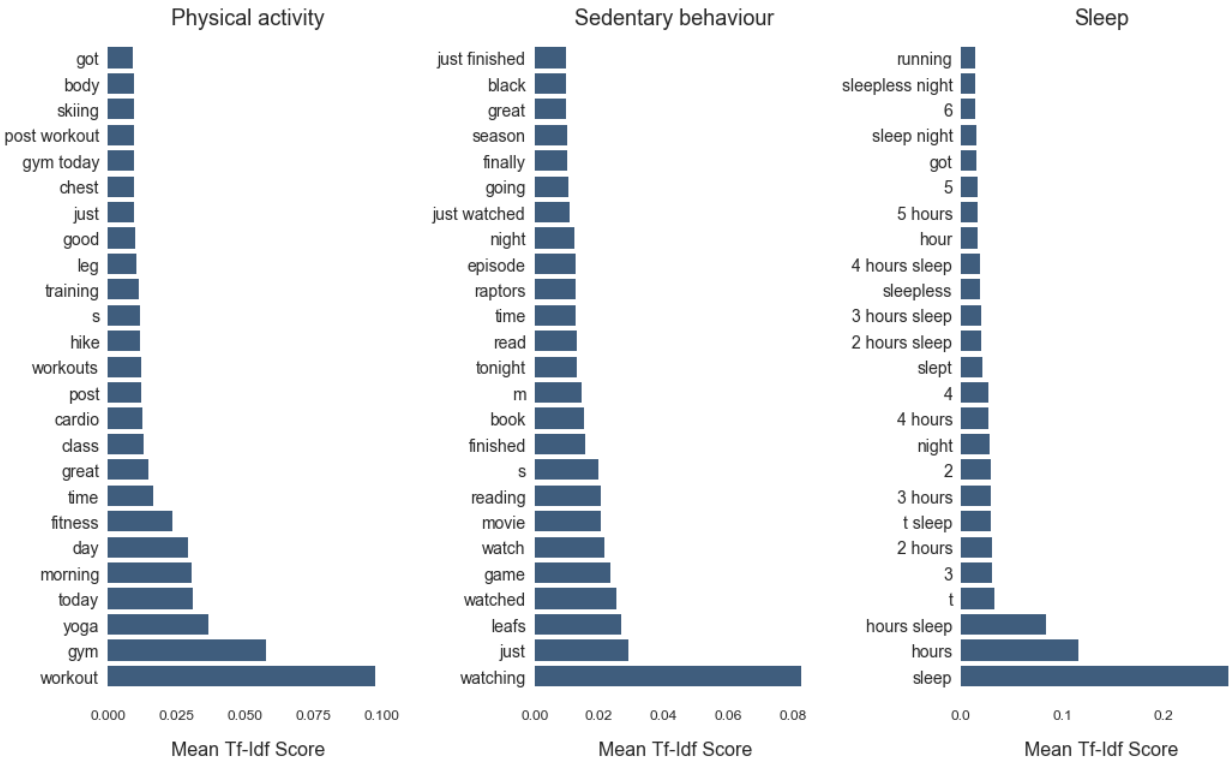

Figure 4.1: Top (1-4) grams from the TF-IDF feature extraction

of the base learners into a final output prediction. In our case, we feed the probability output of these models to another logistic regression model for the final prediction of the PASS labels. In doing so, our model can learn to weigh more accurate models more heavily in the final prediction. This approach allows to combine powerful non-linear classifiers for a greater representational capacity. Our model implementations use the scikit-learn machine learning package in Python [45]. Table 4.2 shows the algorithms and matching Python libraries used for the implementation.

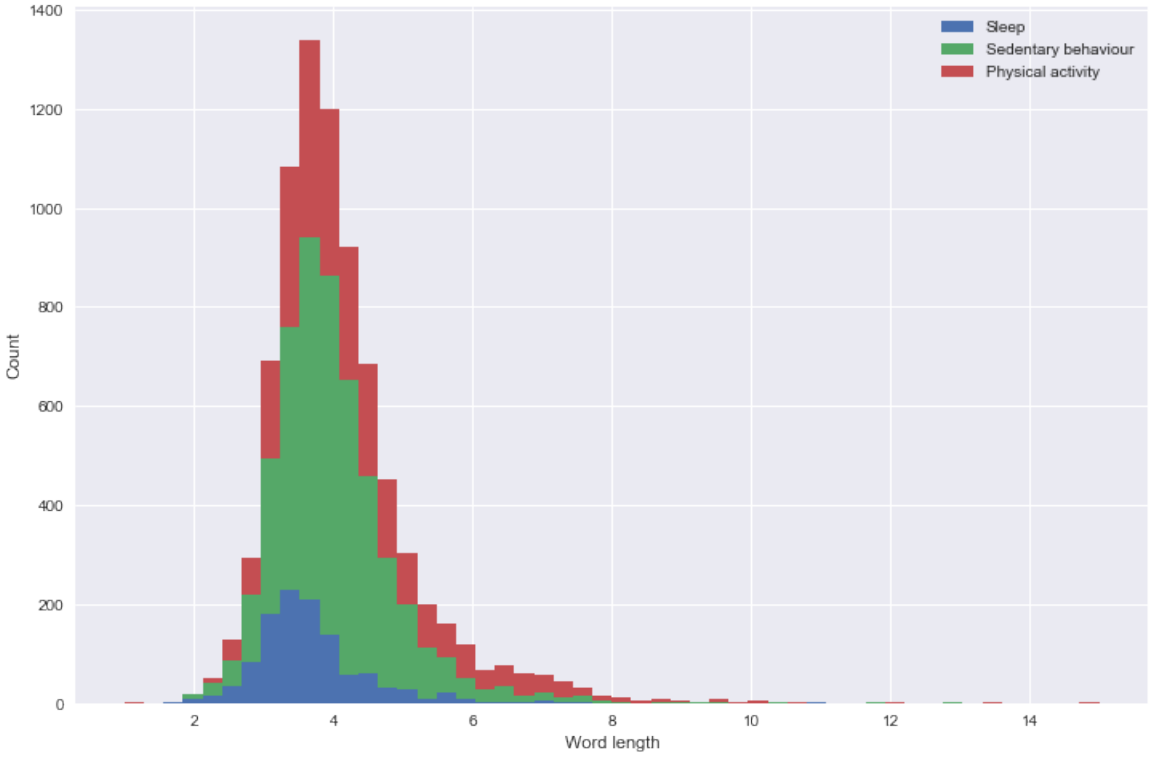

Figure 4.2: Average word length frequency across tweets

4.2.2 Deep Learning

The implementation of our deep learning models is done using the Keras package that offers a high-level and fast deep learning framework [9].

1-D Convolutional Neural Network

In this architecture, we use a one-dimensional CNN with pre-trained GloVe embeddings. Our model uses 1-D CNNs with 3 different filter sizes that are concatenated. This involves sliding a one-dimensional window of three different lengths (size of 1, 2 and 3 in this case). The three convolutional layers can be seen as detectors searching for n-grams of size 1, 2

Figure 4.3: Character frequency across tweets

and 3. The subsequent layers include a max-pooling layers, several fully-connected layers with ReLu activation layers [36]. We employ dropout on the fully-connected layers to prevent overfitting the model [51]. Figure 4.6 shows the model architecture along with the input and output sizes at each layer.

Recurrent Neural Network

The RNN architecture consists of a bidirectional LSTM, which involves stacking two layers: the first layer is the original input kept as is, while the second layer is a reversed copy of the input sequence. This technique adds more context to the model, as each timestep

Figure 4.4: Number of word frequency across tweets

in the network processes data in both directions at the same time. Figure 4.7 shows the architecture of the neural network. Similar to the 1-D CNN architecture previously described, we use pre-trained GloVe word embeddings.

4.3 Active Learning

We use an active learning approach in order to increase the efficiency of labeling our tweets, since we have the challenge of class imbalance. This approach involves labelling data iteratively to avoid tagging redundant tweets and focus on data that are more difficult

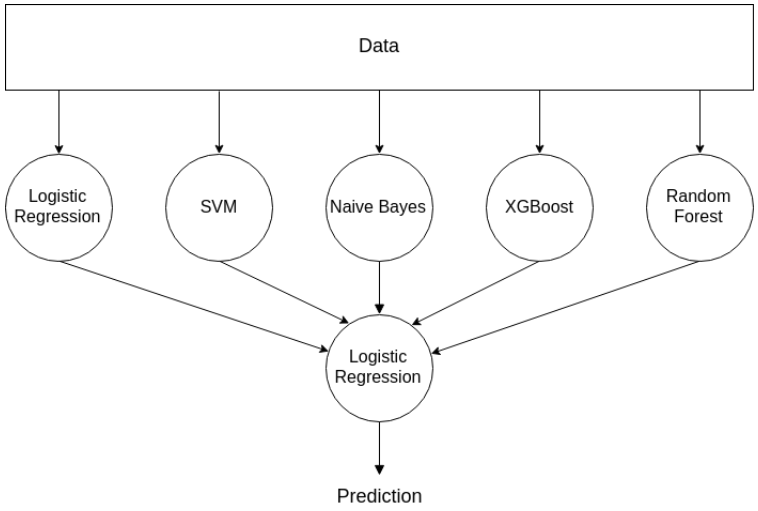

Figure 4.5: Ensemble model high-level overview

to label to increase the performance of our model, similar to Liu et al. [29]. We accomplish this by first training a classifier on the existing labelled data of roughly 1000 tweets that were manually labelled. We then selected all unlabelled tweets that have a probability estimate within a threshold $0.30 <= P(l_i|x) <= 0.70$. The tweets that fall into this decision boundary range are then selected as future tweets to be labelled in the process. The process was repeated 5 times, with approximately 2000 tweets being labelled at each iteration by the same person i.e., the main author of the thesis. The manual annotation rules followed were detailed previously.

4.4 PASS Health Indicators

To obtain PASS health indicators, we classified tweets on our entire dataset of 8.4 million tweets. Afterwards, we aggregated the tweet count for each indicator per province and territory in Canada, and computed the relative proportion for that region. In order to com-

Table 4.2: Algorithms used in the ensemble model with corresponding Python library

Algorithm	**Python libraries**
Random Forest	sklearn.ensemble.RandomForestClassifier
Logistic Regression	sklearn.linear_model.LogisticRegression
RBF SVM	sklearn.svm.SVC
Linear SVM	sklearn.svm.LinearSVC
Naive Bayes	sklearn.naive_bayes.BernoulliNB
XGBoost	xgboost.XGBClassifier

pare our PASS-derived health indicators, we obtained the 2014 CCHS data from Statistics Canada. We correlated our PASS indicators and the 2014 CCHS prevalence of self-reported obesity, diabetes, and mood disorder for each province using Pearson's correlation coefficient. The survey questions used to obtain these statistics can be seen in Table 4.3. The questions emphasized that the survey was interested in conditions diagnosed by a health professional and that are expected to last or have already lasted 6 months or more. The assessment of obesity used guidelines based on Body Mass Index (BMI), a measure that examines the relation between weight and height. BMI was calculated for the population aged 12 to 17 years old, then aged 18 and older, excluding pregnant women, and persons less than 3 feet (0.914 metres) tall or greater than 6 feet 11 inches (2.108 metres) [6]. For the survey variables, we took the percentages of people who reported being overweight or obese, the percentage of self-reported diabetes and mood disorders for every province.

Table 4.3: Survey questions for mood disorders and diabetes

Variable	**Survey question**
Mood disorder	Do you have a mood disorder such as depression, bipolar disorder, mania or dysthymia?
Diabetes	Do you have diabetes?

4.5 Evaluation

For the evaluation of our models, we used 5-fold cross-validation, where 20% of the data is held-out for testing, and this process is repeated K times (5 in this case), with the final scoring metrics averaged across each fold. Due to class imbalance, precision, recall, F1-Score and the Area Under the Receiver Operating Characteristic (AUROC) score were used as evaluation metrics. The training times, predict times, and accuracy were also recorded for additional performance measures. All experiments are performed on a Linux system with an i7-6700 (8-cores @ 3.40 GHz), and one NVidia Titan X Pascal GPU.

Precision is the fraction of True Positives (T_p) over the number of true positives plus the number of false positives (F_p). Precision can be interpreted as the classifier's ability to not label positively a sample that is negative.

$$P = \frac{T_p}{T_p + F_p} \tag{4.1}$$

Recall is the ratio of True Positives (T_p) over the number of true positives plus false negatives (F_n). Intuitively, recall measures a classifiers' ability to find all positive samples in its data.

$$R = \frac{T_p}{T_p + F_n} \tag{4.2}$$

The **F1 score** is the harmonic mean of both the precision and recall, which combines both metric into a single value.

$$F1 = 2\frac{P \times R}{P + R} \tag{4.3}$$

4.5.1 Seasonality Experiments

A set of experiments to investigate the seasonality of the data was also run. Both tweets and PASS-related indicators should vary depending on the period of the year. Indeed, we were interested in the differences between winter and spring/summer periods and how that was reflected in the classification of the models. For this experiment, we separated the data into two time periods: the first part (November 17, 2017 to March 20, 2018) was used as training data, while the other part (April 18, 2018 to May 24, 2018) was used as the testing set.

4.6 Hyper-parameter Tuning

The hyper-parameters for the classifiers were determined through an exhaustive grid-search. The best parameters were selected by performing a nested 5-fold cross-validation and selecting the values that obtained the highest performance scores after being averaged out across folds. To avoid using the cross-validation test data for model selection and to prevent overfitting, hyper-parameters are tuned based on a validation set consisting of data from the inner cross-validation loop, while the outer cross-validation loop is used to estimate the model performance. For logistic regression and the support vector machines, we tuned the C and regularization parameters across a uniform distribution of $(10^{-4}, 10^{4})$.

For the random forest and XGBoost classifiers, we tuned the minimum number of samples required to split an internal node and the maximum depth of the tree over the range [100, 200, 300, 400], [1, 5, 10, 15], and [6, 7, 8, 9, 10] respectively.

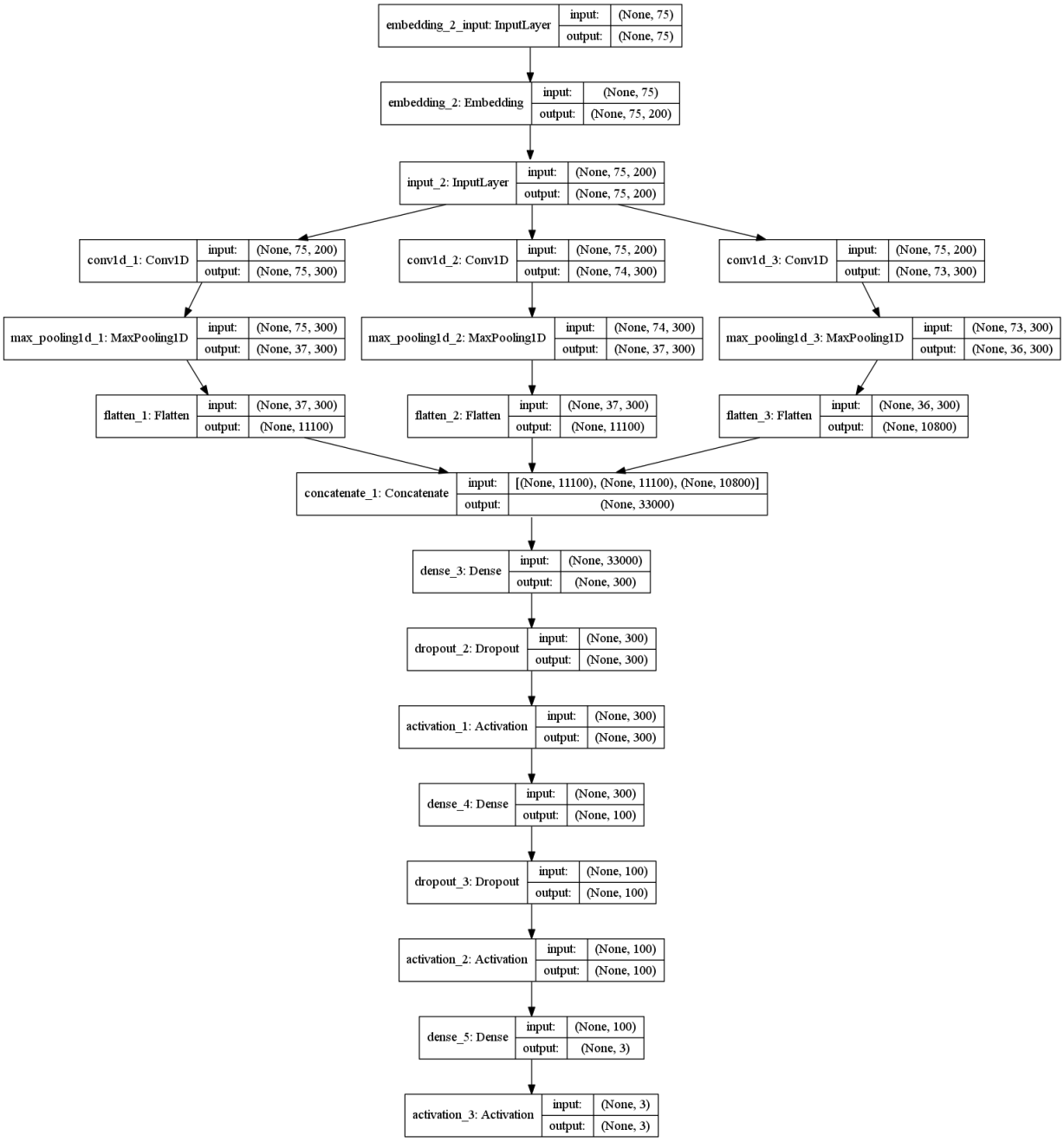

Figure 4.6: 1-D CNN Architecture. Each block contains the name of the layer, and the input and output dimensions. The three convolution layers that are concatenated have filter sizes of 1, 2 and 3.

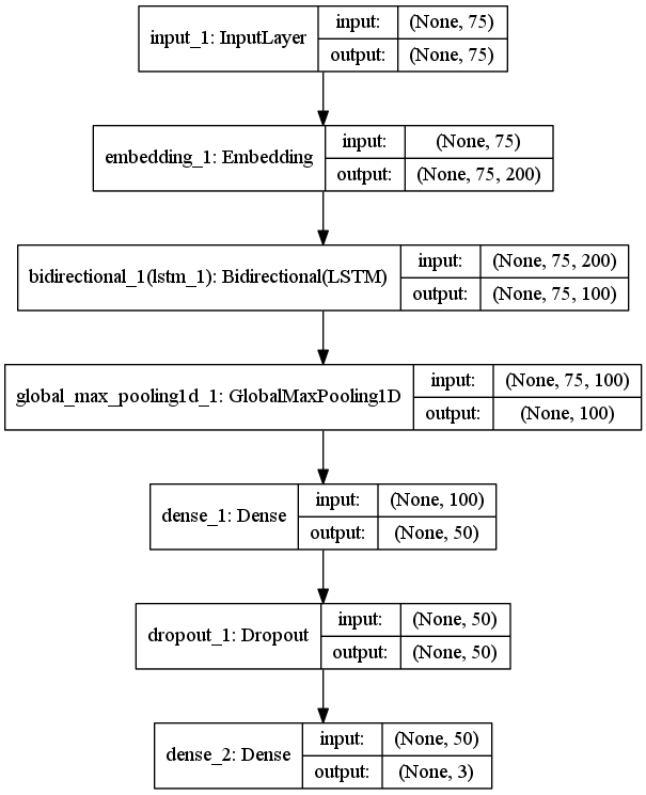

Figure 4.7: LSTM Architecture. Similar to the CNN, each block contains the name of the layer, and the dimensions of the input and output. A dimension size of None indicates that the network accepts inputs of any dimension for the batch sizes.

Chapter 5

Results

5.1 Data

After the manual annotation of our tweet dataset through keyword filtering and active learning, we had a total of 10,283 labelled tweets; 3245, 5657, and 1381 were annotated for physical activity, sedentary behaviour and sleep tweets respectively. The tweets were continuously sampled from the Twitter Streaming API from November 15, 2017 to May 24, 2018, with a gap from March 20, 2018 to April 18, 2018. The input data to the classifiers was the output from the preprocessed tweets described in the previous chapter. We use the binary relevance problem transformation approach to deal with the multi-label classification; a separate classifier was trained for each PASS indicator independently. Table 5.1 shows the distribution of the labeled tweets from our dataset.

Table 5.1: Distribution of labeled data

	Positive tweets	**Negative tweets**
Physical Activity	1524 (46.7%)	1721
Sedentary Behaviour	1945 (34.8%)	3712
Sleep	541 (39.2%)	840

5.2 Model Comparisons

We performed experiments with individual classifiers, then combined them together for the ensemble model. We included the results from the deep learning models for the complete classifier benchmark for the task predicting tweets related to PASS. Table 5.2 shows the experimental results for all the different models we used.

The Naive Bayes classifier obtains the best recall score across each PASS domain, with the highest recall of 0.914, 0.764, and 0.939 for physical activity, sedentary behaviour and sleep. Due to the simplicity of naive bayes, it is a low variance and high bias model, where the model does not achieve high capacity or complexity, but will get most of the predictions right. In addition, naive bayes is a fastest model to train and predict.

For higher complexity, the ensemble model and deep learning models obtain the best performance. The LSTM model scores the highest F1-Score for physical activity, and sedentary behaviour with scores of 0.810 and 0.665 respectively, while falling short as the second highest F1-score for sleep (0.679) behind naive bayes (0.741).

For our deep learning models, we found that training for 10 epochs with batch sizes of 256 gave the best performance. Both deep learning models performed well in comparison to the shallow machine learning algorithms, obtaining strong performance scores in across

most metrics. These results may be seen alongside those of the other algorithms in Table 5.2. While the LSTM model achieves slightly higher performance than the CNN, it takes considerably more time to train the model. This could be explained by the fact that the LSTM processes temporal data, and can only backpropagate through the data sequentially, while operations in the CNN model can be parallelized more efficiently. One point to consider about the CNN model is that it has high space complexity, with 13,720,201 parameters compared to 3,534,301 for the LSTM. Due to the higher capacity of the CNN model, it suffers from overfitting when trained for many epochs and fails to generalize as well as the LSTM. The addition of L2 regularization on the fully-connected layers of the CNN did not seem to help prevent overfitting [51].

5.3 PASS Health Indicators

We used the LSTM model to classify tweets on our entire dataset, as it obtained the best results across all evaluation metrics for the PASS indicators. The total count after aggregating tweet predictions by province and territory is shown in Table 5.3. Applying the LSTM classifier to the 8.4 million tweets resulted in 135,052 PASS-related tweets.

5.3.1 Comparison with Health Surveys

We compared our PASS health indicators with existing Canadian health surveys. Table 5.4 shows the pairwise Pearson's correlation coefficient between the PASS health indicators and the 2014 CCHS data for prevalence of obesity, diabetes and mood disorders. The physical activity indicator was positively correlated with the prevalence of obesity, diabetes and mood disorder. However, there were no statistically relevant ($^*p < .05$) correlations

Table 5.2: Model comparisons using 5-fold cross-validation

Domain	**Algorithm**	**Train time (s)**	**Predict time (s)**	**Test set**				
				Accuracy	F1 Score	Precision	Recall	ROC AUC
Physical activity	Logistic Regression	9.80	3.29	0.912	0.734	0.803	0.720	0.954
	Naive Bayes	**8.19**	3.35	0.917	0.812	0.766	**0.916**	0.954
	Random Forest	10.35	4.13	0.911	0.713	0.821	0.668	0.964
	XGBoost	23.21	4.33	0.926	0.790	0.832	0.799	0.975
	Ensemble	39.51	5.17	0.924	0.775	**0.833**	0.773	0.973
	rbf-SVM	79.11	18.84	0.927	0.793	0.827	0.801	0.964
	LSTM	42.94	2.00	**0.937**	**0.833**	0.811	0.890	**0.979**
	1d-CNN	17.66	**1.44**	0.915	0.767	0.772	0.815	0.956
Sedentary behaviour	Logistic Regression	10.42	3.36	0.828	0.487	0.629	0.421	0.851
	Naive Bayes	**8.41**	3.40	0.771	0.576	0.488	**0.769**	0.859
	Random Forest	10.67	4.12	0.856	0.529	0.742	0.439	0.904
	XGBoost	22.71	4.27	0.863	0.593	0.722	0.527	0.914
	Ensemble	43.32	5.05	0.864	0.575	**0.751**	0.487	0.916
	rbf-SVM	136.23	32.35	0.850	0.562	0.681	0.506	0.883
	LSTM	49.06	3.15	**0.881**	**0.658**	0.739	0.604	**0.924**
	1d-CNN	23.60	**2.67**	0.833	0.601	0.583	0.634	0.857
Sleep	Logistic Regression	10.72	**3.35**	0.952	0.562	0.767	0.521	0.976
	Naive Bayes	**8.23**	3.37	0.934	**0.740**	0.701	**0.937**	0.968
	Random Forest	10.21	4.16	0.941	0.507	0.787	0.498	0.987
	XGBoost	18.01	4.28	0.947	0.629	0.791	0.641	0.982
	Ensemble	29.97	5.07	0.947	0.624	0.798	0.632	**0.988**
	rbf-SVM	43.41	11.04	0.953	0.633	0.769	0.635	0.983
	LSTM	55.88	4.35	**0.954**	0.683	**0.797**	0.711	0.987
	1d-CNN	29.47	3.71	0.951	0.602	0.796	0.596	0.971

Table 5.3: Predicted tweet counts and relative proportion by province

Province	Tweet count				Proportion of tweets (%)		
	Physical activity	Sedentary behaviour	Sleep	Total	Physical activity	Sedentary behaviour	Sleep
Northwest Territories	100	73	4	6,694	1.494	1.091	0.060
Prince Edward Island	274	385	11	28,516	0.961	1.350	0.039
New Brunswick	750	1,458	52	103,699	0.723	1.406	0.050
British Columbia	7,313	10,527	532	1,081,369	0.676	0.973	0.049
Yukon	79	75	1	12,621	0.626	0.594	0.008
Newfoundland and Labrador	546	1,280	52	104,998	0.520	1.219	0.050
Saskatchewan	946	2,172	111	192,221	0.492	1.130	0.058
Nova Scotia	1,420	3,477	175	307,495	0.462	1.131	0.057
Alberta	5,225	12,652	606	1,190,521	0.439	1.063	0.051
Ontario	18,195	50,890	2,227	4,282,422	0.425	1.188	0.052
Québec	2,014	5,144	284	489,935	0.411	1.050	0.058
Nunavut	32	105	4	8,266	0.387	1.270	0.048
Manitoba	1,303	4,321	237	372,593	0.350	1.160	0.064

between the PASS health indicators and the survey data.

5.3.2 Trends

We analyzed the performance of our health indicators by observing the temporal trend of the proportion of all tweets for the PASS domains, as depicted in Figure 5.1 and 5.2. The most noticeable trend we can observe is the lower proportion of tweets related to physical activity after April 2018. This could perhaps be explained by arrival of the new spring season. As most of the data that the classifier was trained on were during the winter, it was biased towards physical activities in the winter e.g., skiing, snowboarding and other similar sports.

Table 5.4: Correlations, means and standard deviations of PASS indicators and 2014 Canadian Community Health Survey variables across all provinces. The survey measures included the percentage of people self-reporting as being obese, having diabetes and mood disorders. These variables were correlated with each PASS domain, shown as their relative proportion of all tweets. The mean and standard deviation for all those variables were reported.

Measure	**Mean (std)**	**Correlations**					
		1.	2.	3.	4.	5.	6.
1. Physical activity tweets (%)	0.61 (0.31)						
2. Sedentary behaviour tweets (%)	1.13 (0.20)	0.033					
3. Sleep tweets (%)	0.05 (0.01)	-0.058	0.575*				
4. Obesity, 18 years old and over (%)	58.09 (6.14)	0.382	0.298	0.117			
5. Obesity, 12 to 17 years old (%)	24.79 (10.57)	0.339	-0.204	-0.120	0.709*		
6. Diabetes (%)	6.60 (2.34)	0.317	-0.035	-0.131	0.670*	0.766*	
7. Mood disorder (%)	7.73 (2.77)	0.236	-0.093	-0.085	0.444	0.480	0.767*

Notes: $^{*}p < .05$

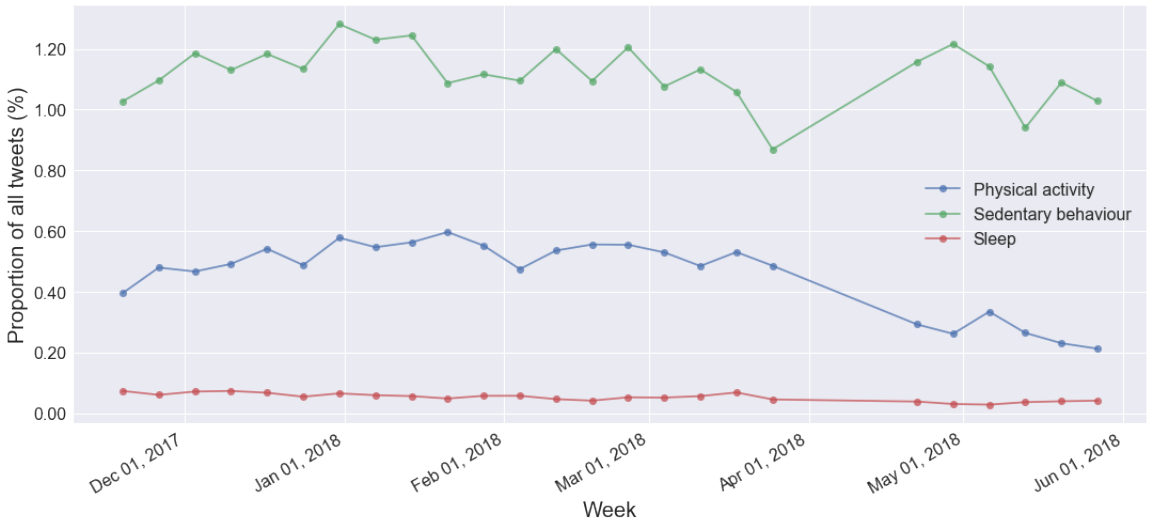

Figure 5.1: Weekly trend of Twitter PASS indicators. Labels are aligned to Sunday.

5.4 Seasonality

We further investigated the effect of seasonality on the data by separating it into 2 different time periods: from November 15, 2017 to March 20, 2018 (Winter), and April 18, 2018 to May 24, 2018 (Spring). The labelled tweets in the first period included a total of 8,070 samples and were used for the training data. The second time frame, which consisted of 2,213 tweets, served as the testing data. Figure 5.3 shows the tweet count from each week since the beginning of data collection. As shown in Table 5.5, seasonality seems to have an effect on the performance of the classifiers. For all three PASS indicators, lower scores than those presented in Table 5.2 were obtained on all performance metrics. The 1-dimensional CNN scored the highest F1-scores on all three domains when trained on tweets within a specific time window. Our models failed to generalize as well to newer data, which contained tweets with information related to the new spring season. As an example, for sedentary behaviour, new movies that are watched and mentioned in later

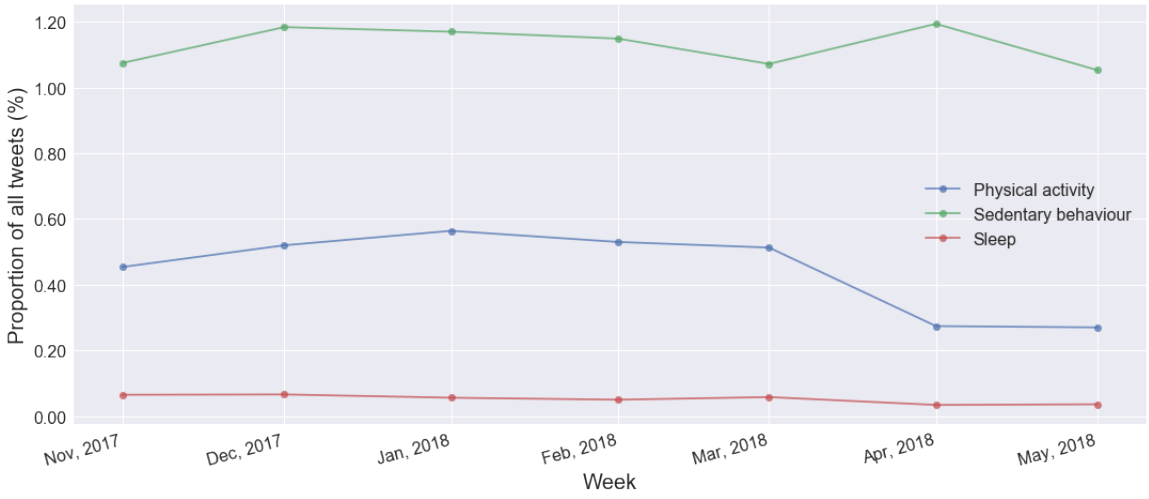

Figure 5.2: Monthly trend of Twitter PASS indicators. Labels are aligned to the first day of the month.

months would fail to be part of the training data. While the LSTM model previously obtained the highest scores on the test sets, it appears to be more susceptible to overfitting when trained on tweets from the winter season, as suggested by the low recall scores.

5.5 Error Analysis

From the experimental results, it can be seen that sedentary behaviour is the most difficult domain to classify, with the lowest scores obtained when compared to the physical activity and sleep classes. The poorer performance for the sedentary behaviour domain could be explained by the tweets in that class, which are much more difficult to classify. In particular, there are a lot of different activities that can fall into the sedentary behaviour category: watching a movie or the television, reading a book, etc. Since the expression of those activities on social media involves mentioning a title, our features may not capture

Table 5.5: Model comparisons using seasonal difference test split

Domain	**Algorithm**	**Train time (s)**	**Predict time (s)**	**Test set**				
				Accuracy	F1 Score	Precision	Recall	ROC AUC
Physical activity	Logistic regression	9.93	3.81	0.940	0.750	0.753	0.747	0.972
	Naive Bayes	**9.07**	3.66	0.927	0.741	0.646	0.868	0.947
	Random forest	10.60	4.41	0.934	0.714	0.743	0.687	0.965
	XGBoost	23.58	5.16	0.943	0.773	0.739	0.811	**0.977**
	Ensemble	39.61	5.92	0.947	0.783	0.770	0.796	**0.977**
	rbf-SVM	79.59	19.64	0.936	0.750	0.708	0.796	0.969
	LSTM	43.67	2.18	0.915	0.730	0.588	**0.962**	0.971
	1d CNN	17.77	**1.78**	**0.948**	**0.779**	**0.801**	0.758	0.976
Sedentary behaviour	Logistic regression	11.30	3.88	0.830	0.457	0.560	0.385	0.834
	Naive Bayes	**8.55**	3.86	0.735	0.491	0.381	**0.690**	0.808
	Random forest	11.58	4.37	0.853	0.444	**0.739**	0.317	0.878
	XGBoost	23.42	4.67	0.860	0.553	0.680	0.466	0.891
	Ensemble	43.50	5.41	**0.863**	0.536	0.720	0.427	**0.886**
	rbf-SVM	136.47	32.91	0.856	0.536	0.667	0.449	0.873
	LSTM	49.52	3.71	0.850	**0.578**	0.605	0.554	0.875
	1d CNN	23.75	**3.00**	0.835	**0.578**	0.549	0.610	0.872
Sleep	Logistic regression	10.94	**3.46**	0.955	0.531	0.691	0.431	0.974
	Naive Bayes	**8.85**	4.17	0.947	0.640	0.533	**0.800**	0.963
	Random forest	10.52	5.14	0.958	0.521	0.806	0.385	0.924
	XGBoost	18.04	5.17	**0.966**	0.638	**0.838**	0.515	0.960
	Ensemble	30.18	5.90	0.964	0.612	0.829	0.485	**0.978**
	rbf-SVM	44.29	11.78	0.960	0.589	0.750	0.485	**0.978**
	LSTM	56.32	4.53	0.955	0.497	0.731	0.377	0.980
	1d CNN	29.98	4.26	0.961	**0.669**	0.662	0.677	0.973

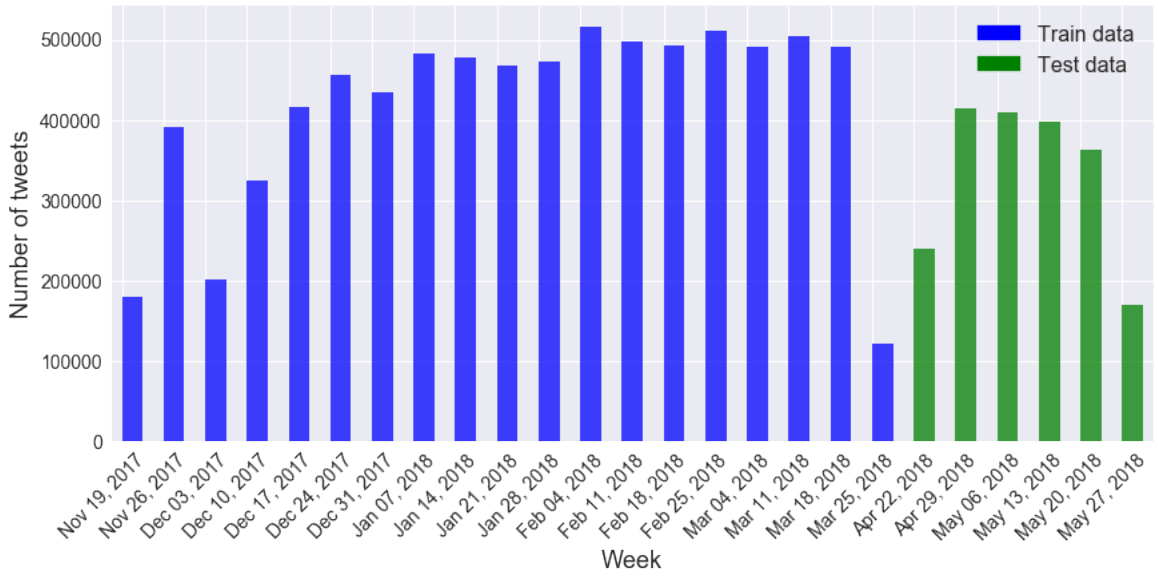

Figure 5.3: Tweet count by week

all possible movies, books, or TV shows, which may explain where our model fails.

The classifiers performed significantly better for predicting physical activity and sleep tweets. From investigating 4.1, the n-grams that are important for physical activity explicitly mentions activity names or sports, which makes it easy for models to capture. Similarly, the most significant n-grams for sleep were all terms that indicated low or poor quality sleep mentioning the amount of time slept, e.g. the sleep duration.

To better understand how our model makes decisions, we analyzed Type I and Type II errors, as well as correct predictions made.

Physical activity. Table 5.6 shows the different types of errors our models made. The false positive example shows a way how the model can fail. As seen from that sample tweet, important keywords like *jumping* and *parkour* were used in the sentence. Since our models rely heavily on word counts (TF-IDF) and word embeddings, the models can

Table 5.6: Physical activity classified tweets

	Physical activity
True positives	today was day 4 out of 90. my workout went great!
False positives	i be jumping from an emotion to another like i'm doing parkour
True negatives	pick an emoji to describe yesterday's workout
False negatives	The worst thing about hiking is when your phone dies before you get to the top and you can't cop any pics for your 'gram!!

mistake this tweet for one related to physical activity when relevant words are used. To properly classify such a tweet requires an understanding of the context, and, for the most part, this is not something our models are designed to capture. TF-IDF features only look at (1, 4)-grams in a tweet which can be insufficient when context is provided further on in a sentence. The use of an LSTM may help circumvent this limitation but generally requires a larger number of examples to perform well. When observing the false negative example, we can see the difficulty that can arise when building models from human text. In this particular case, the user made a statement that implies that the person went hiking shortly before the tweet was posted. While the *hiking* keyword was used, the rest of the tweet must have failed to have enough important words for the models to make the correct prediction or the other words (e.g., phone, pics, gram) confused it.

Sedentary behaviour. The false positive example in Table 5.7 again highlights the difficulty of the task. There are informative keywords (watching, 13 reasons why) that mislead the model to believe that the sample should be positively labelled. However, when looking at the entire context of the tweet, there lacks information to suggest that the user was indeed watching a show recently for an extended period of time.

Table 5.7: Sedentary behaviour classified tweets

	Sedentary behaviour
True positives	watching avengers tonight. so excited!
False positives	this is exactly why i stopped watching 13 reasons why
True negatives	75 % of leaf fans don't truly know the game of hockey
False negatives	guess im going to watch avengers alone

Table 5.8: Sleep classified tweets

	Sleep
True positives	im running on 4 hours of sleep, i have fours hours left in my day
False positives	too bad i can' t sleep until i get home from work in 7 hours
True negatives	it's so nice to sleep for 10 hours
False negatives	almost 2 AM... still awake... need to be up in 4 hours...

Sleep. Errors from our sleep classifiers are reported in Table 5.8. Similar to the model failings of the previous health domains, the false positive could be explained by the use of important keywords, without fully grasping the entire context of the tweet. When investigating the TF-IDF features from Figure 4.1, the amount of hours were important terms in the features. However, as suggested by the figure, the terms do not form a comprehensive coverage of all quantities of sleep that are less than or equal to 7 hours because of the low amount of data. In other words, if the training data does not include the mention of x hours of sleep, then the model will likely fail to classify that tweet properly.

Chapter 6

Discussion and Future Work

6.1 Contributions

We developed a system that uses Twitter data to obtain population-level health metrics in Canada for physical activity, sedentary behaviour, and sleep. The three main contributions of this thesis are as follows:

1. **Twitter data collection and annotation**: We manually annotated our own dataset of tweets using keyword filtering, and an active learning approach to iteratively obtain data more efficiently. Our labeled dataset consisted of a total of 10,283 tweets that are related to the PASS domain.

2. **Machine learning models for the classification of tweets**: We treated this problem as a multi-label classification problem where a classifier has to predict three different PASS indicators as output. We then experimented with different machine learning algorithms. We built an ensemble model consisting of 5 different machine

learning algorithms. The model feeds the probability output from a set of prediction algorithms to a logistic regression classifier to find the optimal combination. We then experimented with deep learning models that included a 1-dimensional CNN and an LSTM.

3. **PASS Health Indicators**: We derived health indicators for every province and territory in Canada by predicting PASS labels on our Twitter dataset of 8.4 million tweets. The health metric was obtained by aggregating the counts for every province and territory in Canada, and computing the proportion of labelled tweets. We then compared our results with the 2014 CCHS survey data using Pearson's correlation.

6.2 Deep Learning for Tweet Classification

To the best of our knowledge, public health surveillance systems that rely on machine learning algorithms for the classification of social media data have not extensively used deep learning. In this thesis, we show the effectiveness of deep learning models with relatively simple architectures. Even with a small tweet dataset, our deep learning models outperformed the traditionally used methods for tweet classification. Once a model is fully trained, deep learning algorithms can further improve the efficiency of predictions as data can be processed in batches and computed in parallel. Finally, there is the benefit of avoiding the manual engineering of features for the classifiers, as the input to the neural networks are simply word embeddings, with the model naturally learning a representation of the data during the training phase. This suggests that the approach is versatile enough for strong performance on other tweet classification problems.

6.3 Comparison with Existing Measures

Current measures from PHAC for PASS indicators provide average statistics such as the average number of minutes per day engaged in physical activity, or the average number of hours per day spent sedentary. These statistics are gathered from individuals using surveys. In comparison, our system measures PASS indicators based on Twitter data, which are aggregated through different regions in Canada. While it is not possible to obtain specific metrics as in the case of PHAC, it allows us to obtain population-level measures that can complement current methods. In particular, passive monitoring of social media data is different such that PASS-related events are observed rather than self-reported. This enables going beyond tracking high level statistics, as Canadians are active at home, at work, and in their communities. Monitoring physical activity, sedentary behaviour, and sleep through social media may be helpful to obtain a clearer picture of the health behaviours of individual Canadians.

Another disadvantage of PHAC's current indicators is that they are updated only every few years. For instance, CCHS is completed annually while the Canadian Health Measures Survey (CHMS) is conducted every two years. The analysis of social media data like Twitter can provide greater temporal resolution and be updated daily, weekly, or monthly. There is also a geographical resolution advantage when using social media data, as this data often comes with latitude and longitude coordinates. While our work computed indicators on a provincial level, it is possible to scale down to even municipalities and neighborhoods. Based on the differences stated above, PHAC can potentially use our indicators to inform their policy decisions with a much shorter latency. Such additional findings can inform revisions for the next versions of the surveys that the existing PASS indicators are based on.

6.4 Implications

Monitoring PASS at the population level complements PHAC. Busy and stressful urban lifestyles that are common in Canada, as well as rising mental health problems such as anxiety and depression, often lead to sleep deprivation and insomnia [26]. Because PASS problems can significantly predispose Canadians to health problems like obesity, diabetes, heart disease, and mental health disorders, it is invaluable for PHAC to have the state-of-the-art ability to monitor the Canadian population's PASS levels. With that said, our PASS surveillance system could provide PHAC with a novel way to continuously monitor the PASS levels in different regions of Canada in a cost-effective manner. Our PASS indicators can complement the existing survey-based indicators by providing higher temporal resolution with weekly updates. Twitter's large userbase enables our indicators to effectively represent the majority of Canadians, although older adults and individuals in remote areas with limited Internet connectivity may be under-represented. This limitation could offer an opportunity; PHAC and Statistics Canada could perhaps target their survey efforts to those under-represented sub-populations rather than conducting surveys of the entire population, which would lead to cost savings. As such, the combination of existing survey efforts and our indicators could together better represent the Canadian population.

6.5 Limitations

While our method of monitoring public health using Twitter data shows promising results, it also comes with limitations.

6.5.1 Data Collection

One limitation in the use of Twitter data for our study is that it is not representative of the entire population, as the majority of users tend to be young male adults [50]. Because our dataset represents a subset of publicly available tweets, our conclusions may not generalize to the full stream of tweets. Obtaining population-level PASS indicators required the collection of geotagged tweets from at most 1% sample of the total tweets at a given time using Twitter's Streaming API. Previous studies suggested that about 1% to 2% of tweets contain GPS location information, and that Twitter's Streaming API may obtain 40% to 90% of all geotagged tweets [5, 35]. Tweets that contain GPS location information may be different from those without. Furthermore, the possibility that privacy concerns differ by age group may pose a problem for interpreting results given a lack of demographic indicators in our current approach. While there are approaches to infer certain demographic statistics from usernames and tweet contents, we did not explore this route. Nonetheless, adoption rates on social media platforms have been steadily increasing [20].

On the subject of privacy concerns, it should be noted that not all users are open to sharing their daily activities and health behaviours on the Internet; hence, the dataset is biased to certain populations. Indeed, some groups are more willing to share health information on public platforms than others [30]. Our technique may only capture certain types of information that users feel comfortable sharing in relation to the PASS domain. Additionally, individuals suffering from mental health or weight-related issues may feel especially uncomfortable disclosing such information due to stigmatization [15]. There is also the possibility that users post socially desirable information that do not reflect their actual lifestyles.

Finally, because our dataset was manually annotated by a single person, there are some

possible errors and biases that can arise, depending on the interpretation of the labeller.

6.5.2 Classifiers

Our dataset was initially collected by using regular expression keyword filtering to select the most relevant tweets to label. While this allows us to obtain most tweets that are related to PASS, there is still a subset that does not go through the filter to be labelled. In part because of the small dataset that we have and the insufficient data for all relevant tweets, there will always be a certain amount of tweets that our classifiers will not be able to detect.

One challenge when dealing with Twitter data is that the text content of tweets often require a lot of cleaning, or comes with problems such as typos, missing characters, and grammatical mistakes. This poses a problem for machine learning algorithms that rely on word frequencies as each token of a tweet is used to build a frequency count for feature extraction. Additionally, the occurrence of rare words, or named entities such as movie titles, book titles, and sports teams or players also becomes a challenging topic. Acquiring a dataset large enough to contain all these named entities is also not a feasible task, hence there should be an approach to deal with rare words or entities, but that share similar meaning. Data augmentation could be used to generate additional training data (e.g., substituting known entities with known rarer ones).

Our method also only considers tweets in the English language, which limits conclusions to English speakers. The verbal expressions that differ across cultures is not taken into account, such as the possible sarcasm and humor that a tweet can contain. While it is still challenging for a human to infer a PASS related tweet, it is even more difficult for a computer to understand the nuances of natural language.

6.6 Future Work

Because our dataset is relatively small, and deep neural networks perform better with large amounts of data, it is very likely that our method of predicting PASS indicators would perform better with additional data. The required scalability of this approach could be accomplished using AMT for the annotation of more tweets, combined with the active learning approach used in this work and others such as Liu et al. [29]. Additionally, GNIP, Twitter's enterprise API, provides access to a substantially higher volume of data from the platform. This allows one to obtain tweets from the past, which goes beyond the regular 1% random sample provided by the streaming real-time API.

Another interesting research route would be to look into other social media outlets that make their data publicly available via an API, such as Instagram and Facebook. Instagram provides an interesting opportunity to apply image recognition technology, which would be an additional modality that can give important insight into the population's actual behaviours and lifestyles. The extension of this work could potentially complement public health data and assist health policymakers in their decision-making, and thus benefit the health of all Canadians.

References

[1] Fatema Akbar and Ingmar Weber. # sleep_as_android: Feasibility of using sleep logs on twitter for sleep studies. In *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*, pages 227–233. IEEE, 2016.

[2] Mohammad Akbari, Xia Hu, Liqiang Nie, and Tat-Seng Chua. From tweets to wellness: Wellness event detection from twitter streams. In *AAAI*, pages 87–93, 2016.

[3] Tim Althoff, Jennifer L Hicks, Abby C King, Scott L Delp, Jure Leskovec, et al. Large-scale physical activity data reveal worldwide activity inequality. *Nature*, 547(7663):336, 2017.

[4] Yves Béland, Lorna Bailie, Gary Catlin, and MP Singh. Cchs and nphs—an improved health survey program at statistics canada. In *Proceedings of the American Statistical Association Meeting, Survey Research Methods*, pages 687–682, 2000.

[5] Scott H Burton, Kesler W Tanner, Christophe G Giraud-Carrier, Joshua H West, and Michael D Barnes. " right time, right place" health communication on twitter: value and accuracy of location information. *Journal of medical Internet research*, 14(6), 2012.

[6] Statistics Canada. Canadian community health survey. *Statistics Canada*, 2014.

[7] Valerie Carson, Mark S Tremblay, Jean-Philippe Chaput, and Sebastien FM Chastin. Associations between sleep duration, sedentary time, physical activity, and health indicators among canadian children and youth using compositional analyses. *Applied Physiology, Nutrition, and Metabolism*, 41(6):S294–S302, 2016.

[8] Maria Chiu, Laura C Maclagan, Jack V Tu, and Baiju R Shah. Temporal trends in cardiovascular disease risk factors among white, south asian, chinese and black groups in ontario, canada, 2001 to 2012: a population-based study. *BMJ open*, 5(8):e007232, 2015.

[9] François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 7(8), 2015.

[10] Rachel C Colley, Valerie Carson, Didier Garriguet, Ian Janssen, Karen C Roberts, and Mark S Tremblay. Physical activity of canadian children and youth, 2007 to 2015. *Health reports*, 28(10):8–16, 2017.

[11] Rachel C Colley, Didier Garriguet, Ian Janssen, Cora L Craig, Janine Clarke, and Mark S Tremblay. Physical activity of canadian adults: accelerometer results from the 2007 to 2009 canadian health measures survey. *Health reports*, 22(1):7, 2011.

[12] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[13] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[14] Samantha Cook, Corrie Conrad, Ashley L Fowlkes, and Matthew H Mohebbi. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. *PloS one*, 6(8):e23610, 2011.

[15] Patrick Corrigan. How stigma interferes with mental health care. *American psychologist*, 59(7):614, 2004.

[16] Janet B Croft. Cdc's public health surveillance of sleep health. *SRS bulletin*, 19(1):15, 2013.

[17] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

[18] Munmun De Choudhury, Sanket Sharma, and Emre Kiciman. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, pages 1157–1170. ACM, 2016.

[19] Virgile Landeiro Dos Reis and Aron Culotta. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 182–188, 2015.

[20] Maev Duggan. Social media update 2014. 2015. http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf.

[21] Bernard MFM Duvivier, Nicolaas C Schaper, Michelle A Bremers, Glenn Van Crombrugge, Paul PCA Menheere, Marleen Kars, and Hans HCM Savelberg. Minimal

intensity physical activity (standing and walking) of longer duration improves insulin action and plasma lipids more than shorter periods of moderate to vigorous exercise (cycling) in sedentary subjects when energy expenditure is comparable. *PloS one*, 8(2):e55542, 2013.

[22] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*, 2015.

[23] Centre for Chronic Disease Prevention and Control. Improving health outcomes : a paradigm shift : Centre for chronic disease prevention strategic plan 2016 – 2019. 2015.

[24] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[25] Ronald Gravel and Yves Béland. The canadian community health survey: mental health and well-being. *The Canadian Journal of Psychiatry*, 50(10):573–579, 2005.

[26] Ellen T Kahn-Greene, Desiree B Killgore, Gary H Kamimori, Thomas J Balkin, and William DS Killgore. The effects of sleep deprivation on symptoms of psychopathology in healthy adults. *Sleep medicine*, 8(3):215–221, 2007.

[27] Emre Kıcıman and Matthew Richardson. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 547–556. ACM, 2015.

[28] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[29] Jason Liu, Elissa R Weitzman, and Rumi Chunara. Assessing behavioral stages from social media data. In *CSCW: proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work*, volume 2017, page 1320. NIH Public Access, 2017.

[30] Yingjie Lu, Yang Wu, Jingfang Liu, Jia Li, and Pengzhu Zhang. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. *Journal of medical Internet research*, 19(4), 2017.

[31] Charles E Matthews, Maria Hagströmer, David M Pober, and Heather R Bowles. Best practices for using physical activity monitors in population-based research. *Medicine and science in sports and exercise*, 44(1 Suppl 1):S68, 2012.

[32] David J McIver, Jared B Hawkins, Rumi Chunara, Arnaub K Chatterjee, Aman Bhandari, Timothy P Fitzgerald, Sachin H Jain, and John S Brownstein. Characterizing sleep issues using twitter. *Journal of medical Internet research*, 17(6), 2015.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[34] Elizabeth M Morgan, Chareen Snelson, and Patt Elison-Bowers. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6):1405–1411, 2010.

[35] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *ICWSM*, 2013.

[36] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[37] Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. Building a national neighborhood dataset from geotagged twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance*, 2(2), 2016.

[38] Quynh C Nguyen, Matt McCullough, Hsien-wen Meng, Debjyoti Paul, Dapeng Li, Suraj Kath, Geoffrey Loomis, Elaine O Nsoesie, Ming Wen, Ken R Smith, et al. Geotagged us tweets as predictors of county-level health outcomes, 2015–2016. *American journal of public health*, 107(11):1776–1782, 2017.

[39] Thin Nguyen, Mark Larsen, Bridianne O'Dea, Hung Nguyen, Duc Thanh Nguyen, John Yearwood, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Using spatiotemporal distribution of geocoded twitter data to predict us county-level health indices. *Future Generation Computer Systems*, 2018.

[40] Thin Nguyen, Duc Thanh Nguyen, Mark E Larsen, Bridianne O'Dea, John Yearwood, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Prediction of population health indices from social media using kernel-based textual and temporal features. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 99–107. International World Wide Web Conferences Steering Committee, 2017.

[41] Public Health Agency of Canada. Physical activity, sedentary behaviour, and sleep (pass): a new way of tracking healthy daily activity. 2017. https://infobase.phac-aspc.gc.ca/datalab/pass-blog-en.html.

[42] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272, 2011.

[43] Michael J Paul and Mark Dredze. A model for mining public health topics from twitter. *Health*, 11:16–6, 2012.

[44] Michael J Paul and Mark Dredze. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183, 2017.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[46] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[47] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[48] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142, 2003.

[49] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 254–269. Springer, 2009.

[50] Aaron Smith and Monica Anderson. Social media use 2018: Demographics and statistics. 2018. http://www.pewinternet.org/2018/03/01/social-media-use-in-2018.

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[52] Statista. Number of social media users worldwide from 2010 to 2021 (in billions). 2018. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[53] Statista. Social networking in canada - statistics & facts. 2018. https://www.statista.com/topics/2729/social-networking-in-canada/.

[54] Mark S Tremblay, Valerie Carson, Jean-Philippe Chaput, Sarah Connor Gorber, Thy Dinh, Mary Duggan, Guy Faulkner, Casey E Gray, Reut Gruber, Katherine Janson, et al. Canadian 24-hour movement guidelines for children and youth: an integration of physical activity, sedentary behaviour, and sleep. *Applied Physiology, Nutrition, and Metabolism*, 41(6):S311–S327, 2016.

[55] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. A review of multi-label classification methods. In *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006)*, pages 99–109. Citeseer, 2006.

[56] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.

[57] Ted Vickey and John G Breslin. Do as i tweet, not as i do: comparing physical activity data between fitness tweets and healthy people 2020. *mHealth*, 1, 2015.

[58] Pamela Wicker and Bernd Frick. The relationship between intensity and duration of physical activity and subjective well-being. *The European Journal of Public Health*, 25(5):868–872, 2015.

[59] Zhijun Yin, Daniel Fabbri, S Trent Rosenbloom, and Bradley Malin. A scalable framework to detect personal health mentions on twitter. *Journal of medical Internet research*, 17(6), 2015.

[60] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[61] Ni Zhang, Shelly Campo, Kathleen F Janz, Petya Eckler, Jingzhen Yang, Linda G Snetselaar, and Alessio Signorini. Electronic word of mouth on twitter about physical activity in the united states: exploratory infodemiology study. *Journal of medical Internet research*, 15(11), 2013.

[62] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.