AN EXAMINATION OF EXERCISE EFFECTS ON SELF-REPORTED WELL BEING USING SOCIAL MEDIA DATA

by
MUTLU SORUKLU

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
July 2023

ABSTRACT

AN EXAMINATION OF EXERCISE EFFECTS ON SELF-REPORTED WELL BEING USING SOCIAL MEDIA DATA

MUTLU SORUKLU

DATA SCIENCE M.S. THESIS, JULY 2023

Thesis Supervisor: Asst. Prof. ONUR VAROL

Keywords: exercise, social media, sentiment analysis, causal inference, time-series, well-being

The increasing usage of social media has opened up new avenues for studying the relationship between physical exercise and well-being. While exercise has long been recognized for its positive impact on physical health, its effects on mental and emotional well-being have gained significant attention. In this context, Twitter, as a popular social media platform, provides a valuable source of real-time data that reflects individuals' thoughts, emotions, and behaviors in their daily lives. Furthermore, the rise of wearable devices has allowed researchers to collect detailed activity reports, including exercise data. By leveraging Twitter data and sentiment analysis, we explored the effects of exercise on daily life well-being. Through the analysis of sentiment scores of tweets posted before, during, and after exercise periods, we compared the average sentiment scores of exercise and non-exercise periods to gain insights into the impact of exercise on overall sentiment. Our methodology involved applying causal inference models to time series data by using propensity score matching methods, revealing how exercise periods influenced people's sentiment status. The results of our study highlight the constructive influence of regular physical activity on mental well-being. We have identified the positive effect of exercise on the daily posted content in terms of sentiment during the exercise periods. This research contributes to the growing body of knowledge on the relationship between exercise and daily life well-being.

ÖZET

EGZERSİZİN SOSYAL MEDYA VERİLERİNİ KULLANARAK KİŞİNİN BİLDİRDİĞİ İYİ OLUŞU ÜZERİNDEKİ ETKİSİNİN İNCELENMESİ

MUTLU SORUKLU

VERİ BİLİMİ YÜKSEK LİSANS TEZİ, TEMMUZ 2023

Tez Danışmanı: Dr. Ögr. Üyesi ONUR VAROL

Anahtar Kelimeler: egzersiz, sosyal medya, duygu analizi, nedensellik analizi, zaman serisi

Sosyal medyanın artan kullanımı, fiziksel egzersiz ve sağlık arasındaki ilişkiyi incelemek için yeni yollar açtı. Egzersizin fiziksel sağlık üzerindeki olumlu etkisi uzun süredir kabul edilirken, zihinsel ve duygusal esenlik üzerindeki etkileri önemli ölçüde dikkat çekmiştir. Bu bağlamda popüler bir sosyal medya platformu olan Twitter, bireylerin günlük yaşamlarındaki düşünce, duygu ve davranışlarını yansıtan gerçek zamanlı değerli bir veri kaynağı sağlamaktadır. Ayrıca, giyilebilir cihazların yükselişi, araştırmacıların egzersiz verileri de dahil olmak üzere ayrıntılı etkinlik raporları toplamasına olanak sağlamıştır. Twitter verilerinden ve duyarlılık analizinden yararlanarak, egzersizin günlük hayattaki refah üzerindeki etkilerini araştırdık. Egzersiz dönemlerinden önce, sırasında ve sonrasında gönderilen tweetlerin duyarlılık puanlarının analizi yoluyla, egzersizin genel duyarlılık üzerindeki etkisine ilişkin içgörüler elde etmek için egzersiz ve egzersiz yapılmayan dönemlerin ortalama duyarlılık puanlarını karşılaştırdık. Metodolojimiz, eğilim puanı eşleştirme yöntemlerini kullanarak zaman serisi verilerine nedensel çıkarım modellerini uygulamayı ve egzersiz sürelerinin insanların duygu durumlarını nasıl etkilediğini ortaya çıkarmayı içeriyordu. Çalışmamızın sonuçları, düzenli fiziksel aktivitenin zihinsel esenlik üzerindeki yapıcı etkisini vurgulamaktadır. Egzersizin günlük yayınlanan içerik üzerindeki olumlu etkisini, egzersiz dönemlerinde paylaşılan içeriklerin duygu durumuna olan pozitif etkisini açığa çıkararak göstermiş olduk. Bu araştırma, literatürdeki egzersiz ve günlük yaşam refahı arasındaki ilişki hakkındaki çalışmalara katkıda bulunmaktadır.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitudes to my thesis supervisor Asst. Prof. Onur Varol. He was so dedicated to my academic growth and was willing to share his invaluable knowledge with me throughout my thesis process. I have learned so much from him which I will bear by my side. Even at times when I fall into darkness in my research, he encouraged me to find the light again and continue. I consider myself fortunate to have had the opportunity to work under his supervision and benefit from his wealth of knowledge and experience. His mentorship has not only contributed to the successful completion of my thesis but has also shaped me into a more confident and capable researcher.

I am forever indebted to my family for their immense support, and I am incredibly fortunate to have such a loving and caring family by my side. Their belief in me and their continuous support have been instrumental in my personal and academic growth. I want to express a special acknowledgment to my beloved mother, Fatma, for her unwavering love, encouragement, and abiding support that she has bestowed upon me without any conditions. Her presence has been a constant source of strength and inspiration.

I would like to extend my heartfelt appreciation to all the members of VRL Lab for their unwavering technical and emotional support throughout my master's journey. Their friendship, encouragement, and constructive feedbacks have been instrumental in making this experience truly meaningful.

To all my beloved family...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Social media usage has been increasing significantly over the years. Regardless of gender, educational level, economic status or ethnicity, average social media usage jumped from 7% to 65% in the period 2005-2015.(Perrin, 2015) Recently, there has been a growing interest in understanding the relationship between physical exercise and well-being(Zhang, Chen & Chen, 2021), (Jacob, Tully, Barnett, Lopez-Sanchez, Butler, Schuch, López-Bueno, McDermott, Firth, Grabovac & others, 2020), (Maugeri, Castrogiovanni, Battaglia, Pippi, D'Agata, Palma, Di Rosa & Musumeci, 2020) and (An, Chen, Wang, Yang, Huang & Fan, 2020). Exercise has long been recognized as a beneficial activity for improving physical health (Fox, 1999), but its impact on mental and emotional well-being has gained significant attention in the scientific community (Nienhuis & Lesser, 2020), (Belcher, Zink, Azad, Campbell, Chakravartti & Herting, 2021). In the era of social media, where individuals openly share their thoughts and experiences, there is a vast amount of data available that can be analyzed to explore the effects of exercise on daily life well-being (Vickey, Ginis, Dabrowski & Breslin, 2013).

Amongst different types of social platforms, Twitter provides insights about experiences, feelings of people. Twitter provides a unique opportunity to observe individuals' real-time thoughts, emotions, and behaviors, making it an invaluable resource for researchers interested in studying various aspects of human life. By leveraging social media data, particularly tweets, we can gain insights into the subjective experiences of individuals in their everyday lives.

In addition, increase in the number of wearable devices allowed researchers to gather activity reports.(Holst, 2021) With the recent developments on Global Positioning System (GPS) technology, providing confidence to studies that track activity. Connection that these devices have with social media apps provides the opportunity to use Twitter as the source of data to investigate the relation between exercise and feelings.

Various researches have been conducted in the health and exercise field, emphasizing

how exercise affects mental and physical well-being.(Althoff, White & Horvitz, 2016; Fox, 1999) These studies focus that physical activity has a positive impact on health using sensor data and search engine query logs. However, investigating the relation between exercise and mental well-being with Twitter data have yet to be explored.

In this research, we investigated how exercise affects daily life well-being using sentiment scores of tweets posted before, during and after exercise periods. Taking into account that sentiment scores of posts on Twitter hint at the well being of a person, we compared average sentiment scores of exercise and non-exercise periods to understand how exercise affected the overall sentiment. Base methodology to analyze the effects of exercises on daily life is to use causal inference models on time series data. Comparing sentiment scores for exercise and non-exercise periods gives clues about how exercise periods affected the sentiment status of people. We aim to show being regularly active has a constructive influence on mental well being.

Research Objectives:

- To collect a substantial data-set of tweets from users who disclose their exercise routines and compare them with respect to exercise period and non-exercise periods.
- To perform sentiment analysis on the collected tweets to assess and compare the emotional states expressed during exercise and non-exercise periods.
- To analyze and interpret the results to determine the impact of exercise on daily posts of users.
- To provide insights and implications for public health interventions and strategies aimed at improving overall mental health and well-being through exercise.

1.1 General Flow of the Thesis.

A general literature review about studies that focuses on the exercises and well-being and causality analysis background will be given in Chapter 2. Chapter 3 describes the dataset, clarifies how it is collected and gives insights from the the data. Besides it demonstrates processing methods applied on the raw data to create the exercise dataset. In Chapter 4, our methodology will be explained in detail and a comprehensive background on causal inference models that were used in this

research will be given then we explain how these inference models applied on our dataset. Later, results of the methods and main findings will be provided in Chapter 5. Finally, conclusions and discussion will be presented in Chapter 6.

2. LITERATURE REVIEW

The relationship between exercise and well-being has been widely explored in both physical and mental health research. Traditionally, studies have relied on self-report measures or controlled experiments to investigate the effects of exercise on well-being outcomes (Sims, Smith, Duffy & Hilton, 1999), (Meevissen, Peters & Alberts, 2011). However, with the advent of social media platforms and the abundance of user-generated content, researchers now have the opportunity to tap into a vast amount of real-time data to gain insights into the subjective experiences of individuals in their daily lives. In our study we deployed a comparative analysis on social media data to assess the exercise-related tweets and its effects on shared content on social profiles on daily routine. On this aspect, at this section we will analyze the work studied that are linked to our work from the social media data leveraging to methods that followed with similar fashion.

2.1 Exercise and Well-Being

Numerous studies have demonstrated the positive impact of exercise on overall well-being. For example, physical activity has been linked to reduced symptoms of depression and anxiety (Stanton & Reaburn, 2014), improved cognitive function (Hillman, Erickson & Kramer, 2008), and enhanced self-esteem (Fox, 1999). These findings have underscored the importance of incorporating exercise into daily routines for promoting mental and emotional well-being. Exercise is to be behave as an antidepressant for patients suffering depression proven by (Schuch, Vancampfort, Richards, Rosenbaum, Ward & Stubbs, 2016). They conducted this research by including patients with Major Depressive Disorder (MDD) and they also showed that publication bias usually underestimates the effect of exercise on depression suffering patients. Another work related to the positive effect of exercise, recommended that

30 minutes physical activity under supervision 3-4 times a week is recommended to improve the health conditions of MDD patients (Nyström, Neely, Hassmén & Carlbring, 2015).

2.2 Social Media and Well-Being

The rise of social media platforms has provided researchers with a new avenue to explore the relationship between exercise and well-being. Social media platforms offer a unique space where individuals openly share their thoughts, emotions, and behaviors, allowing researchers to analyze user-generated content and gain insights into their subjective experiences. Studies have shown that social media can provide valuable data for examining well-being outcomes, including emotional states, social support, and health behaviors. Analysis by using Twitter data around the globe showcased that people exhibits happier moods on weekends (Golder & Macy, 2011). Study on data collected from Facebook messages from a group of volunteers revealed that males use the word "my" higher while talking about their wife and girlfriends compared to women. This study based on word usage comparisons and reports that emotionally unstable persons use the words "sick of" and "depressed" drastically more. (Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, Agrawal, Shah, Kosinski, Stillwell, Seligman & Ungar, 2013). Using Twitter to early detect the depression was studied with a supervised learning technique by using the context of the tweets posted by depressed and non-depressed users and pointed out that depression might be detected prior to diagnosis (Reece, Reagan, Lix, Dodds, Danforth & Langer, 2017).

2.3 Sentiment Analysis and Social Media

Sentiment analysis, a subfield of natural language processing, enables researchers to assess the emotional tone or subjective opinions expressed in text data. This technique has been widely applied in analyzing social media content to understand the sentiments associated with various topics. For instance, researchers have used sentiment analysis to study emotional responses to political events (Tumasjan, Sprenger,

Sandner & Welpe, 2010), brand sentiment (Godes & Mayzlin, 2002). By applying sentiment analysis to social media data, researchers can gain insights into the emotional experiences. An example usage of such data is portrayed on an analysis that Twitter feeds can be used as a complementary tool for classic polling by measuring the public mood (O'Connor, Balasubramanyan, Routledge & Smith, 2010). Even for stock exchange price predictions, sentiment analysis on social media data can give insights. Researchers found out that aggregate public mood conditions calculated from Twitter feeds improved the prediction of closing price of Dow Jones Industrial Average stock market (Bollen, Mao & Zeng, 2011). Similarly, sentiment analysis from Twitter on macro level emotions can have high correlations with the actual surveys taken by the same community indicating that social media can give correlative results with the surveys which might be hard to collect in general (Pellert, Metzler, Matzenberger & Garcia, 2022).

2.4 Causal Inference

Causal inference is a fundamental concept in statistical analysis that seeks to understand the causality between variables. It goes beyond mere correlation to exploit the underlying mechanisms and drivers of observed phenomena. In essence, causal inference aims to answer the question, "What would have happened if a particular cause or intervention had been different?" By employing rigorous methods and drawing on various statistical techniques, researchers in diverse fields, such as social sciences, medicine, and economics, strive to uncover causal relationships and make informed decisions based on reliable evidence. With its emphasis on identifying causal links, causal inference plays a crucial role in shaping policy decisions, optimizing interventions, and advancing our understanding of the world around us. Figure 2.1 underlines the big picture of causal inference studies.

Based on the book "Observation and Experiment: An Introduction to Causal Inference" by (Rosenbaum, 2017), causal inference methods can be defined as a set of statistical and analytical techniques used to draw causal conclusions from observational or experimental data. These methods aim to uncover cause-and-effect relationships and provide insights into the impact of interventions or treatments on outcomes of interest.

The book discusses various causal inference methods, including:

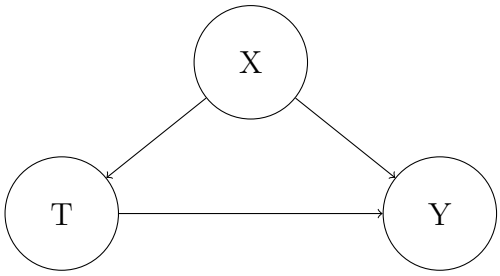


Figure 2.1 Causal impact of treatment (T) to the outcome (Y) under cofounders that define the sample points (X). Aim is to measure the effect of T on Y. Usually the path X $\rightarrow$ T $\rightarrow$ Y corresponds to treatment group while X $\rightarrow$ Y is corresponding to control group behaviour.

$$ATE = \sum_{x}([Y \mid T = 1, X = x] - [Y \mid T = 0, X = x])$$

where ATE is the average treatment effect caused by T (Nogueira et al., 2022)

- **Propensity Score Methods**

  These methods involve estimating the propensity score, which is the conditional probability of receiving a treatment given a set of observed covariates. Propensity score methods allow for the adjustment of confounding variables and help to estimate treatment effects in observational studies.

- **Instrumental Variables**

  Instrumental variables (IV) are used to address endogeneity, where the treatment and outcome variables are simultaneously determined. IV methods utilize instrumental variables that are correlated with the treatment but not directly associated with the outcome, providing a way to estimate causal effects.

- **Matching Methods**

  Matching methods aim to create comparable groups by matching treated and control units based on their observed covariates. This allows for a more balanced comparison between the treated and control groups, mitigating the influence of confounding factors.

- **Sensitivity Analysis**

  Sensitivity analysis assesses the robustness of causal inference results to potential unobserved confounding. It involves examining the impact of varying assumptions or scenarios on the estimated causal effects to evaluate the stability of the conclusions.

The book also covers additional topics such as regression discontinuity designs, difference-in-differences, and causal mediation analysis, which provide further tools for causal inference.

Overall, the causal inference methods discussed in the book provide researchers with a toolkit to overcome the challenges of establishing causal relationships in observational settings. These methods enable the identification of causal effects, control for confounding factors, and enhance our understanding of the causal mechanisms underlying observed phenomena.

In our study, we used propensity score methods and matching methods together to build the treatment and control samples and finally calculate the exercise (treatment) effects on the sentiment score shifts (outcome).

2.5 Comparative Studies on Time Series

Mainly, this study focuses on the effects of exercise on the posts shared on Twitter in terms of sentiment scores. In other terms, we aim to show impact of exercise as an intervention on user's daily life and their post on social world on a context where the time is one of the underlying dimensions. Studies following the similar approach, analyze the effects of an any interventions on status-quo. A study that questions how usage of a mobile game affected users daily steps that they take. Pokemon Go (a mobile game that requires users to go to different places to collect Pokemon) influence on physical activities of users studied and proved that Pokemon Go has a positive impact on average daily steps taken by users after the treatment group started to play the game (Althoff et al., 2016). In addition, in the study of emotional dynamics at the minute level, researchers used time series sentiment data of users before and after the affect labelling event to calculate the influence of affect labelling on the emotional state of users at aggregate level (Fan, Varol, Varamesh, Barron, van de Leemput, Scheffer & Bollen, 2019). Another study about the effect of mindfulness training on psychology students reports that the treatment can promote the therapy quality on patients given by the treated students and well-being of caregivers by improving problem focused coping skills compared to controlled subjects (de Vibe, Solhaug, Rosenvinge, Tyssen, Hanley & Garland, 2018).

2.6 Advantages and Limitations of Social Media Data

The use of social media data in studying exercise and well-being offers several advantages. It provides a large-scale and real-time dataset, allowing researchers to capture the experiences of a diverse range of individuals. Moreover, social media data can overcome recall biases often associated with self-report measures, providing more ecologically valid insights into daily life experiences. However, it is important to acknowledge the limitations of social media data, including issues of representativeness, data quality rooted from high noise in the data, and privacy concerns. Researchers should carefully consider these limitations when interpreting findings. In our study, we created a data pipeline that the results are based on the aggregate level not individual level.

3. DATASET AND DATA DESCRIPTION

3.1 Dataset Description

As mentioned, we used social media data in this project to analyze the effect of exercises in daily life. We can break the data into two category. First category is being the tweets that consists exercise information on them posted on Twitter. These type of tweets will be called exercise tweets in this report. A set of example exercise tweets can be found in the following table.

id	**Tweet Text**
1	I just finished running 4.21 km in 29m:18s with #Endomondo #endorphins
2	I just ran 4.08 km with Nike+. #nikeplus
3	I just finished a 3.95 km run with Nike+ Running. #nikeplus
4	I just finished running 6.35 km in 39m:14s with #Endomondo #endorphins
5	I just ran 10.0 km @ a 5'15"/km pace with Nike+. #nikeplus

Table 3.1 Exercise Tweet Examples

Table 3.1 shows a set of example tweet texts that we considered as exercise tweets. Hashtags that appear on the table was the main source of data collection part. The data collection part will be explained with detail in the upcoming section where we define the data collection pipeline.

Second category in the data is the daily tweets that were posted by users who posted exercise tweets. These tweets were posts that users on the Twitter platform shares in their daily life. These type of tweets in the data will be called daily tweets throughout this report. A set of daily tweets can be found in the below table.

id	Tweet Text
1	Just migrated from LastPass to Bitwarden. So far, so good!
2	And there was I thinking the bird was going to sever her wrist or peck her eye out....
3	Totally! Some sense has returned to the world... (Still more to do, but it's a start!)
4	Does anybody know if car washes (like @IMOcarwash) are open during #Lockdown2?
5	What do you do when you've had a stressful time talking to the Child Maintenance Service? Playstation, beer and cheese.

Table 3.2 Daily Tweet Examples

As seen on the table 3.2, these are the set of posts that users post on Twitter in their daily life.

3.2 Data Collection Pipeline

Here the process of data collection will be explained in detail. While defining the dataset in the previous part we said that we have two category in the whole data. Data collection process started with the first category which is collecting the exercise tweets. Twitter Academic API[1] access endpoints were used in order to collect exercise tweets. All tweets that contain these hashtags were gathered by using historical search.

Data is collected using Twitter's API with the following query:

```
query = "(strava OR fitbit OR mapmyrun OR runkeeper OR nikeplus OR
    garmin  OR endomondo) (run OR ran OR running OR jog OR jogging
    OR workout OR exercise OR fitness OR training OR gym OR cardio)
    (mile OR miles OR km OR  kilometer) lang:en has:links -is:
    retweet"
```

Listing 3.1 Twitter API Search Query

The query has six parts:

- Tweet should have at least one of the words (strava, mapmyrun, runkeeper, nikeplus, garmin, endomondo). These keywords are names of some popular

[1] https://developer.twitter.com/en/products/twitter-api

fitness apps that people use to track their exercises.

- Tweet should have at least one of the words (run, ran, running, jog, jogging, workout, exercise, fitness, training, gym, cardio). These are the words that appear in exercise reports to indicate an exercise made.
- Tweet should have at least one of the words (mile, miles, km, kilometer). This one is to get running or walking exercises.
- Tweet language should be English. We have used English sentiment analyzer to assign score for tweets so we have collected users who post in English language.
- Tweet should have a link inside to identify it was posted from a connected app or linked to it.
- Tweet must not be a retweet.

This search of tweets was used to create a user set $(A) = u_1, u_2, u_3, ..., u_{30000}$ with length 30000. Then we saved collected user ids to gather historical tweets of each individual person in the dataset.

Next step was to collect all tweets of users whose ids are in the database. At this step we used Tweepy[2] package from Python. Process of collecting the daily tweets are summarized in Algorithm 1.

Algorithm 1 Collect Historical Tweets

```
procedure CollectHistoricalTweets(A)
    C ← ∅
    for u_i in A do
        tweets ← GetHistoricalTweets(u_i)
        if tweets is not empty then
            C ← C ∪ {u_i}
    return C
```

Third step was to divide exercise reported tweets from daily tweets for the whole database. To do this, following set of keywords were used. If the tweet text contains any of the word mentioned below, it was labelled as exercise tweet. Otherwise it was considered as daily tweet.

```
exercise_keywords = ['run','ran', 'jog', 'walk', 'hike', 'swim', '
    bike', 'cycle', 'yoga', 'pilates', 'lift', 'gym', 'workout', '
    training', 'marathon', 'fitness', 'fitbit', 'strava', 'nikeplus'
    ,'calories', 'steps', 'activity', 'sweat', 'cardio','endurance',
     'aerobic', 'anaerobic', 'burn', 'energy','exercise', 'health',
```

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

```
'healthy', 'heart', 'muscle', 'strength', 'runners', 'cycling',
'weight', 'training', 'athlete', 'motivation', 'personal record'
, 'fitbit', 'strava', 'endomondo', 'mapmyrun', 'runkeeper']
```

Listing 3.2 Exercise Report Filter Keywords

At the end, we had gathered two set of tweets one being the collection of exercise tweets and the other being the daily tweets. An example table describing what fields are included in the collected sets of tweets. Explanation of the fields can be found in the figure 3.3. User ids and tweets ids are especially used to establish the connection between Exercise and Daily tweet sets. In order to process an exercise tweet or daily tweet we used the text field which the details of this processing will be explained in later parts of the document.

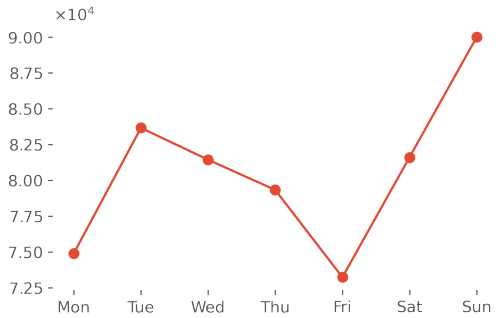
Field Name	Data Type	Explanation
id	string	The unique identifier of the tweet.
created_at	string	The date and time when the tweet was created.
text	string	The content of the tweet.
screen_name	string	The screen name of the user who posted the tweet.
user_id	string	The unique identifier of the user who posted the tweet.
followers_count	integer	The number of followers the user has.
friends_count	integer	The number of friends (users followed by the user) the user has.
retweet_count	integer	The number of retweets the tweet has received.
favorite_count	integer	The number of times the tweet has been favorited.
hashtags	array	An array of hashtags used in the tweet.
mentions	array	An array of users mentioned in the tweet.

Table 3.3 Fields of collected tweets (Eo et al., 2016)

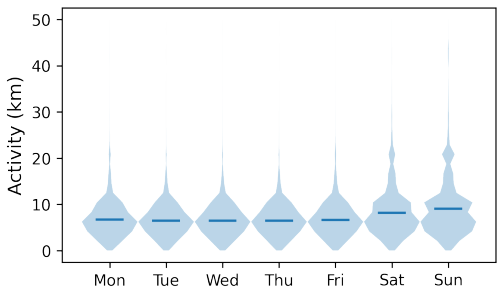
3.3 Statistics on Collected Exercise Dataset

In this section visualizations and basic statistics from the collected exercise dataset will be illustrated. Exercise dataset that we have collected constains many users and each user has many number of exercise reports. Some representations of the dataset will be shown below.

- Daily & Monthly Stats



(a) Count of Exercise Reports



(b) Amount of Exercises

Figure 3.1 Daily Exercise Stats

It can be seen from the figure 3.1 that most of the reported exercises fall on Sundays which makes sense in terms of people' spare time. Usually weekends are the times when people spare on themselves. Additionally amount of exercise in terms of distance is slightly higher towards the weekend indicating that exercises on weekend has higher duration compared to the rest of the week.

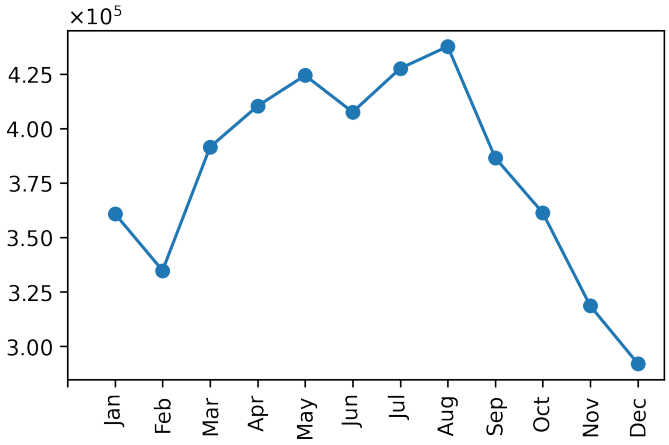


Figure 3.2 Monthly Counts reflects that summer has higher number of reports from users which also indicates that people are going on a walk or running on summer more than other seasons of the year. Starting is taken from 2015 January till the end of 2022 December to eliminate seasonal duplication inconsistency.

- User Time Spans

Figure 3.3 outlines that user time span in the database pile up more than four years signalling that the users mostly been on the Twitter platform for a long time.

- Exercise Report Interevents

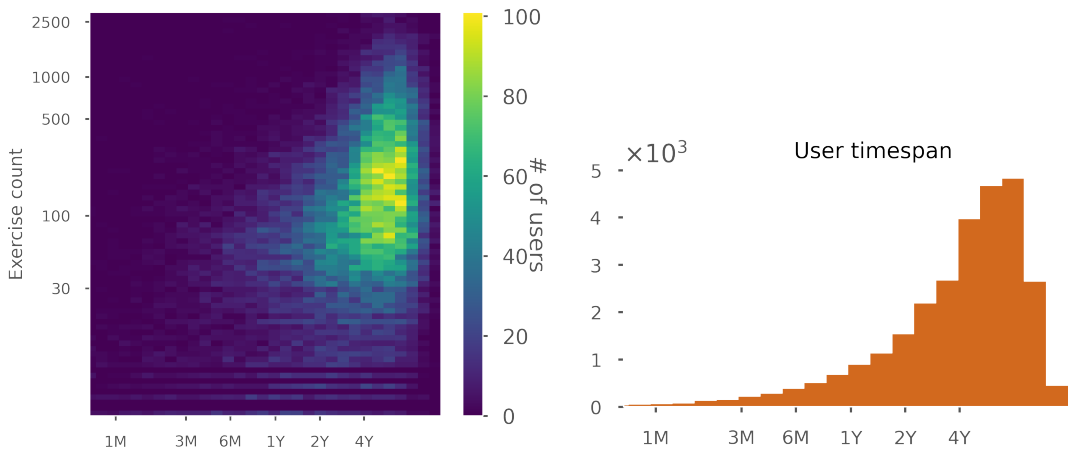


Figure 3.3 Users Time Span Distributions

In order to check frequency of the exercise reporting we can check interevent distribution of the exercise reports. Figure 3.4 reveals that most users are reporting their exercise events on Twitter on the basis of days.

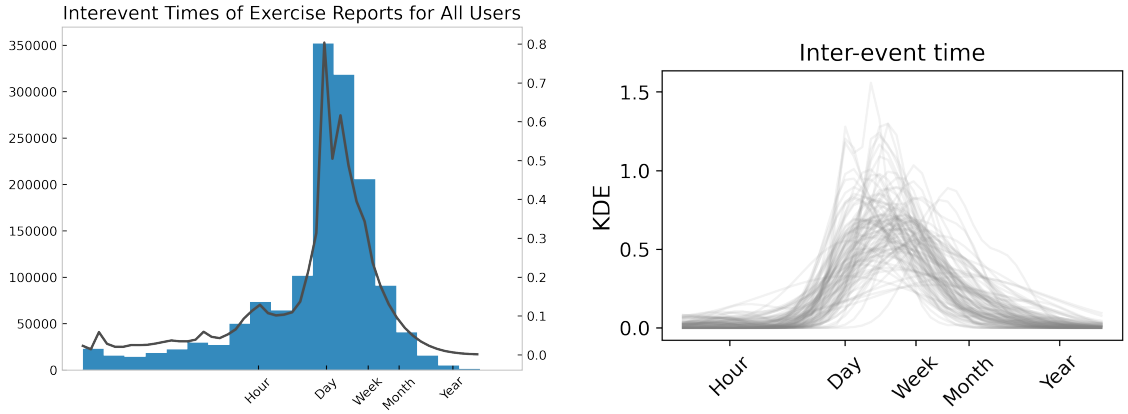


Figure 3.4 User Interevent Distributions

3.4 Exercise Session Preparation

In this section, how user exercise reports are clustered into sessions will be explained in detail. In the rest of the document an exercise session will be called as a period where a user has been doing regular exercises and sharing them on Twitter. Considering a ascending sorted timestamp list for exercise reports of each user, the aim is to assign session labels to each exercise report using the Algorithm 2.

Given the list of exercise report timestamps of user i as:

$$T_{u_i} = [t_1, t_2, t_3, ..., t_n]$$

Sessions are structured as:

$S_{u_i} = [[t_1, t_2], [t_3, t_4, ..., t_{12}], [t_{13}, t_{14}, ..., t_{n-3}], [t_{n-2}, t_{n-1}, t_n]]$ where S : list of clusters for u_i

Algorithm 2 Exercise session is defined as periods in which any consecutive report time does not exceeds a certain number of days. Inertia in the algorithm refers to within-cluster sum of squares.

procedure CLUSTERER(T, d)
    $CurrentCluster \leftarrow 0$
    $ClusterLabels \leftarrow []$
    for all i **in enumerate**(T) **do**
        if $t_{i+1} - t > d$ **then**
            $CurrentCluster \leftarrow CurrentCluster + 1$
        $ClusterLabels \leftarrow ClusterLabels \cup CurrentCluster$
    return $ClusterLabels$

procedure CLUSTERACTIVITYTIMES(T)
    $D \leftarrow$ range(20, 45, 3)
    $InertiaList \leftarrow []$
    for d **in** D **do**
        $ClusterLabels \leftarrow$ CLUSTERER(T, d)
        $InertiaScore \leftarrow$ WCSS$(T, ClusterLabels)$
        $InertiaList \leftarrow InertiaList \cup \{InertiaScore\}$
    $BestDay \leftarrow$ ELBOWPOINT$(InertiaList)$
    $BestClusters \leftarrow$ CLUSTERER$(T, BestDay)$
    return $BestClusters$

Algorithm 2 is used for each user to get the session clustering. Set of days used is selected from 20 to 44 with 3 increments to make the algorithm more efficient in terms of calculation time. With the Algorithm 2 sparse reports for some users are not considered as valid session. In order for a session to be valid we used a threshold $thr = 20$. So if a session has number of reports $< thr$ then, that session was not considered as a valid exercise session in the analysis.

Figure 3.5 exhibits one user and his/her session assignment for the exercise reports.

Figure 3.6b illustrates the resultant distribution of days used to assign exercise sessions to each user as a result of the Algorithm 2. For high number of users, algorithm resulted with using 20 and 23 days as the threshold for interevents to begin a new session.

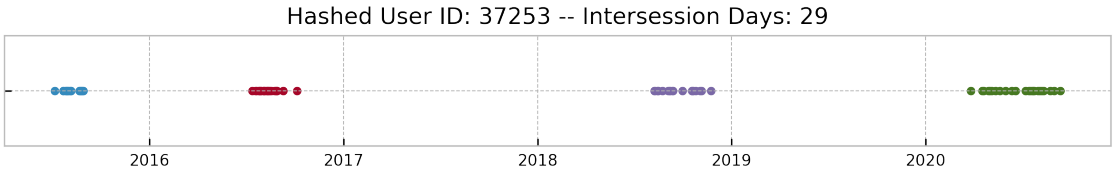
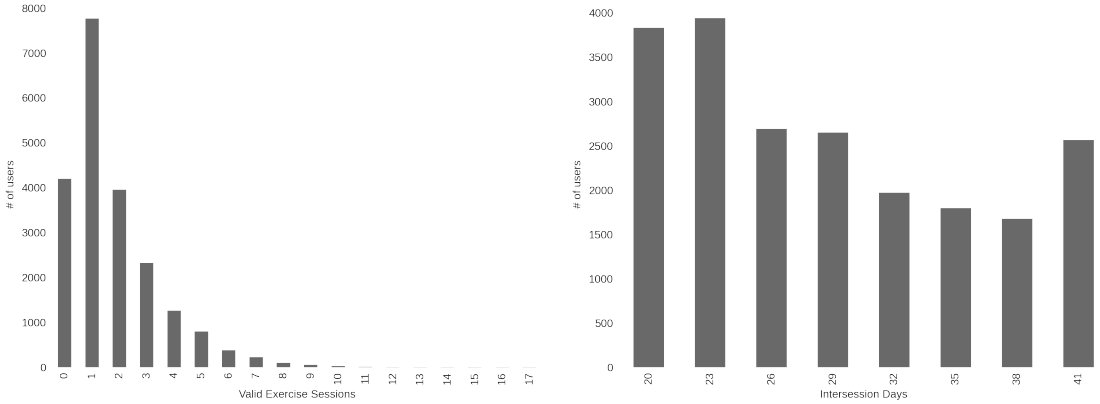


Figure 3.5 Sample Output of Algorithm 2



(a) Number of exercise session per user (b) Days used to create sessions for users

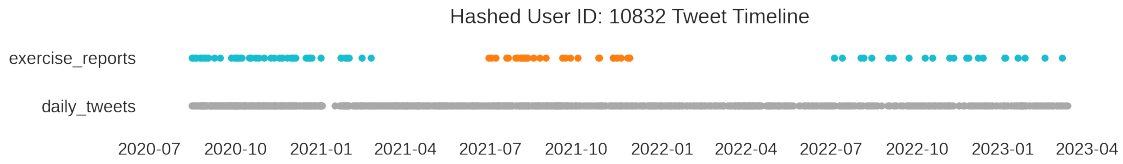
Figure 3.6 Summary statistics from constant clustering algorithm. Session per user (3.6a), Used ϵ values (3.6b)

Figure 3.6a depicts that most of the users in the database has 1-2 valid session. Users with 0 valid session has either sparse reports or they are fast quitters. Taking result 3.6 into account, we might conclude that the majority of users are reporting their activities frequently which makes the data used in the analysis more reliable.

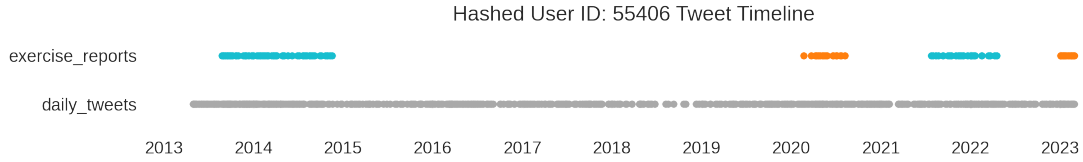
After assignment of exercise sessions to each user, we can inspect some example users with their timelines showing both daily tweets and exercise tweets. Figure 3.7 illustrates some example users. One of the challenges in this study was hinted that exercise session periods for each users has arbitrary starting times with arbitrary duration.

3.4.1 Alternatives to Constant Clustering Exercise Session

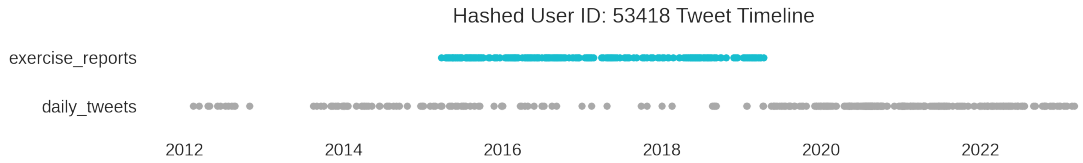
Constant clustering as explained in Algorithm 2 is using a set of days from a list to find the best day that divides timeline into sessions. It has a very similar working logic to DBSCAN algorithm on 1D data where best days is the epsilon parameter in the DBSCAN algorithm. Another trial to find exercise sessions was to apply Jenkspy (Jenks & of Kansas. Department of Geography, 1977) algorithm. This



(a) Sample User 1



(b) Sample User 2



(c) Sample User 3

Figure 3.7 Sample users with timelines. Each user and their timeline is unique in terms of density of the tweets and start-end times.

algorithm tries to find natural breaks in the data. Trials with Jenkspy produced very similar results with the constant clusterer that was applied as a final choice. Performance comparisons for both these algorithms can be found in the Appendix A. Another approach was to try kMeans algorithm on the 1D data. However taking into account that majority of users have 1 session and number of clusters starts from 2, kMeans performed poorly. Additionally, constant clustering algorithm outperformed Jenkspy in terms of computation time.

3.5 Random Users Timeline Collection

Second leg of the dataset consisted of tweets collected for completely random users from Twitter. To get a random set of users we have used Twitter feed. This dataset consisted of captured tweets from live stream. We have extracted a set of users from this stream dataset to further retrieve the historical timeline of random users.

Limitations on API rate limits prevented to collect large number of users. That is why we were able to collect 3000 random users which is very low compared to the exercise sharing users. These users assigned to control group which will we described in the methodology section 4.4 by applying propensity score matching method. With this method, we overcame the issue of low number of users in the control group. Since this random users data consisted of accounts who do not not share any exercise activity on Twitter, we aim to use this portion of the data as the baseline to evaluate the exercise activities on the socially shared content.

4. METHODS

4.1 BERT Models

BERT (Bidirectional Encoder Representations from Transformers) models are built on the Transformer architecture, which was introduced by (Vaswani & et al., 2017). The Transformer architecture incorporates a self-attention mechanism that allows the model to focus on different parts of the input sequence, capturing dependencies between words. BERT, specifically, utilizes the Transformer architecture as its backbone, as described in the original BERT paper by (Devlin & et al., 2018).

The key innovation of BERT lies in its bidirectional nature, which sets it apart from traditional language models. Unlike models that process text in a left-to-right or right-to-left manner, BERT considers both the left and right context of each word. This bidirectional training approach was proposed by (Devlin & et al., 2018), and it enables BERT to have a deeper understanding of the context and meaning of words.

Pretraining BERT involves training on large-scale unlabeled text corpora using a masked language modeling objective and next sentence prediction task. These pretraining techniques were introduced by (Devlin & et al., 2018), and have proven effective in learning robust representations of words, capturing semantic and syntactic information in the data.

Fine-tuning BERT for specific downstream tasks is a common practice. This involves adding task-specific layers and training the model on labeled data. Fine-tuning techniques for BERT have been explored in various studies, including the work by (Sun & et al., 2019) and (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer & Stoyanov, 2019).

Introduction of RoBERTa done by (Liu et al., 2019), a refined version of BERT that

incorporates several modifications and optimizations to improve BERT's robustness, performance and generalization capabilities.

4.2 Twitter-roBERTa-base for Sentiment Analysis

In this study a RoBERTa-base model finetuned for sentiment analysis with the TweetEval benchmark model (Loureiro, Barbieri, Neves, Anke & Camacho-Collados, 2022) was used. The base model was trained on 124M tweets from January 2018 to December 2021. Predictions done by the model produces positive, negative and neutral scores for a given tweet. Score for every category is in the $[0-1]$ range. Closer to the 1 for any category indicates that the model was more confident on the prediction.

Here we made predictions on daily tweets in the database using the model (Loureiro et al., 2022). Table 4.1 reveals sample predictions with positive, negative and neutral scores for each input tweet.

High level model usage example is shown in Figure 4.1. As Figure 4.1 expresses, raw text needed to be preprocessed before feeding into the model. Tokenization step is used to create tokens from words inside the text. Using the Figure 4.1, we have parsed all daily tweets for each user to get associated scores and created the sentiment dataset.

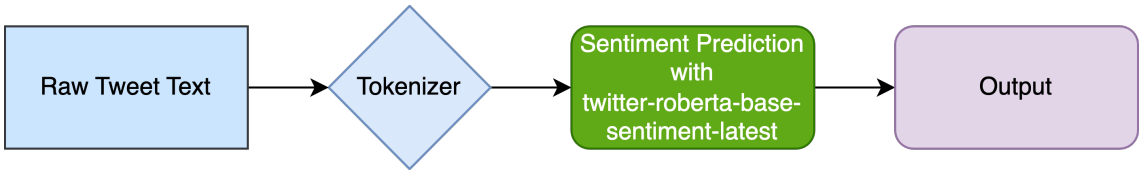


Figure 4.1 Model Usage Flow

4.3 Daily Tweet Sentiment Score Alignment with Exercise Sessions

As explained in section 3.4, session start and end times were mapped to daily tweets to get daily tweets sentiment falling into each exercise and non-exercise periods. This

Table 4.1 twitter-roberta-base-sentiment-latest model scores for sample inputs

Text	Positive	Negative	Neutral
`Hello sis, so you'll be in Davao this coming Sunday. That's nice. I'll bring my family to hear your testimony. God bless...`	**0.983**	0.002	0.014
`Thank you Kemdi for allowing me to be your teacher!! And for the bundt cake!`	**0.986**	0.002	0.012
`To put things in perspective it just took me best part of 30 minutes to change from shorts to joggers`	0.162	0.121	**0.717**
`Tunes for a tenner this Friday on BenMinsterFM 6am til 10am with all money going to MacNEYorks fundraising total so get in and requesting.`	0.213	0.004	**0.683**
`Thank you for highlighting this issue HaringeyCP The pool closures have caused untold disruption to our members and their families as well as our staff.`	0.035	**0.653**	0.312
`Unbelievably, we have now been notified the pool is CLOSED tonight 20/1 and tomorrow Saturday 21/1. We cannot apologise enough even though this is completely out of the control of the club.`	0.007	**0.847**	0.146

alignment was done to identify stats and trends of sentiment scores before, during and after exercise session periods. As a result of the alignment process we had session and non-session periods in each user's timeline. Figure 3.7a depicts there are three session periods and 2 non-session periods for duration between exercise sessions. User in Figure 3.7b on the other hand, has a single session period and two non-session periods.

4.3.1 Unification Of Sentiment Scores

Table 4.1 demonstrates that output of the model is three different scores for three different category. At this part, we aimed to create a unified sentiment score for every single tweet. We call the output of this process is compound score. Simply taking the difference of positive and negative score to while ignoring the neutral score was the approach.

Table 4.2 Compound Score Illustration

Positive	Negative	Neutral	Compound
0.213	0.004	0.683	**0.209**
0.035	0.653	0.312	**-0.618**
0.162	0.121	0.717	**0.41**

Compound column in Table 4.2 is the calculated as $score_{positive} - score_{negative}$. Compound values fall into space $[0-1]$. Calculating the compound score for each single tweet allowed to evaluate any tweet easily without doing multiple analysis for all score types.

4.3.2 Normalization of Sentiment Scores on User Level

Another score transformation is done by finding Δ scores for each tweet. The reasoning this normalization is that every user has her/his sentiment score range. For instance some users might usually share sad posts while others usually share happier posts. Simply we subtracted compound score from average compound score for each user. With this methodology, we will be analyzing how the users' score changes in time with respect to their own average. Following formula describes how we normalized the scores on user basis.

$$\Delta_j = score_j - \mu_i \quad \text{for } i \in Users, j \in Tweets_i$$

4.3.3 Label Smoothing

As mentioned by (Müller, Kornblith & Hinton, 2019), label smoothing is used as a regularization method to overcome the generalization of model prediction. It is noted that label smoothing can distribute the the prediction mass from higher class to dominated classes. In our case, pretrained model that we have used tend to give

high score for *positive* class. We have applied the smoothing methodology described in the Algorithm 3 to distribute to scores towards *negative* and *neutral* classes. Figure 4.2 illustrates the compound score distribution for each smoothing constant (*alpha*). As a rule of thumb $alpha = 0.2$ had been chosen for the further analysis. This way we keep the model prediction close to the original by distributing mass from the *positive* class to rest. Compound score is defined like $score_{positive} - score_{negative}$ as described in section 4.3.1.

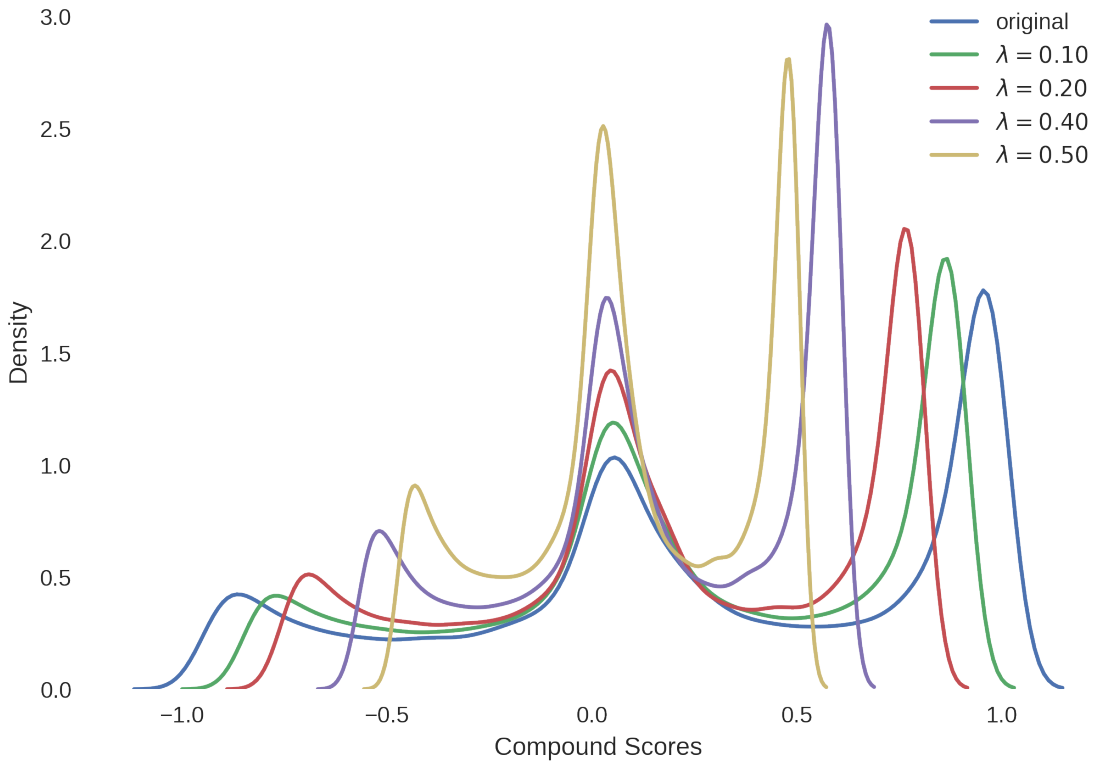


Figure 4.2 Original model scores were extreme. Higher the smoothing constant, the more shrank the compound score is.

Algorithm 3 Linear label smoothing method used

1: **Input:** Dictionary *scores*
2: **Output:** Dictionary *scores_smoothed*
3: **procedure** LABELSMOOTHING(*scores, alpha*)
4: Initialize *scores_smoothed* as an empty dictionary
5: **for each** *label* **in** *scores* **do**
6: $scores\_smoothed[label] \leftarrow (1 - alpha) \times scores[label] + \frac{alpha}{\#\ \textbf{of labels}}$
7: **return** *scores_smoothed*

4.3.4 Bucketization of Periods

Even tough we map the exercise sessions to daily tweets timeline, one other challenge was to create aggregate sentiment trends. In pursuance of same base timeline for each user we bucketized periods into n bins. In other words, we converted all session and non-session periods for all users into n bins of equal spaced time buckets. Each bucket is represented as the mean of sentiment scores for tweets that fall into each bin. Figure 4.3 is an example of bucketization process where. Pseudo-code 4 portrays the methodology for score calculation mean for bins where the input is the result of the bucketization from Figure 4.3. As explained in the algorithm, input is the list of lists from the period's buckets. this methodology allowed us to represent every period with equal length where each item in a period represents the average sentiment score for a given user. Introduction of bucketization method led to shrinkage of longer periods of users and expansion of shorter periods to the same base. Considering the fact that users have distinct time interval of exercise periods, unifiying them into same base allowed us to aggregate users in terms of their sentiment trends.

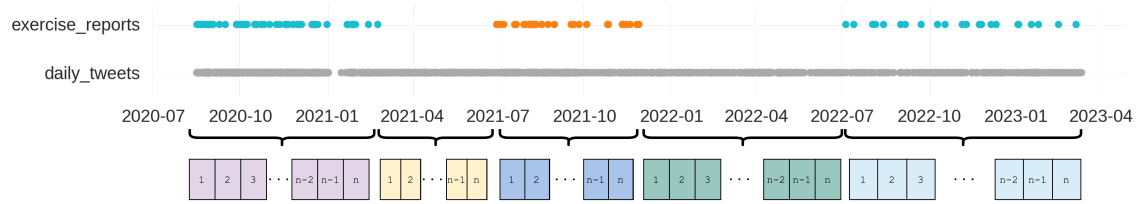


Figure 4.3 User periods seperated into equal length buckets. The user has three session periods and two non-session periods. Each period here is divided into equal-length bins in terms of duration.

Algorithm 4 Bucketization Averaging
$P = [[bin_{1_1}, bin_{1_2}, ..., bin_{1_k}], [bin_{2_1}, bin_{2_2}, ..., bin_{2_l}], ..., [bin_{n_1}, bin_{n_2}, ..., bin_{n_m}]]$ where k,l,m: tweet count falls into bins 1,2, n respectively

procedure BINMEANS(P)
    $bin_{means} \leftarrow []$
    for b **in** P **do**
        $bin_{avg} \leftarrow$ AVERAGE(b)
        $bin_{means} \leftarrow bin_{means} \cup \{bin_{avg}\}$
    return bin_{means}

Since each user's timeline is unique to itself, bucketization logic creates a same base timeline for all users. Now, we might take a look at the distance distribution of buckets to their relative period ends. Figure 4.4 demonstrates how much middle of buckets are away from the period end. It can be deducted from the graph that duration of bucketized periods are 1 year long on average across all users.

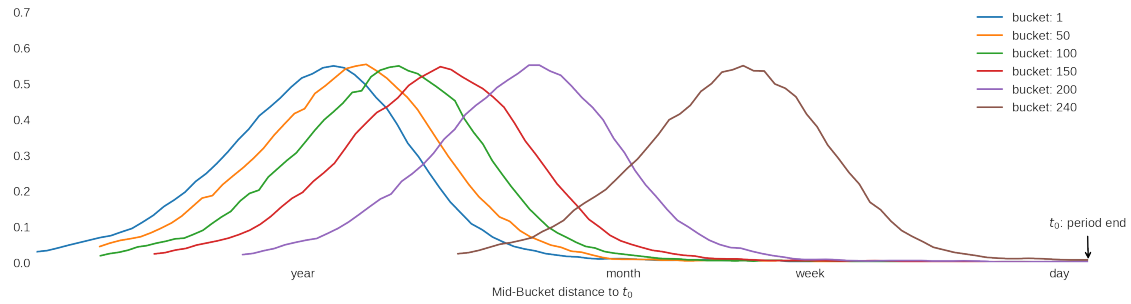


Figure 4.4 Distribution of real time distance of buckets to the relative period end. For instance on the average first bucket is at 1 year away from its period end while bucket number 240 is 1 week away in average.

4.4 Propensity Score Matching

Comparison of user sentiment on exercise and non-exercise periods requires two set of users one being the treatment group and other being the control group. In the context of this study, people who share exercise activities on Twitter were labelled as treatment users where doing exercise is the treatment itself. Control users on the other hand are people who does not share any exercise activity on Twitter. Although this assignment does not guarantee that those control users are not doing any exercise, but they are not sharing their exercise activities on social platforms. Eventually, we focus our analysis on the assumption that control users posts on Twitter were not affected by their exercise activities if there are any.

In construction of control and treatment group, we have conducted a method called Propensity Score Matching(PSM) in the literature. Aim of PSM is to create similar users for both group in terms of their social profile. Instead of using pure random users for the control group, by applying PSM, we aimed to create a control group with similar profiles to the treatment users.

4.4.1 Overview of Propensity Score Matching

The concept of Propensity Score Matching (PSM) was brought to literature with the study by (Rosenbaum & Rubin, 1983). In statistical analyses of a treatment on a particular sample, Propensity Score Matching (PSM) is being used by matching similar samples from treatment and control group in terms of the propensity score that describes the covariates of samples as a singular value. Usually unmatched

samples are disregarded as noted by (Little & Rubin, 2000). Usage of nonrandomized data should be done by selecting the controlled units carefully as discussed by (Rubin, 1974).

Some definitions on the literature:

- ATE: Average Treatment Effect is the average treatment effect across all units in the population both treatment and control group.
- ATT: Average Treatment Effect on the Treated units
- ATC: Average Treatment Effect on the Controlled units at the counterfactual scenario where the control group receives the treatment

4.4.2 Covariate Selection

In this study, the treatment and control group are users who do exercise and share on Twitter and users who do not share their exercise activities on Twitter respectively. To find the similar users from each group, we applied PSM with the following covariates. These covariates can be described as the variables that define a social profile from a broad perspective.

- number of followers
- number of friends
- average tweet count per week
- user time span
- average sentiment score of tweets
- std sentiment scores of tweets

At this step we applied logarithmic transformation on covariates. Log-transforming features before normalization is a common data preprocessing technique used to handle skewness in the data. Skewness refers to the asymmetry of the data distribution, where the tail of the distribution is stretched out towards one side more than the other. When dealing with skewed data, log transformation can be beneficial for several reasons:

- **Symmetry:** Taking the logarithm of data can often make the distribution more symmetric. This can be helpful because many statistical methods and

machine learning algorithms assume that the data is normally distributed or close to it.

- **Variance Stabilization:** For instance variance for number of followers increases with its mean. Log transformation can help stabilize the variance, making it more consistent across different levels of the feature.
- **Linearization:** Logarithmic transformations can convert multiplicative relationships into additive ones. In other words, taking the logarithm of a feature can make relationships between variables more linear, which was beneficial for logistic regression that assumes linearity.
- **Outliers Mitigation:** Log transformation can mitigate the impact of extreme outliers. For example, user time span is one such case where outliers exists. By compressing large values towards the mean, log transformation can make the data less sensitive to extreme observations.

4.4.3 Propensity Score Calculation

Here at this step, we have the data to train a model to predict propensity scores for users. The target for to be trained Logistic Regression model is binary label that user being in treatment:1 or control:0 groups. Figure 4.5 shows the high level flow we used to calculate the propensity scores.

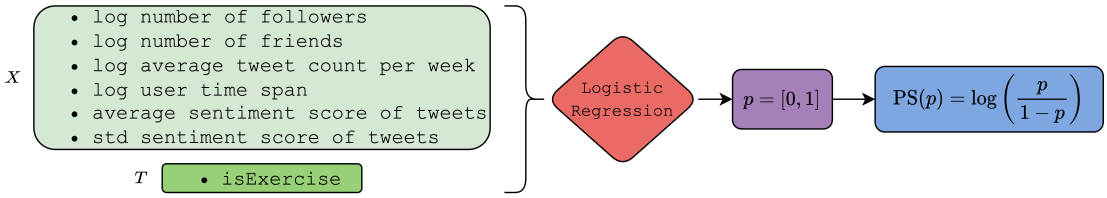


Figure 4.5 Steps followed to calculate propensity scores (PS) for each user

After fitting a Logistic Regression model on the whole data, we predicted the training data to get the model predictions. Although this method of predicting the train data is not applied on many prediction task, the goal was not to predict which class is predicted correctly. The goal was to get prediction probabilities for each instance. Still, we can take a look at the confusion matrix created from model training in Figure 4.6.

As we can see from the confusion matrix, the mode did perform quite good in terms of f1-score even tough the data was imbalanced.

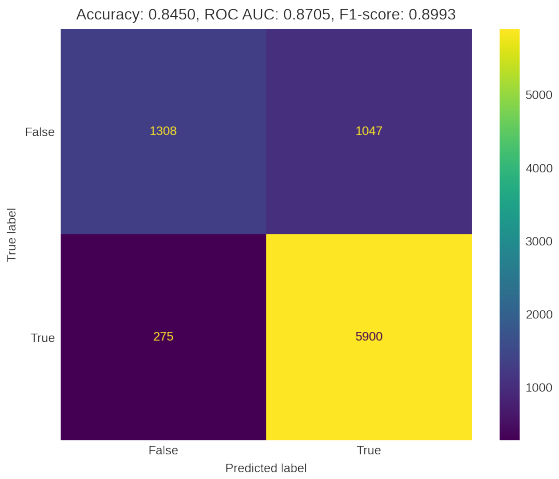


Figure 4.6 Model Prediction Results

After we get the prediction probabilities, we used logit transformation using the Formula 4.1 to transform probabilities from $[0, 1]$ to $[-\infty, \infty]$. Logit transformation is quite handy in these situations since it also takes the model confidence into account. Higher the model confidence, higher the logit transformed value. Finally, we used the logit transformed value of prediction probabilities as propensity score for the matching algorithm as the Figure 4.5 covered.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{4.1}$$

4.4.4 Matching

Now we have done the propensity score part of the PSM, it is time to match samples from treatment to control. There are different types of matching logics in the literature such as Nearest Neighbor, Mahalanobis Metric or Caliper Matching as described by (Thavaneswaran & Lix, 2008). Effectiveness of each matching strategy is depending on the experiment specifications. In our study, the treatment group was the dominating side regarding the population size. So we have chosen to continue with Caliper Matching to oversample the control group.

As mentioned by (Sianesi, 2001), caliper matching is done by using an epsilon value to find close users. We used quarter of standard deviation of propensity scores of the treatment group as the epsilon constant. Below is the logic for Caliper Matching algorithm.

$$\text{Range} = |P_i - P_j| < \epsilon$$

where: P_i is the estimated propensity score for the treated subjects i,

P_j is the estimated propensity score for the control subjects j,

ϵ is the constant that we used as $\frac{\sigma}{4}$ with σ: standart deviation of propensity scores

Figure 4.7 lays out the example matching using the Caliper Matching.

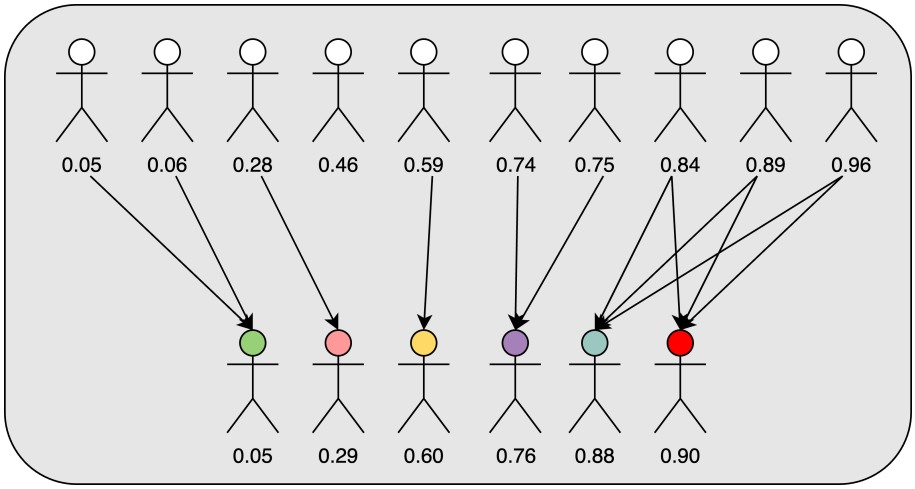


Figure 4.7 Top&Bottom rows are Treatment&Control Users with their corresponding propensity scores.
Treated users are matched with controlled users using Caliper Matching

4.4.5 Evaluating the Results of PSM

To evaluate the results of the matching, we can take a look at the distribution of covariates and the propensity score distributions for both treatment and control groups at before and after the matching. Keeping in mind that the purpose is to create similar group of users in terms of covariates, the more similar the distribution after the match is the better matching performance. Figure 4.9 is complete comparison for covariates that were used to calculate the propensity scores and their distributions before and after the matching applied. It can be deducted from the distributions that matching brought variable distributions closer. For instance, similarity for mean sentiment scores of users after the matching is considerably well.

4.5 Time Series Construction

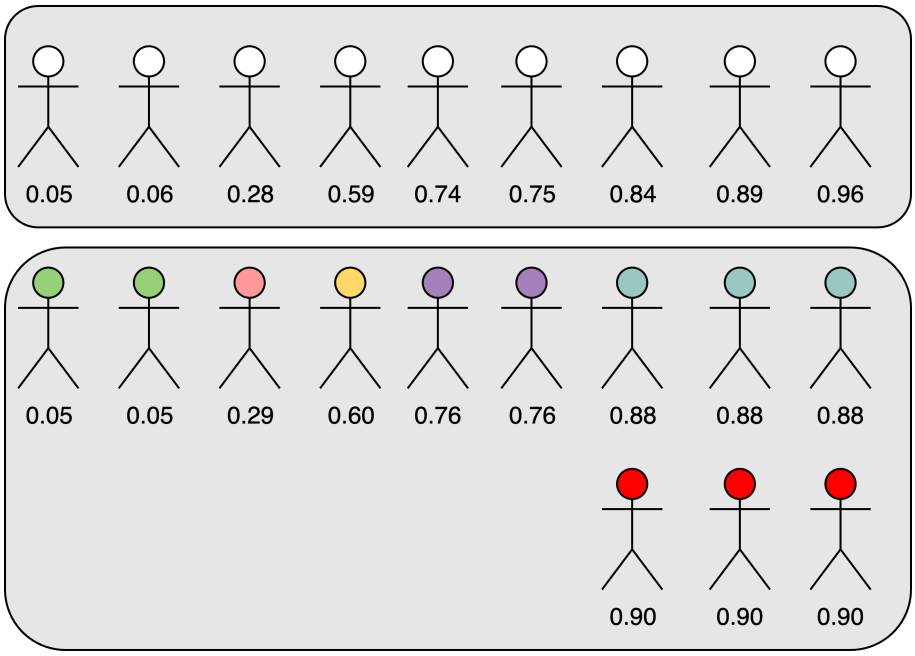


Figure 4.8 Treatment and control populations after matching. For instance, red and turquoise samples were matched with multiple treated samples as shown in Figure 4.7 so they are used in the matched population as the number of matches. Treated users that does not have any match are also eliminated.

To evaluate the aggregate sentiment score shifts at exercise periods we have used time series that is created from the buckets of each user. Compared to usual timeseries where the data has constant time intervals, we have constant buckets in the data. So in other terms, buckets that we have created are considered as time slots for the timeseries analysis. Created bucket-based time series is at the aggregate level by aggregating on user level. Figure 4.10 clarifies the approach. Regardless of the number of buckets, we aggregated user buckets into a single bucket for each period. To give an example, consider 10 users with only one exercise period in their timeline. Then these user will have three periods named before, during and after exercise period. Then we will aggregate these 10 users into 3 periods by taking average of each bucket in corresponding periods.

Similar to work from (Fan et al., 2019) where they aligned users based on affect labeling times, we have aligned all bucketized non-session and session periods separately for all users based on the bucket numbering. With this aggregation we would be able to construct time series from user level data to aggregate level. Buckets with no tweets fallen into, we used the interpolation method to fill the empty buckets. On the interpolation step, we used the average scores of the closest non-empty buckets from both left and right side of the corresponding bucket. Figure 4.11 depicts the filling process.

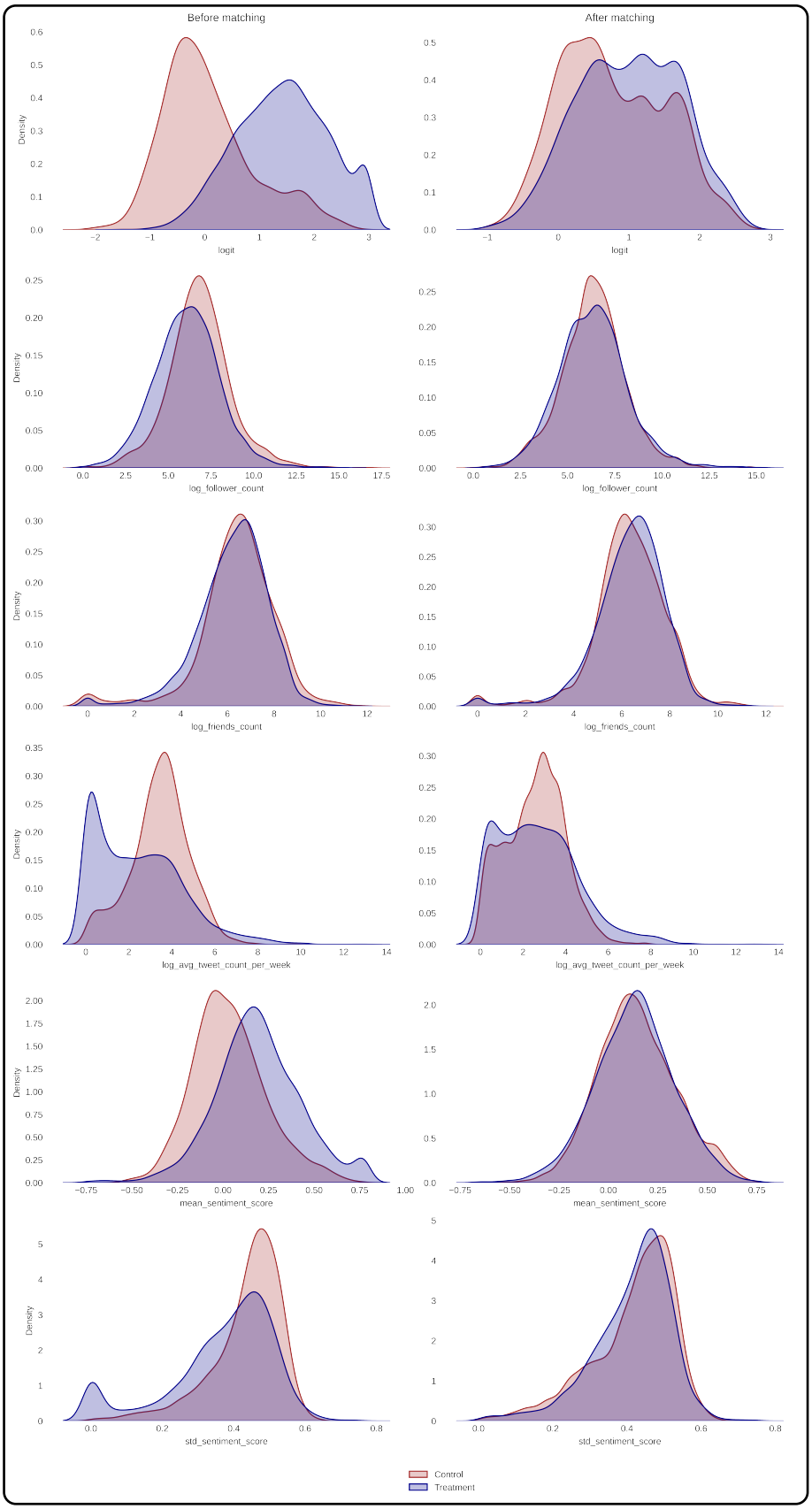


Figure 4.9 Distributions of propensity scores (logit), and covariates in Treatment and Control groups before and after the PSM. Matching produced promising results especially in logit, follower counts, mean and std sentiment scores by making distribution of both group much similar.

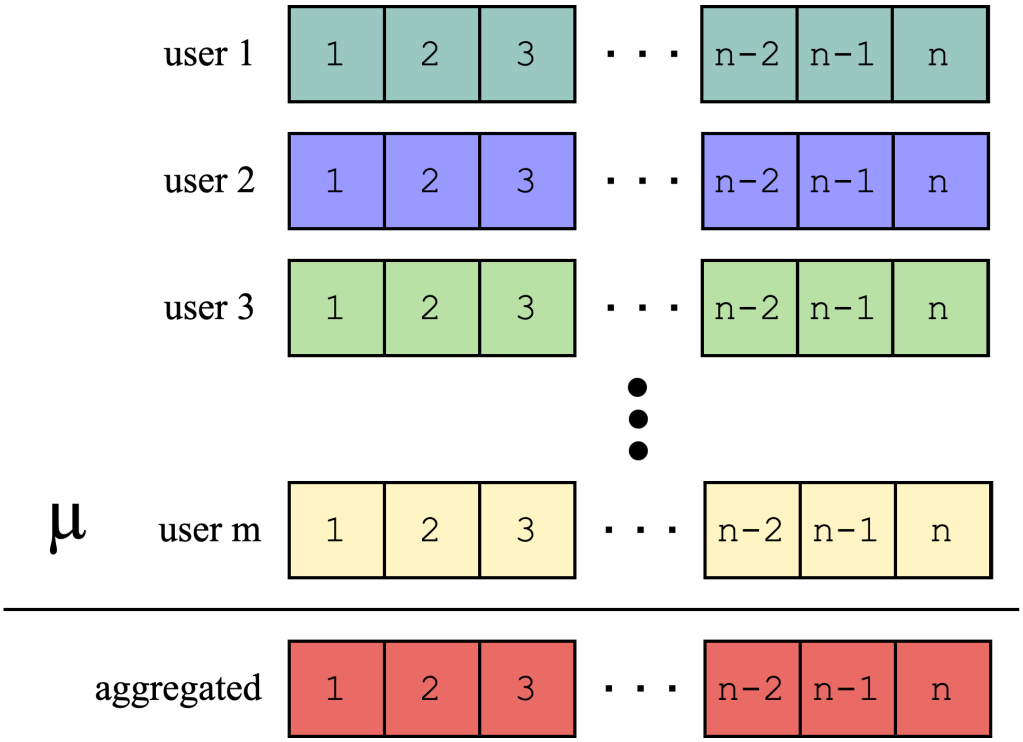


Figure 4.10 Aggregation of users with on bucket level. Aggregation is done by taking the average of each bucket within itself.

Figure 4.11 Interpolation of empty buckets(j) with average score of closest non-empty buckets $(j-1)$ and $(j+1)$ using the formula $S_j = \mu(S_{j-1}, S_{j+1})$

5. RESULTS

5.1 Propensity Score Matching

Considering that the treatment is being sharing exercise activities on Twitter, treatment effect we have calculated is amount of shift from the daily baseline for users. In other words, effect of sharing exercise activities on social media is sentimental increase or decrease in usual tweets. To calculate the treatment effect in our study, we used nonrandomized data by applying propensity score matching as discussed in the previous chapter. Table 5.1 summarizes the estimated effects of the treatment on the units.

We can read the table by pointing out that users have 0.011 uplift in the sentiment scores of posted contents on average (ATE). ATT is telling us that treated users sentiments going into positive direction from their sentiment base during their exercise periods with an estimate of 0.013. ATC 0.010 tells us that the in a counterfactual scenario where the controlled units get the treatment, they would have share more positive content in social media. Similarly we can observe from Figure 5.1 that the average score shift for treatment group while the control's average is staying around 0 which is also supporting the estimates for the treatment effect.

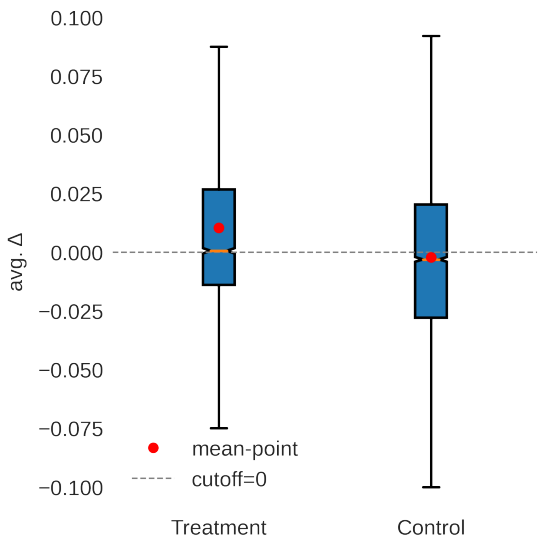


Figure 5.1 Showing the average Δ scores for treatment&control groups

Table 5.1 Treatment Effect Estimates: Matching

	Est.	S.e.	z	P> |z|	[95% Conf. int.]
ATE	0.011	0.004	2.747	0.006	0.003–0.019
ATC	0.010	0.005	1.915	0.056	-0.000–0.021
ATT	0.013	0.004	3.330	0.001	0.005–0.021

5.2 Different Periods Sentiment Comparison

Here we can take a look at the Figure 5.2. As we observed from the plot that Δ score for users on average is following an increased trend during the exercise period compared to before and after the exercise periods. Sub-areas commenting on the Figure 5.2:

- Before Exercise Period (Left lightblue region)

  This period corresponds to the aggregate bucketized time of users before they started sharing any exercise activities. Each dot in the scatter represents the average Δ score of treated users. Each point's (j) value is formulated as

  $$\mu_j = \frac{\Sigma_{i=1}^{m} \Delta_i}{m} \text{ for } i \in \{1,2,3,...,m\}, j \in \{1,2,3,...,n\}$$

  where m represents the number of users and n represents the number of buckets. It can be seen that there is a stationary trend on this region around 0 which means sentiment scores are not affected by any intervention on the overall scale.

- During Exercise Period (Middle green region)

  Here in this region, it is clear that there is an uplift from the base. Δ scores on average follows an improved trend with an addition of a δ amount. Average sentiment shift in here can be formulated as

  $$\Delta + \delta$$

  where δ represents the treatment effect.

- After Exercise Period (Right darkblue region)

  After period clearly shows there is a downside trend which can be thought as going back to the base. The downtrend suggest that in time the value of treatment uplift (δ) is going back to 0.

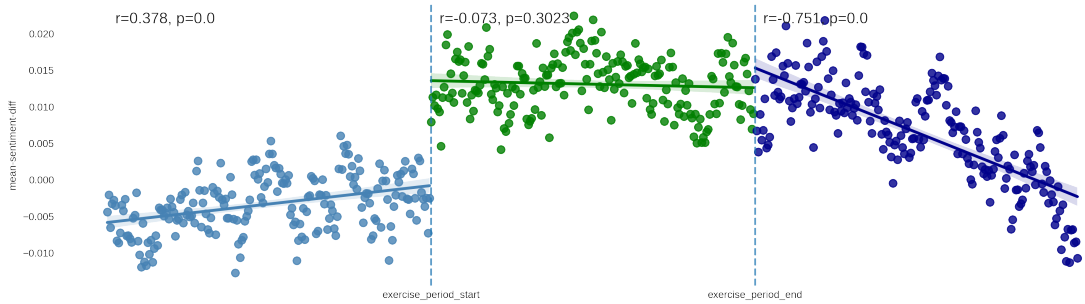


Figure 5.2 Average Δ levels for periods before, during and after on aggregate level

Similarly, we can investigate to the scatter points and their trends around exercise start and exercise end periods including the control group which we know that scores in the group are not affected by the exercise sharing. Figure 5.3 clearly points out that the control group follows a similar trend without any δ uplift after the change point while the treatment group (blue and green colored parts) explicitly indicate the clear δ uplift on the overall. Figure 5.4 details what trend is followed by the treatment and control groups in the same frame. Downside trend for the treatment can be seen while the trend shift did not occur for the control around the change-point. Treatment group's trend can be explained as going back to the normal state. One thing to note in here is that control group's delta scores are higher compared to the treatment before the change-point. This is expected as the Δ itself is created by subtracting the sentiment score from the whole-time average for all users.

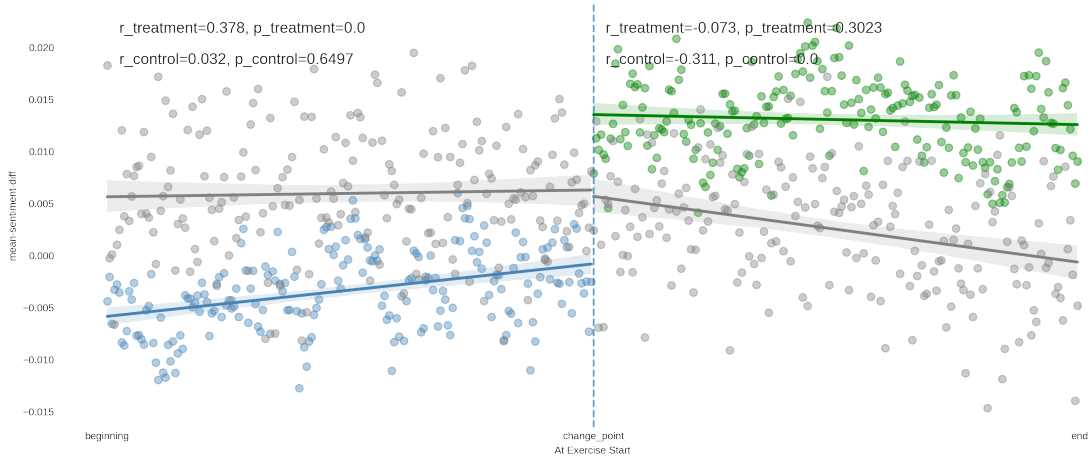


Figure 5.3 Aggregated Δ scores at exercise start change-point comparison. Gray scatters are representing the control group. Greens are the points during the exercise periods and lightblue scatters are for before exercise period.

More detailed version of the Figure 5.3 by separating the treatment and control groups around change point portrayed at the Figure 5.5a and Figure 5.5b respectively. Again, while the δ elevation can be seen clearly for the treatment, we can not see any elevation or demotion for the control after the change-point. Disconti-

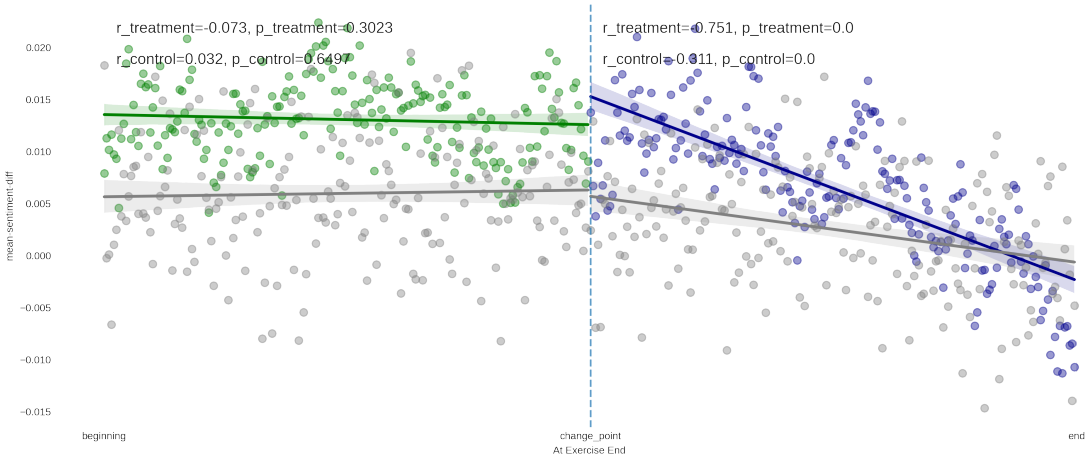
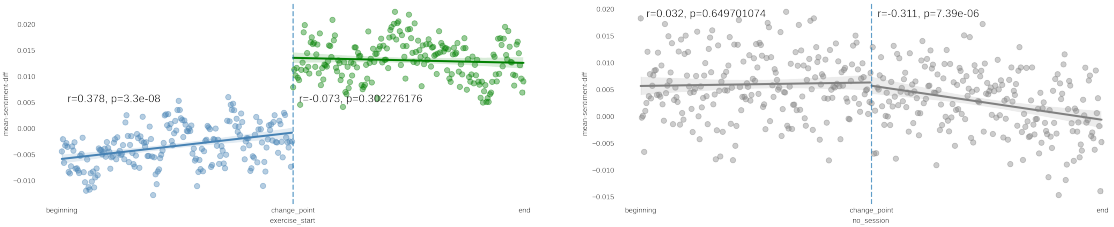


Figure 5.4 Exercise end change-point comparison. Grays represent the control group. Green and darkblue points are during and after exercise period respectively.

nuity in the regression line on the treatment is visually clear to be noticed in Figure 5.5a. To elaborate the results on Figure 5.5a we applied regression discontinuity using linear regression with the formula $\Delta \sim x * isExercise$ where x is time content of the regression to distinguish between before and during exercise periods. Table 5.2 summarizes the regression analysis results while pointing out the coefficient for $isExercise$ being 0.0204 representing the treatment effect calculated by the regression discontinuity analysis. Again we can see the constructive effect of treatment on the Δ scores.



(a) Exercise start period Δ score trends for treatment group

(b) Change point Δ score trends for control group

Figure 5.5 Startline detailed for Treatment(a) and Control(b) groups.

Table 5.2 Regression Discontinuity with Linear Regression

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.0059	0.001	-11.145	0.000	-0.007	-0.005
x	2.527e-05	4.54e-06	5.562	0.000	1.63e-05	3.42e-05
isExercise	**0.0204**	0.001	13.721	0.000	0.017	0.023
x:isExercise	-3.012e-05	6.43e-06	-4.687	0.000	-4.27e-05	-1.75e-05

5.3 Time Series Exploration

The Figure 5.6 provides an overview of the average delta shifts allowing for a quick assessment of the effectiveness of the intervention. The uplift on the treatment group supports the key message that the exercise sharing has positive influence on the tweets posted by the treated users in terms of the sentiment scores. It gives a visual representation of the impact of the treatment on the delta score which is the amount of distance from the baseline of individuals. This figure significantly contributes to the understanding and evaluation of the intervention's effectiveness.

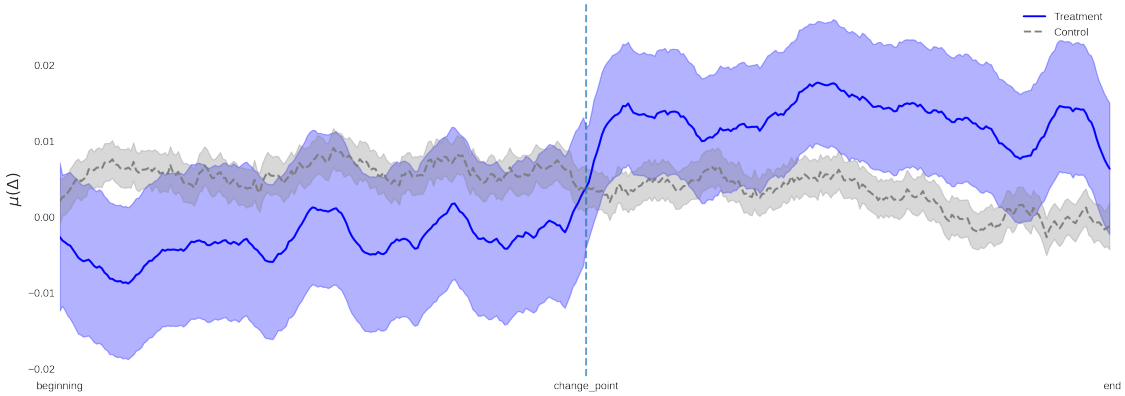


Figure 5.6 Treatment&Control groups Δ score shifts. x-axis reveals the beginning of a period where both group did not share any exercise activity. Then treated users starts their exercise periods that we labelled as the change-point. Shaded areas are 95% CI.

It should also be noted that the time series has been built by applying a moving average method with 20 window size on bucket averages using number of buckets as 200. Smoothed lines depict the above-mentioned uplift with an amount of $\delta \sim 0.13$ which was calculated as the average treatment effect (ATT) by causal inference methodology with the propensity score matching. It is also interesting that that the confidence interval is much wider for the treatment group scores which is caused by the matching method since we over-sampled the control group while removing the treated users who does not have any match in the control group.

5.4 Sentiment Score Shifts

In this section, we delve into the journey of sentiment scores over time, specifically examining their path before, during, and after exercise periods. To visually represent this trajectory, Figure 5.7 illustrates how the sentiment scores evolve through these different phases.

Initially, before the exercise period commences, the scores are concentrated in the bottom-left region of the plot. Here, the average Δ (change) in scores hovers just below 0, indicating a slightly negative sentiment shift, with an average standard deviation of approximately 0.35. This suggests that, on average, users' sentiments are slightly on the negative side, but there is a variability in their individual sentiment scores.

As the exercise period begins, a notable shift in sentiment scores occurs. They move towards the top-right region of the plot, and both the mean Δ scores and the standard deviation experience an uplift. The shift towards the top-right signifies a positive change in sentiment, indicating that users' mood is improving during the exercise period. Additionally, the increase in standard deviation implies that users' sentiments become more diverse, with some experiencing substantial positive changes while others may have more moderate shifts.

During the exercise period, the plot takes on a mountain-like shape, with a noticeable ramp-up region. This ramp-up region indicates a significant increase in mean Δ scores during this phase. It implies that as users engage in exercise activities, their sentiment scores undergo a more substantial positive change on average. This suggests that exercise is associated with a mood improvement, leading to more positive sentiment expressions on the platform during this period.

As we progress into the after-exercise period, the sentiment scores return to the vicinity of where they started in terms of the mean Δ scores. The return journey shows that after the exercise session concludes, users' sentiment tends to revert to a similar level as before the exercise period. This suggests that the positive sentiment boost experienced during exercise is sustained for the post-exercise period duration after the exercise ends. One other conclusion we can make from this journey is that it might trigger the users to go into another exercise period to retain their uplifted mood.

In summary, Figure 5.7 provides a comprehensive representation of the sentiment score trajectory throughout the exercise timeline. It shows the initial slightly negative sentiment with variability, the positive sentiment shift during exercise, and the subsequent return to the pre-exercise sentiment levels after the exercise period. This analysis sheds light on how exercise impacts users' sentiment on the platform and

how sentiment scores evolve in response to exercise-related activities.

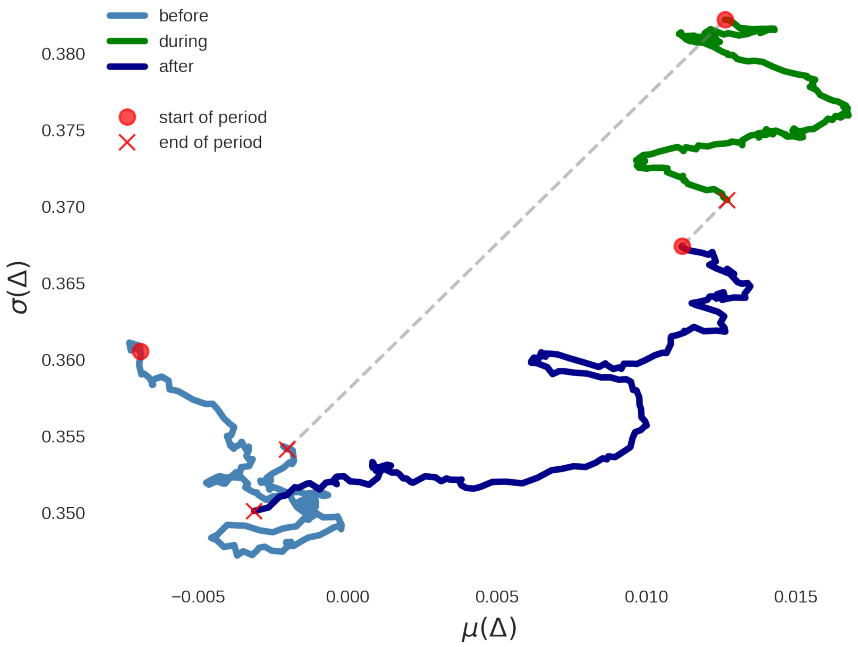


Figure 5.7 Sentiment score path captured on aggregate level. x-axis shows the average Δ scores during each period while the y-axis showing the standard deviation.

5.5 Score Shift Variances

The standard deviation timeseries plot, shown in Figure 5.8, reveals a fascinating insight into the variability of sentiment scores for posted content during exercise periods. This plot represents the average standard deviation on user buckets, and it becomes apparent that during the exercise period, there is a significant increase in the standard deviation. This finding suggests that the mood of tweets exhibits much more variation compared to the periods before and after exercise.

One intriguing observation is that there is a sharp uplift in standard deviation as soon as the exercise period starts. This implies that users' moods begin to fluctuate more right after they commence their exercise activities. This increase in standard deviation indicates that the overall mood variability among users is boosted after they start sharing their exercise experiences and activities on the Twitter.

However, it is important to note that this heightened mood variability is not a

permanent state. After the exercise session ends, the fluctuation in mood gradually returns back to a more normal state, which was approximately at a standard deviation of ∼0.35 at the very beginning.

Based on this finding, it can be concluded that exercise reporting serves as a mood variability booster. This indicates that users are sharing both extreme and non-extreme content related to their exercise experiences. When we refer to "extreme" content, we mean content that exhibits sentiment scores strongly leaning towards the positive side. In other words, users experience instant mood boosts caused by exercise, which is reflected in their posts during exercise periods.

In summary, the standard deviation timeseries plot shows how the mood of posted content varies throughout different periods, with a notable increase in mood variability during exercise periods. This suggests that users experience fluctuating emotions and heightened positivity during exercise, which is evident from the sentiment scores shared on the platform.

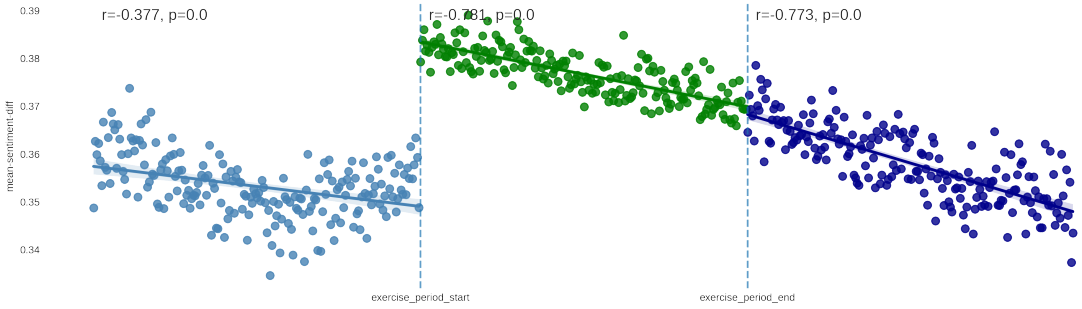


Figure 5.8 Aggregated(average) standard deviation of the Δ scores.

6. DISCUSSION & CONCLUSIONS

In conclusion, our research focused on investigating the effects of exercise sharing on a social platform, specifically analyzing the impact on daily shared content. By utilizing data obtained from Twitter's research API, we were able to explore and identify these effects. Our data processing approach incorporated a novel method called bucketization, which allowed for aggregated comparisons and analysis.

Through the application of propensity score matching, we addressed data distribution mismatches between the treatment and control groups, enabling a more reliable comparison. Utilizing the time series data created for exercise and non-exercise periods, we compared the treatment and control groups based on the average shifts from user-level baselines.

Our study revealed two main findings. Firstly, we discovered that users who shared exercise activities on social platforms also exhibited a higher tendency to share more positive content during their exercise periods, as indicated by sentiment scores calculated using a pre-trained RoBERTa model. This suggests a potential positive association between exercise engagement and the overall sentiment of shared content.

Secondly, we observed mood jumps during exercise periods, indicating instant changes in sentiment scores influenced by the exercises. These instantaneous jumps provide valuable insights into the immediate effects of exercise on social media platforms. Further analysis of these instantaneous effects could provide a deeper understanding of the dynamics between exercise and mood on social platforms.

For future work, it would be beneficial to explore the underlying mechanisms driving the observed effects. Conducting qualitative research, such as interviews or surveys, could provide insights into users' motivations for sharing exercise-related content and their perceptions of the impact on mood and sentiment. Additionally, expanding the analysis to other social media platforms and considering a broader range of exercise types and user demographics could contribute to a more comprehensive understanding of the relationship between exercise sharing and social media engagement.

In summary, our research contributes to the growing body of knowledge on the effects of exercise sharing on social platforms. The findings suggest that exercise engagement on these platforms may have positive implications for shared content and mood. This has implications for individuals, public health initiatives, and the design of social media platforms. Future studies can delve deeper into the topic, addressing potential limitations and exploring additional factors that influence the relationship between exercise and social media engagement.

BIBLIOGRAPHY

Althoff, T., White, R. W., & Horvitz, E. (2016). Influence of pokémon go on physical activity: study and implications. *Journal of medical Internet research*, *18*(12), e315.

An, H.-Y., Chen, W., Wang, C.-W., Yang, H.-F., Huang, W.-T., & Fan, S.-Y. (2020). The relationships between physical activity and life satisfaction and happiness among young, middle-aged, and older adults. *International journal of environmental research and public health*, *17*(13), 4817.

Belcher, B. R., Zink, J., Azad, A., Campbell, C. E., Chakravartti, S. P., & Herting, M. M. (2021). The roles of physical activity, exercise, and fitness in promoting resilience during adolescence: effects on mental well-being and brain development. *Biological psychiatry: Cognitive neuroscience and neuroimaging*, *6*(2), 225–237.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1–8.

de Vibe, M., Solhaug, I., Rosenvinge, J. H., Tyssen, R., Hanley, A., & Garland, E. (2018). Six-year positive effects of a mindfulness-based intervention on mindfulness, coping and well-being in medical and psychology students; results from a randomized controlled trial. *PloS one*, *13*(4), e0196053.

Devlin, J. & et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eo, S.-W., Lee, Y., Yu, K., & Park, W. (2016). Establishing the process of spatial informatization using data from social network services. , *34*(2), 111–120.

Fan, R., Varol, O., Varamesh, A., Barron, A., van de Leemput, I. A., Scheffer, M., & Bollen, J. (2019). The minute-scale dynamics of online emotions reveal the effects of affect labeling. *Nature Human Behaviour*, *3*(1), 92–100.

Fox, K. R. (1999). The influence of physical activity on mental well-being. *Public health nutrition*, *2*(3a), 411–418.

Godes, D. & Mayzlin, D. (2002). Using online conversations to study word of mouth communication. *SSRN Electronic Journal*.

Golder, S. A. & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881.

Hillman, C. H., Erickson, K. I., & Kramer, A. F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Reviews Neuroscience*, *9*(1), 58–65.

Holst, A. (2021). Iot connected devices worldwide 2019-2030.

Jacob, L., Tully, M. A., Barnett, Y., Lopez-Sanchez, G. F., Butler, L., Schuch, F., López-Bueno, R., McDermott, D., Firth, J., Grabovac, I., et al. (2020). The relationship between physical activity and mental health in a sample of the uk public: A cross-sectional study during the implementation of covid-19 social distancing measures. *Mental health and physical activity*, *19*, 100345.

Jenks, G. & of Kansas. Department of Geography, U. (1977). *Optimal Data Classification for Choropleth Maps*. Occasional paper. University of Kansas.

Little, R. J. & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual*

review of public health, *21*(1), 121–145.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loureiro, D., Barbieri, F., Neves, L., Anke, L. E., & Camacho-Collados, J. (2022). Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Maugeri, G., Castrogiovanni, P., Battaglia, G., Pippi, R., D'Agata, V., Palma, A., Di Rosa, M., & Musumeci, G. (2020). The impact of physical activity on psychological health during covid-19 pandemic in italy. *Heliyon*, *6*(6).

Meevissen, Y. M., Peters, M. L., & Alberts, H. J. (2011). Become more optimistic by imagining a best possible self: Effects of a two week intervention. *Journal of behavior therapy and experimental psychiatry*, *42*(3), 371–378.

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? *Advances in neural information processing systems*, *32*.

Nienhuis, C. P. & Lesser, I. A. (2020). The impact of covid-19 on women's physical activity behavior and mental well-being. *International journal of environmental research and public health*, *17*(23), 9036.

Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., & Gama, J. (2022). Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *12*(2), e1449.

Nyström, M. B., Neely, G., Hassmén, P., & Carlbring, P. (2015). Treating major depression with physical activity: a systematic overview with recommendations. *Cognitive behaviour therapy*, *44*(4), 341–352.

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the international AAAI conference on web and social media*, volume 4, (pp. 122–129).

Pellert, M., Metzler, H., Matzenberger, M., & Garcia, D. (2022). Validating daily social media macroscopes of emotions. *Scientific reports*, *12*(1), 11236.

Perrin, A. (2015). Social media usage: 2005-2015.

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, *7*(1), 13006.

Rosenbaum, P. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Schuch, F. B., Vancampfort, D., Richards, J., Rosenbaum, S., Ward, P. B., & Stubbs, B. (2016). Exercise as a treatment for depression: a meta-analysis adjusting for publication bias. *Journal of psychiatric research*, *77*, 42–51.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791.

Sianesi, B. (2001). Implementing propensity score matching estimators with stata.

In *UK Stata Users Group, VII Meeting*, (pp. 1–15).

Sims, J., Smith, F., Duffy, A., & Hilton, S. (1999). The vagaries of self-reports of physical activity: a problem revisited and addressed in a study of exercise promotion in the over 65s in general practice. *Family Practice*, *16*(2), 152–157.

Stanton, R. & Reaburn, P. (2014). Exercise and the treatment of depression: A review of the exercise program variables. *Journal of Science and Medicine in Sport*, *17*(2), 177–182.

Sun, Y. & et al. (2019). Ernie: Enhanced language representation with informative entities. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Thavaneswaran, A. & Lix, L. (2008). Propensity score matching in observational studies. *Canada: University of Manitoba*.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proceedings of the International AAAI Conference on Web and Social Media*, *4*(1), 178–185.

Vaswani, A. & et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Vickey, T. A., Ginis, K. M., Dabrowski, M., & Breslin, J. G. (2013). Twitter classification model: the abc of two million fitness tweets. *Translational behavioral medicine*, *3*(3), 304–311.

Zhang, Z., Chen, B., & Chen, W. (2021). The mediating effect of perceived health on the relationship between physical activity and subjective well-being in chinese college students. *Journal of American College Health*, *69*(1), 9–16.

APPENDIX A

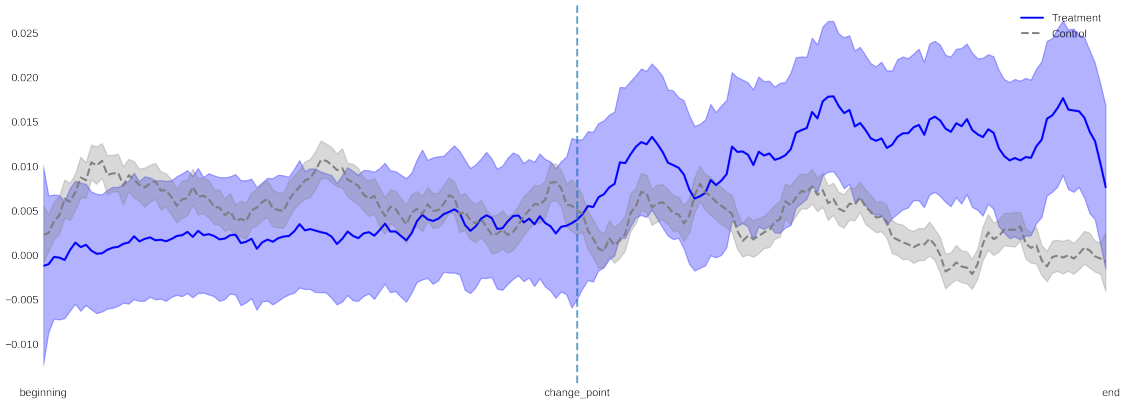


Figure A.1 Time series comparisons with 100 buckets

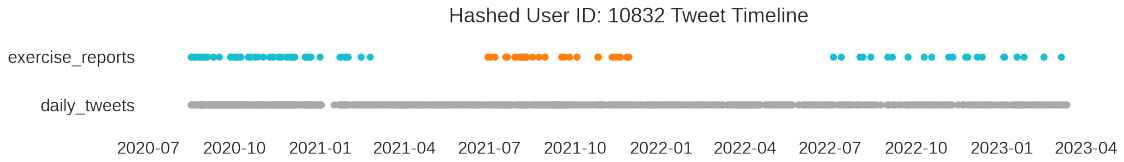


Figure A.2 Sample user tweet timeline

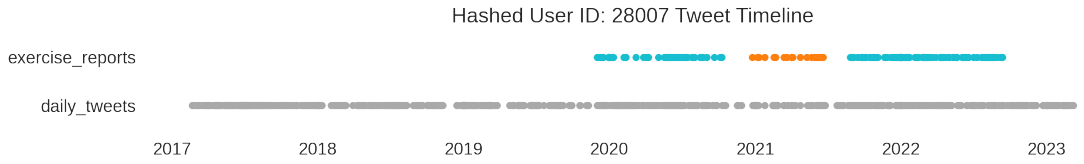


Figure A.3 Sample user tweet timeline

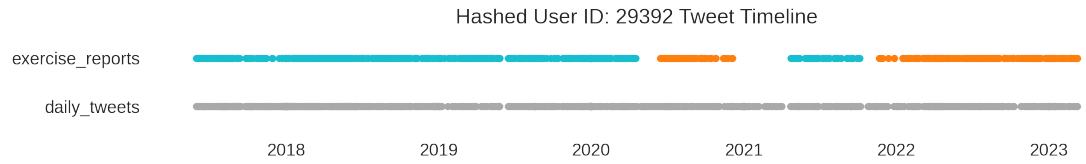


Figure A.4 Sample user tweet timeline

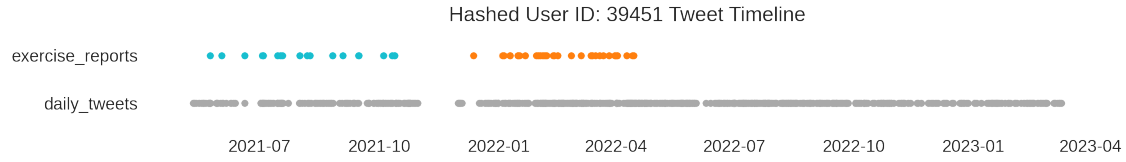


Figure A.5 Sample user tweet timeline

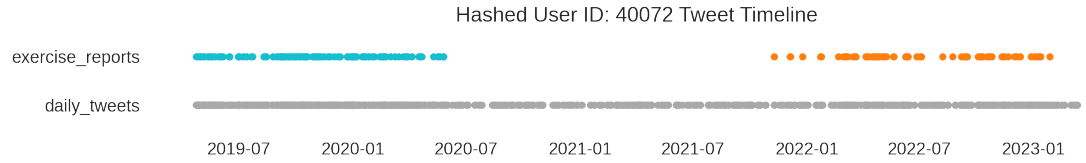


Figure A.6 Sample user tweet timeline

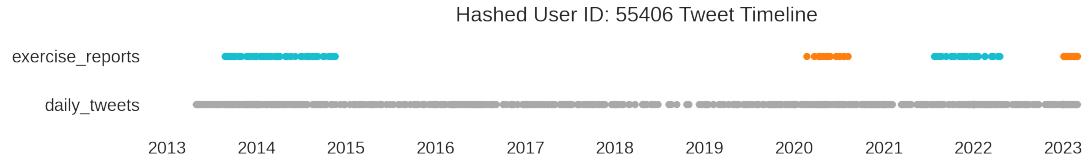


Figure A.7 Sample user tweet timeline

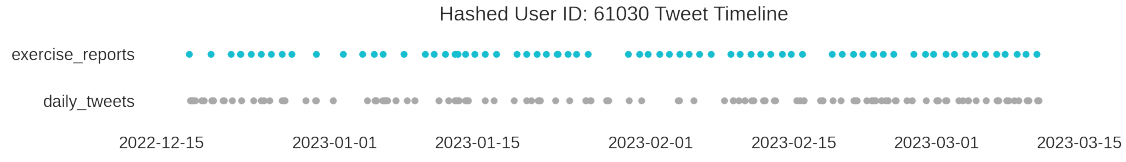


Figure A.8 Sample user tweet timeline

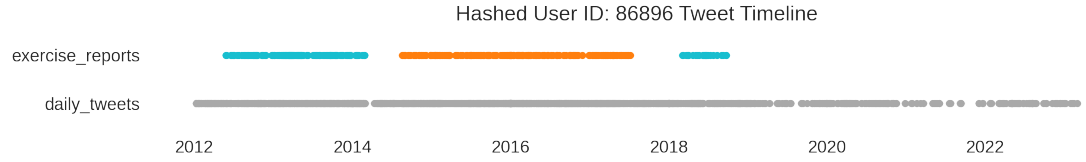


Figure A.9 Sample user tweet timeline

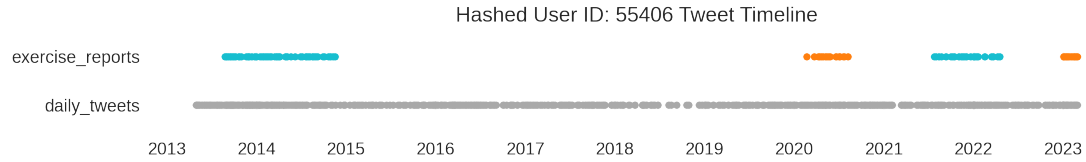


Figure A.10 Sample user tweet timeline

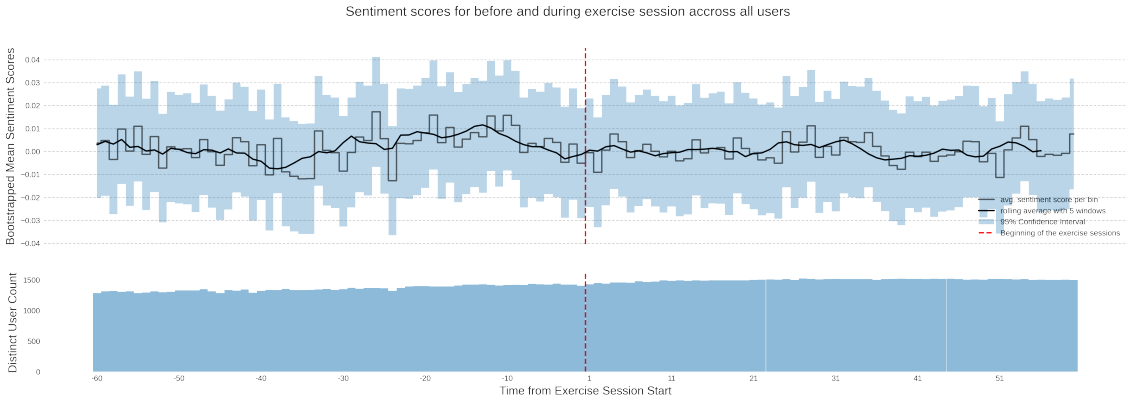


Figure A.11 Without bucketization sentiment score changes at exercise start periods

Figure A.12 Without bucketization sentiment score changes at exercise end changepoint

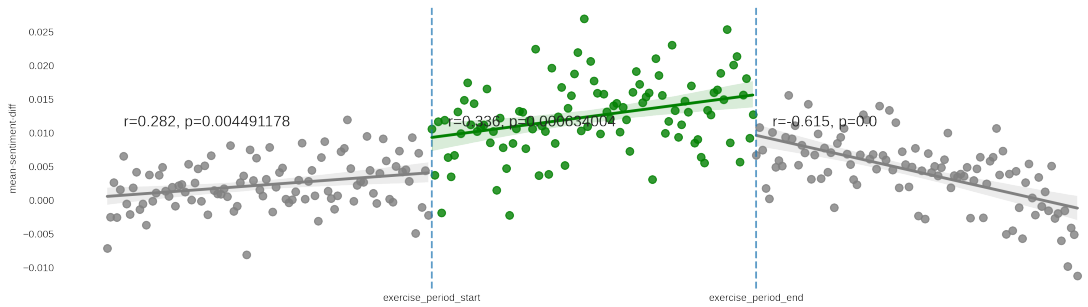


Figure A.13 Exercise before-during and after sentiment scores aggregated with 100 buckets.

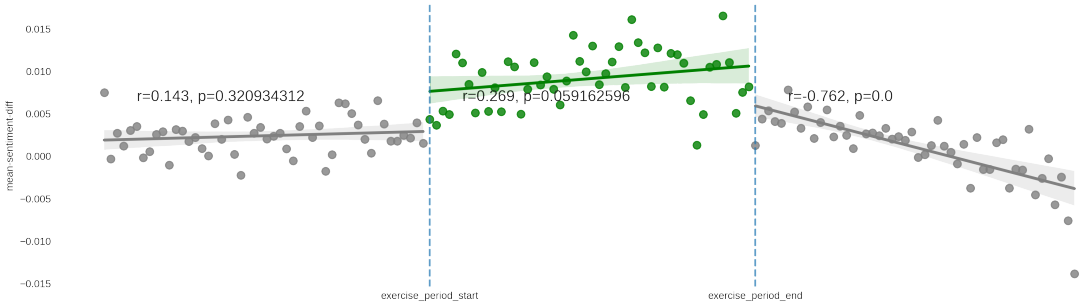


Figure A.14 Exercise before-during and after sentiment scores aggregated with 50 buckets.

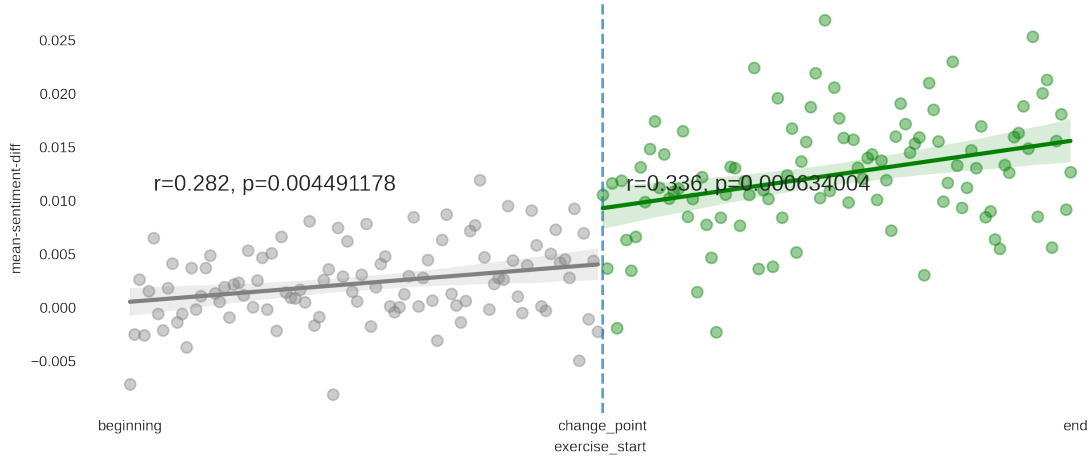


Figure A.15 Scatter plot of exercise before and during for treatment with 100 buckets.

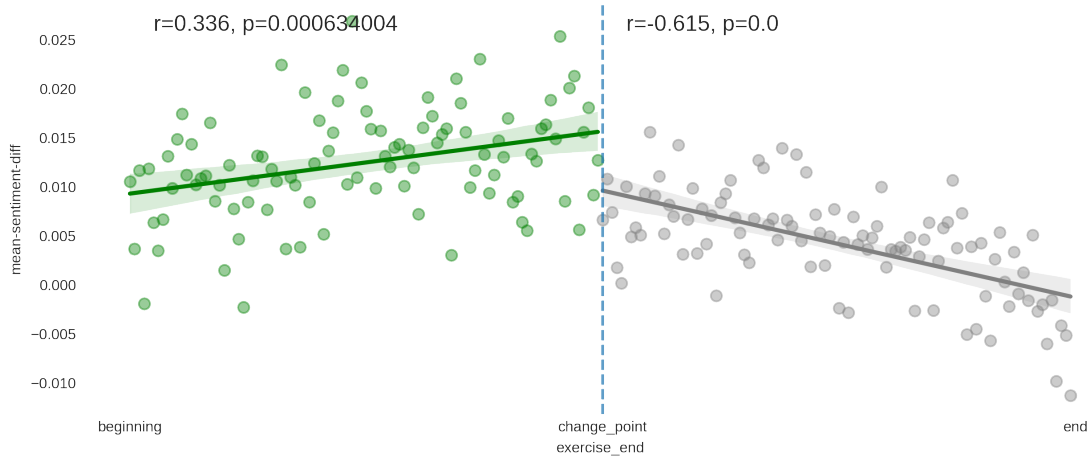


Figure A.16 Scatter plot of exercise during and after for treatment with 100 buckets.

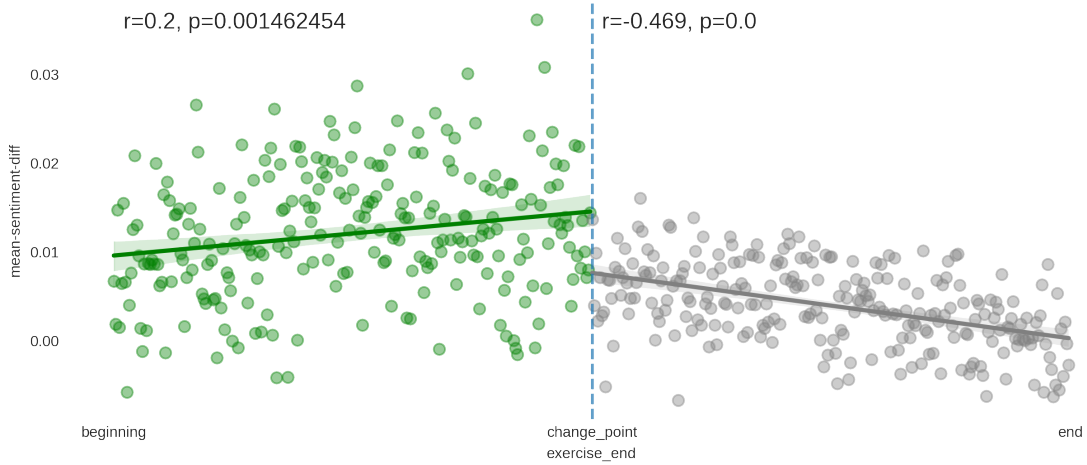


Figure A.17 Scatter plot of exercise during and after for treatment with 250 buckets.

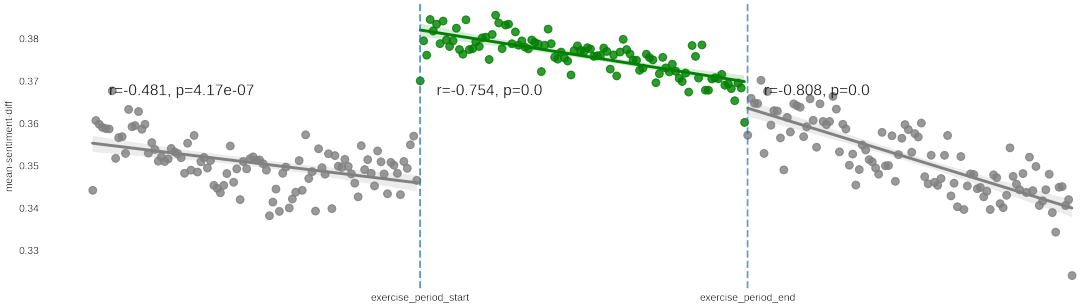


Figure A.18 Scatter plot of standard deviation of exercise during and after for treatment with 100 buckets.

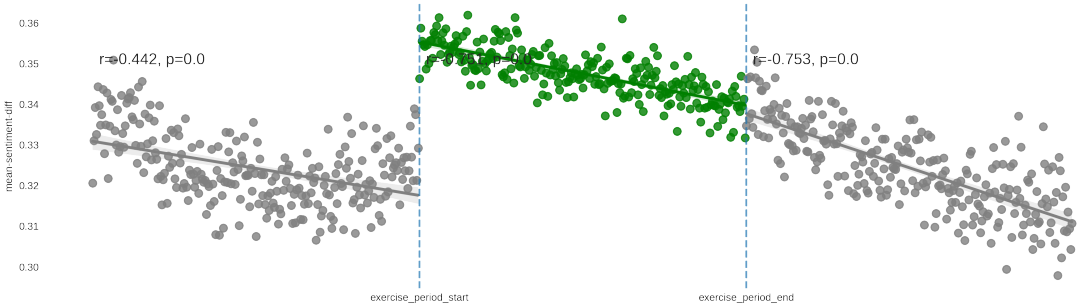


Figure A.19 Scatter plot of standard deviation of exercise during and after for treatment with 250 buckets.

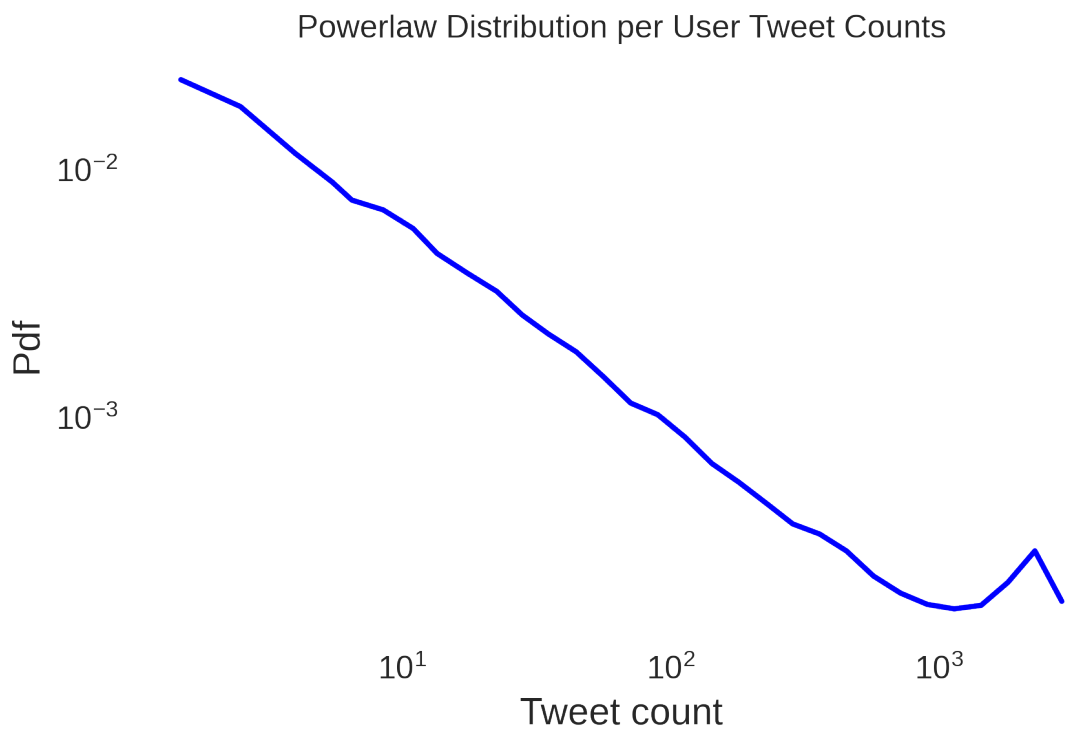


Figure A.20 Number of tweets per user powerlaw distribution. Peak at the end is due to the API limits that we can get at most 3200 tweets from a single user.

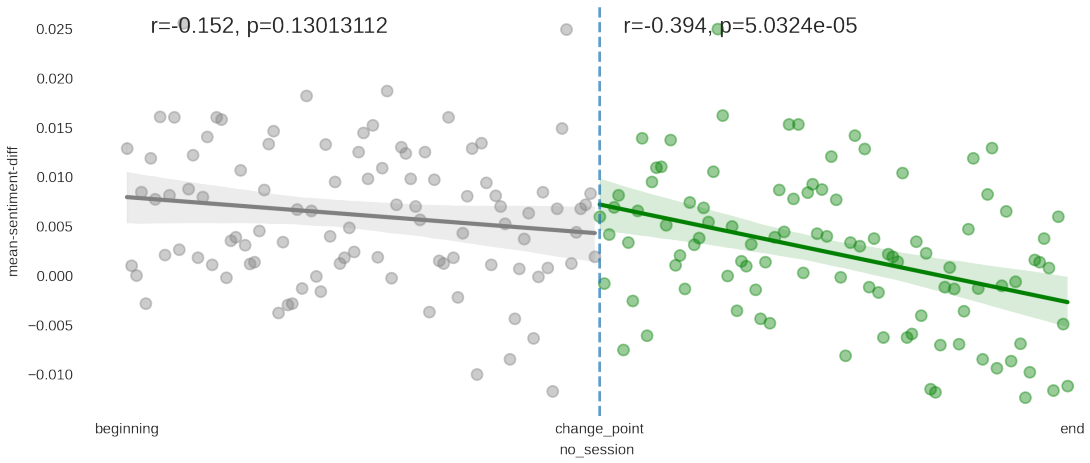


Figure A.21 Scatter plot of sentiment scores for control group with 100 buckets.

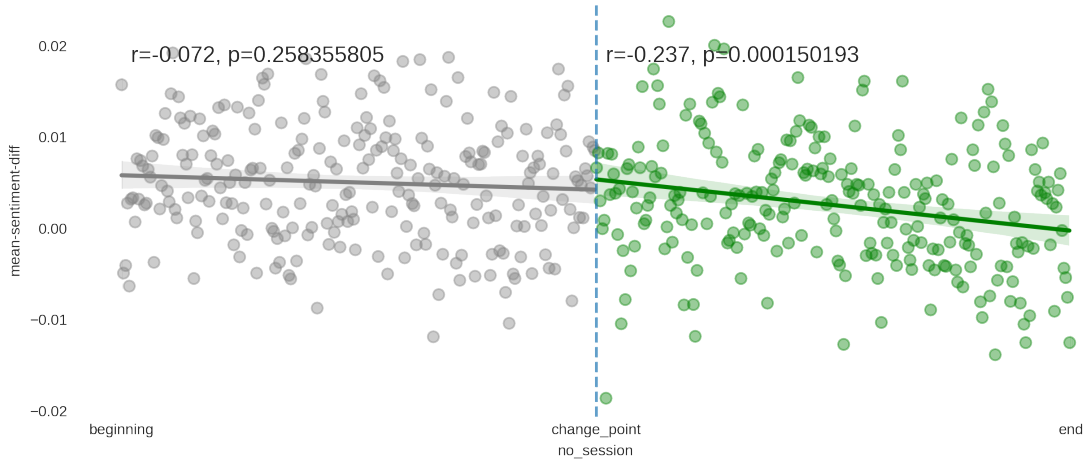


Figure A.22 Scatter plot of sentiment scores for control group with 250 buckets.

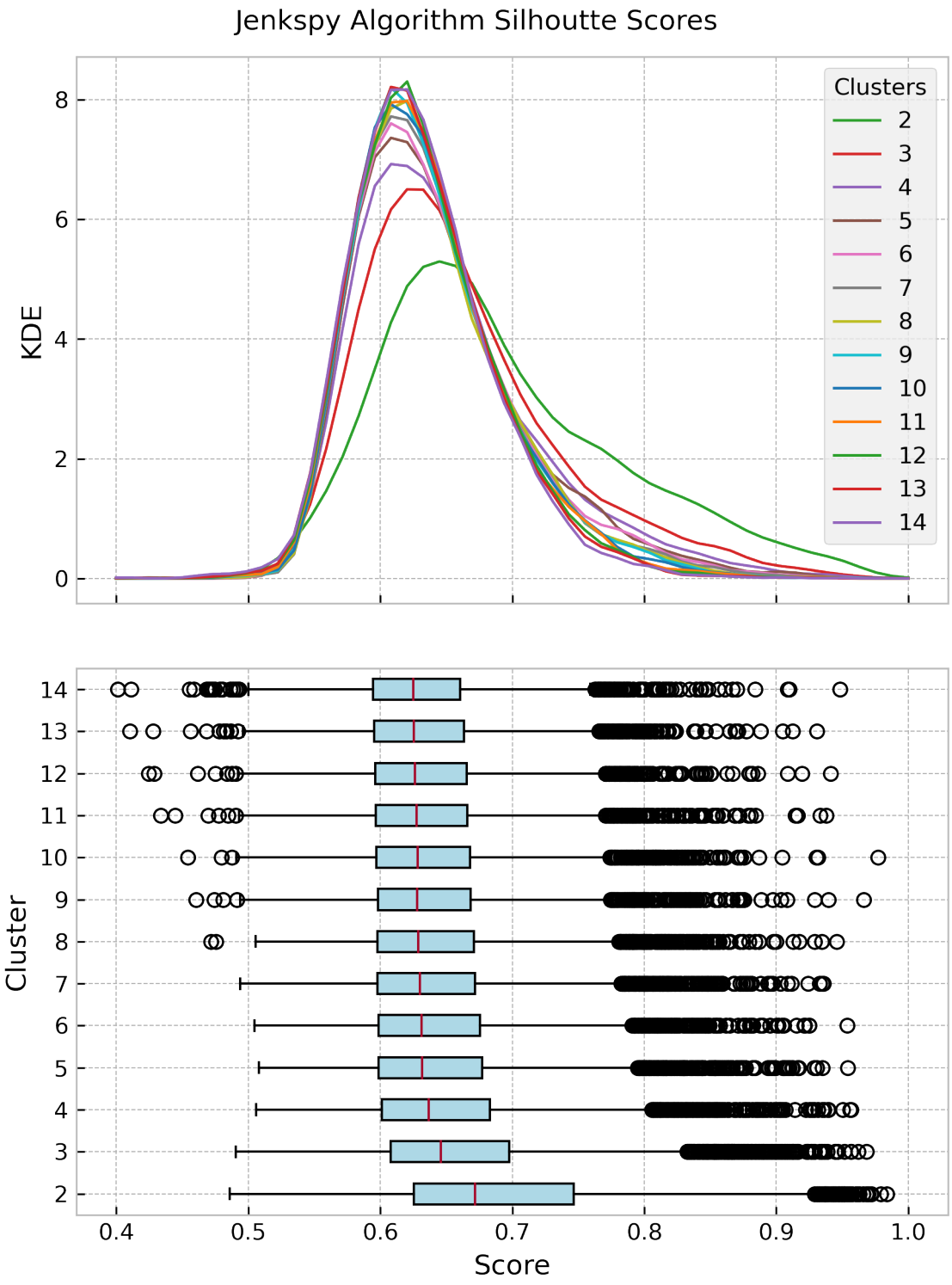


Figure A.23 Jenkspy Clustering for session detection silhoutte scores

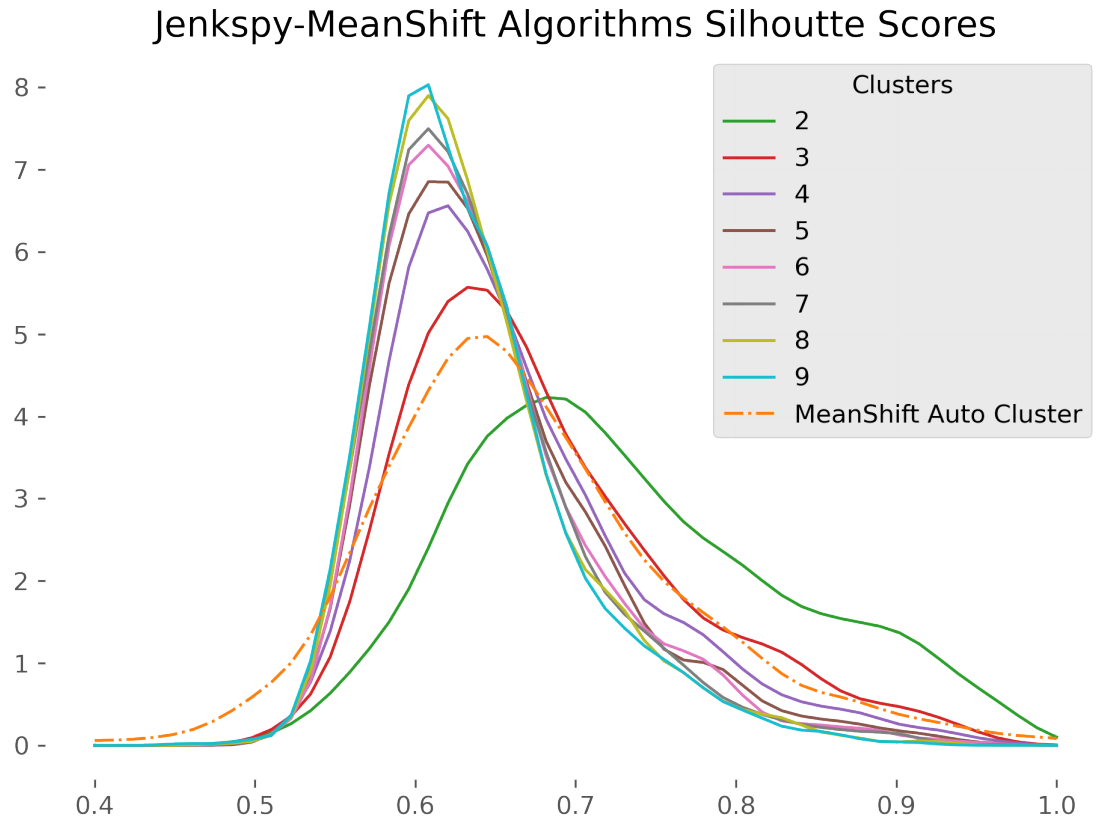


Figure A.24 Session Clustering Silhoutte Scores

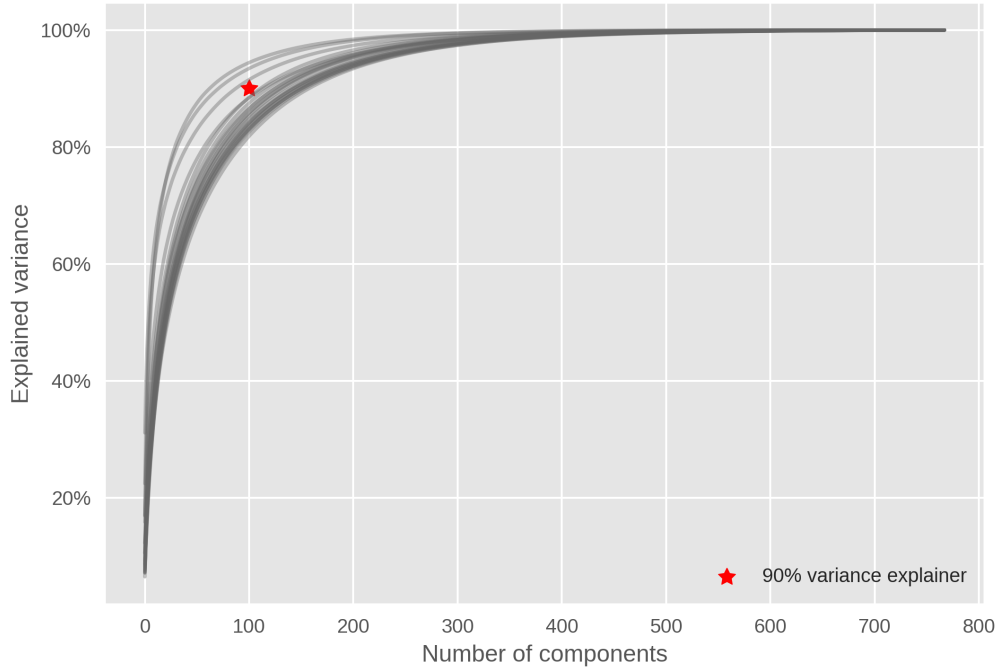


Figure A.25 User historical tweets aggregated to get the user based embeddings with a pretrained BERT model. Then we applied PCA to reduce model output vector length while preserving the variance. Original output had vector length at 720 then PCA reduced them to 100 while preserving the variance pointed with the red star.

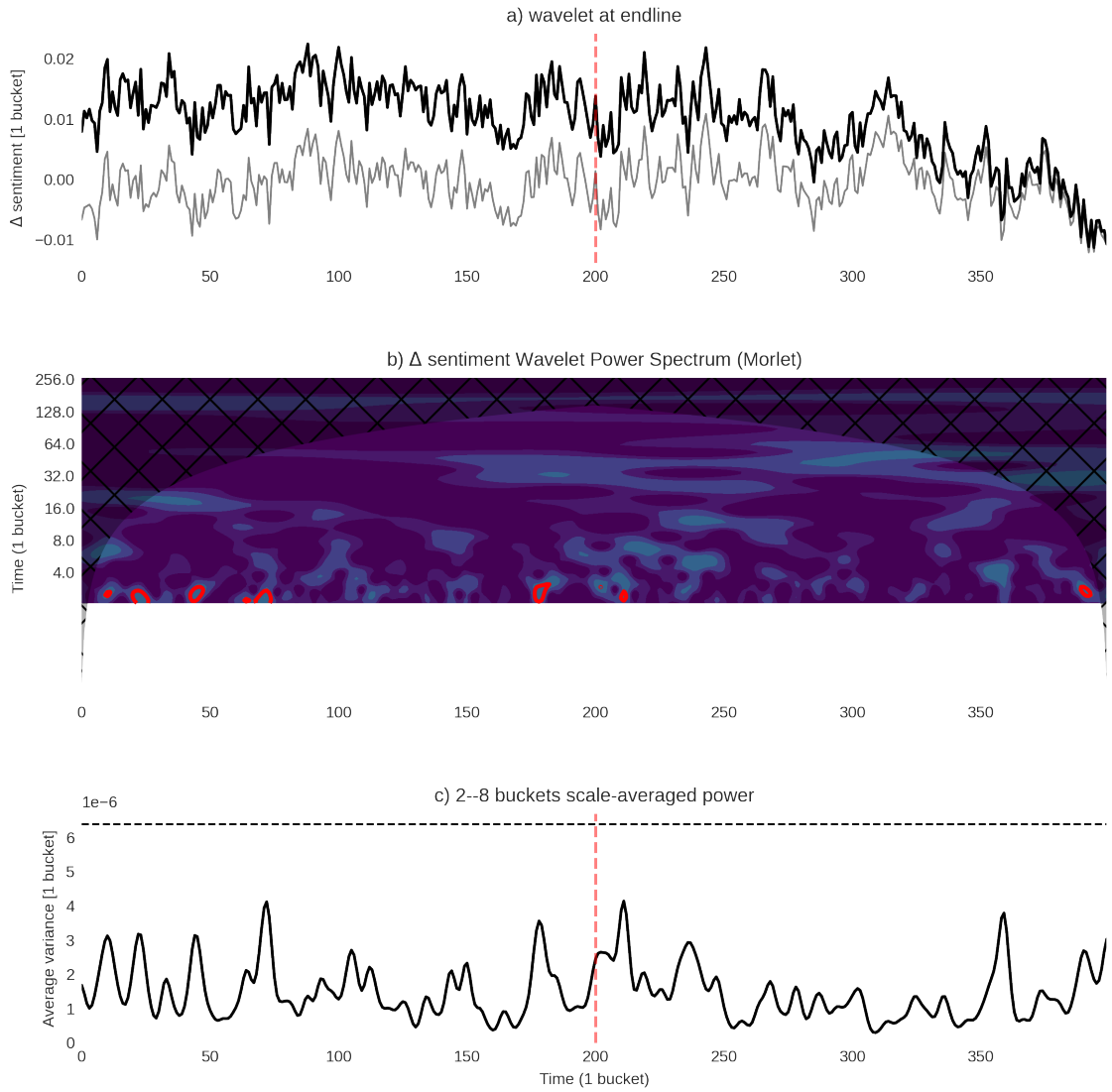


Figure A.26 Spectral analysis of control users at exercise end change-point with 250 buckets: (a) Time- series (solid black line) and inverse wavelet transform (solid grey line), (b) Normalized wavelet power spectrum of the Δ scores using the Morlet wavelet ($\omega = 6$) as a function of time and of Fourier equivalent wave period.(d) Scale-averaged wavelet power over the 2–8 buckets band (solid black line) and the 95% confidence level (black dotted line).