Quantification of change in online activity patterns through the course of the COVID-19 pandemic

Rico Pircklén

School of Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 22.11.2021

Supervisor

DSc Talayeh Aledavood

Advisor

DSc Talayeh Aledavood
Postdoc Ted Chen

Aalto University
School of Engineering

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of the master's thesis

Author Rico Pircklén

Title Quantification of change in online activity patterns through the course of the COVID-19 pandemic

Degree programme Mechanical engineering

Major Mechanical engineering **Code of major** ENG25

Supervisor DSc Talayeh Aledavood

Advisor DSc Talayeh Aledavood
Postdoc Ted Chen

Date 22.11.2021 **Number of pages** 51 **Language** English

Abstract

Human activities follow temporal patterns such as daily, weekly, and seasonal rhythms that are related to physiology, natural cycles, and social constructs. One such rhythm is the near 24-hour, circadian rhythm. The frequency of this rhythm is somewhat similar among people, but its phase depends on individual.

The availability of social media data has provided a fine-grained view of temporal patterns of activity on both individual and population levels. In this thesis, these digital daily cycles are studied by using an extensive 3-year online activity dataset around the time of the COVID-19 pandemic from over 55 thousand Finnish Twitter users, as well as another 2-year dataset of over 15 thousand users from Houston area around the time of Hurricane Harvey crisis. The main focus of this thesis is to quantify changes in the temporal patterns during the crisis events in order to better predict the effect of crises on public health, as well as on the behaviour of individuals.

We begin by illustrating differences between average daily rhythms on population level as well as showing that there is a strong individual-level variation beyond averages before moving on to investigating changes in these patterns through the courses of the crises. To extract interpretable activity patterns, an unsupervised method called non-negative matrix decomposition (NMF) is used.

The population level activity patterns in Finland experienced significant changes between the time of the pandemic and the year before it. These changes had a moderate correlation of 0.61 with the overall COVID-19 government response index that follows the level of measures Finnish government had taken regarding the pandemic. It was also shown how during the first five weeks of the lockdown period, major changes in activity patterns were experienced before returning to the normal. Additionally, significant changes on individual level were demonstrated. While the effects of Hurricane Harvey were clearly visible from the increased usage of Twitter in general, no changes in population level activity patterns were detected. The achieved results provide optimism towards the possibility of better understanding the effects of crises on population and individuals through social media activity patterns.

Keywords COVID-19, temporal data, social media, public health , non-negative matrix factorization

Aalto University
School of Engineering

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Rico Pircklén	
Työn nimi Sosiaalisen median käytön muutosten kvantifiointi COVID-19-pandemian aikana	
Koulutusohjelma Konetekniikka	
Pääaine Konetekniikka	**Pääaineen koodi** ENG25
Työn valvoja TkT Talayeh Aledavood	
Työn ohjaaja TkT Talayeh Aledavood Postdoc Ted Chen	
Päivämäärä 22.11.2021 **Sivumäärä** 51	**Kieli** Englanti

Tiivistelmä

Ihmisen toiminta noudattaa toistuvia, ajallisia rytmejä, jotka liittyvät fysiologiaan, luonnollisiin sykleihin ja sosiaalisiin rakenteisiin. Yksi tällainen rytmi on noin 24 tunnin vuorokausirytmi, joka on tavanomaisesti taajuudeltaan yhtäläinen eri ihmisten välillä, mutta johon liittyy yksilöllisiä vaihe-eroja.

Sosiaalisen median datamassat tarjoavat korkearesoluutioinen kuvan yksilöiden ja väestön ajallisesta toiminnasta. Tässä opinnäytetyössä näitä digitaalisia syklejä tutkitaan käyttämällä laajaa 3 vuoden data-aineistoa COVID-19-pandemiasta yli 55 tuhannelta suomalaiselta Twitter-käyttäjältä sekä toista 2 vuoden aineistoa yli 15 tuhannesta Houstonin alueen asukkasta hurrikaani Harveyn aikoihin. Tämä tutkimus keskittyy kvantifioimaan kriisitapahtumien aiheuttamia muutoksia ihmisten ajallisessa toiminnassa kyetäksemme paremmin ennustamaan kriisien vaikutuksia kansanterveyteen sekä yksilöiden käyttäytymiseen.

Aloitamme tutkimuksen havainnollistamalla keskimääräisten päivärytmien eroja väestötasolla, sekä osoittamalla, että yksilötasolla voidaan nähdä keskiarvoista merkittävästi poikkeavaa vaihtelua. Tämän jälkeen siirrymme tutkimaan kriisien vaikutusta näihin ajallisiin toimintarytmeihin.

Tutkimuksen perusteella Suomen väestötason aktiivisuusrytmit kokivat merkittäviä muutoksia pandemian ajan ja sitä edeltävän vuoden välillä. Näillä muutoksilla oli kohtalainen, 0,61:n korrelaatio COVID-19-rajoituksia mittaavan indeksin kanssa, joka seuraa Suomen hallituksen pandemian suhteen toteuttamien toimenpiteiden tasoa. Lisäksi opinnäytetyössä osoitettiin, kuinka maaliskuun 2020 rajoitusten alkaessa viiden ensimmäisen viikon aikana koettiin merkittäviä muutoksia aktiivisuusrytmeissä ennen palautumista normaaliin. Vastaavia muutoksia havaittiin myös yksilötasoilla. Vaikka hurrikaani Harveyn vaikutukset olivat nähtävissä Twitterin lisääntyneen käytön perusteella, väestötason aktiivisuusrytmeissä ei havaittu muutoksia. Opinnäytetyön tulokset toimivat perustana mahdollisuudelle ymmärtää tulevaisuudessa paremmin kriisien vaikutuksia väestöön ja yksilöihin sosiaalisen median aktiivisuusrytmien pohjalta.

Avainsanat COVID-19, aikasarja, sosiaalinen media, kansanterveys

Preface

A journey is coming to an end soon. It has been a long one that started over a decade ago as a computer science -freshman, eventually taking me to mechanical engineering, and abroad to Russia, Sweden, Thailand, Germany, and Slovenia. Now, eleven years later, I've found myself back in my roots, in Finland, finalizing this thesis on data science. I would not have come this far without support, and I have spent much time on thinking about all the people that I've met during these years.

First of all, I would like to thank my family, friends, and colleagues for supporting me, and trusting in me. I am forever grateful for having you all in my life and for sharing this epic adventure with me. I also want to thank Seri and Cora for being good girls - at least most of the time.

A big thanks goes to DSc. Talayeh Aledavood for providing the topic, and making it possible for me to write this thesis in the field that I love so much. I also want to thank Talayeh for her guidance and supervision on this work. Lastly, I would like to acknowledge Ph.D. Ted Chen for providing the invaluable hurricane Harvey dataset for this thesis, as well as the countless ideas and advice on the research work.

Although this journey is coming to an end, my childlike curiosity on science is still there, stronger than ever, looking for the next adventure.

Helsinki, 22.11.2021

Rico Pircklén

Contents

Symbols and abbreviations

Symbols

H Unknown parameter matrix in NMF
V Non-negative data matrix in NMF

Abbreviations

COVID-19	coronavirus disease of 2019
CSS	computational social science
KL	Kullback–Leibler
EMD	The Earth mover's distance
ICA	independent component analysis
IoT	Internet of Things
MF	matrix factorization
NMF	non-negative matrix factorization
OxCGRT	Oxford COVID-19 Government Response Tracker
PCA	principal component analysis
PTSD	post-traumatic stress disorder
RMSE	root mean squared error
RSS	residual sum of squares
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SVD	singular value decomposition
WHO	The World Health Organization

1 Introduction

Human activities follow temporal patterns such as daily, weekly, and seasonal rhythms that are related to physiology, natural cycles, and social constructs. Circadian rhythms are 24-hour cycles that are part of the body's internal clock, and known to be important drivers of human activity. One of the most important and well-known circadian rhythms is the sleep-wake cycle: most people function with the diurnal cycle, sleeping at night and being active through the day. However, within the day-night pattern, there are individual differences with people broadly categorised into morning-type, evening-type, and intermediate-type groups called chronotypes. [3]

Social networking services are increasingly integrated into individuals' daily lives, thus providing a fine-grained view of temporal patterns of activity on a large scale. In this study, these digital daily cycles are studied by using an extensive 3-year online activity dataset around the time of the COVID-19 pandemic from 55 thousand Finnish Twitter users, as well as another 2-year dataset of 16 thousand users from the Greater Houston area around the time of Hurricane Harvey crisis. Specifically, we use tweeting patterns as a measure of individual- and population-level behaviour. We expect that disruptions such as physical life events in individuals' lives will demonstrate abnormalities in their social media behaviour.

The main objective of this thesis is to analyze if social media data can be used to understand individual and population level temporal activity patterns. Part of this objective is to investigate how crisis events, such as the COVID-19 pandemic and Hurricane Harvey, affect these rhythms. Detecting and quantifying changes in temporal patterns might help us to better predict the effect of crises on public health, as well as on the behaviour of individuals. While being out of scope for this study, this research opens also a possibility for future continuation on how different crises affect different population sub groups, such as different genders. Therefore, this thesis works as a foundation for future research towards understanding the effects of crises on population and individuals through social media activity patterns.

This thesis will be structured in a way that allows us to explore online activity patterns and their changes in both population and individual level for two different crises. This is, however, not a trivial task, and requires cross disciplinary understanding on both temporal rhythms and online behaviour of human beings, as well as the data science and methods behind analyzing hundreds of millions of messages and their time series aspects. For this reason, we first include a general chapter on the theoretical background of these topics and the existing literature related to them, followed by a chapter representing the research materials and methods. In the fourth chapter, we use the materials and methods to detect and quantify changes in the online activity patterns during crisis periods. In the final chapter, these achieved results are discussed. In more detail:

Chapter 2 begins with discussions on human behaviour on social media, focus being on Twitter and its usage. After this, a theoretical background to the study of temporal patterns of activity, including time series analysis and the temporal rhythms of human beings is provided. Lastly, both investigated crises, the COVID-19 pandemic in Finland, as well as Hurricane Harvey in the USA, are presented, and their effects on the population are addressed.

The next Chapter 3 is devoted to the research materials and methods. First, the data used for this study, as well as its collection and processing are described in detail. This is followed by an exploratory data analysis. To do so, we analyse user activities and gender, in order to better understand the potential limitations, biases, and possibilities the data provides. Lastly, we will focus on the theoretical background of the used methods and metrics, such as non negative matrix factorization (NMF), and Pearson correlation coefficient.

Then, Chapter 4 presents the achieved results. We structure this chapter into four main parts. For the first part, we will focus on the optimization and robustness testing of NMF. Then, we inspect differences among the temporal patterns of individuals. To do so, we utilize the components derived from an NMF model. After this, we quantify the effects of the COVID-19 pandemic on the activity patterns of the population. Lastly, the results of similar analyses on the Hurricane Harvey case are shown.

Finally, in Chapter 5, we discuss the results, and present concluding remarks, including an outlook of further research directions.

2 Background

This chapter begins by discussing the theoretical background of the human behaviour on social media. This includes the description of digital footprints and their forming, as well as the general usage of social media networks, with a focus on the Twitter usage in Finland. After this, in Section 2.2, a brief overview of time series as a temporal data structure is given. In Section 2.3, the COVID-19 pandemic and its effects on human behaviour are explored. Closer look will be taken on the pandemics effects on Finland and its population. Lastly, Hurricane Harvey that hit the Houston area in summer 2017, is presented.

2.1 Human behaviour on social media

Uncovering the patterns that characterize human behavior is of outstanding importance for understanding phenomena related to public health. Since such phenomena are driven by people, they are closely related to the way in which individuals are connected and interact with each other. [32]

Originally, the term social networks referred to graphs that try to capture the concept that individuals are embedded in webs of social relations and interactions [11]. Typically, in such graphs individuals are represented as nodes that are joined together by links or ties. Nowadays, however, social networks may be more commonly seen as a synonym for social media platforms which individuals use for social networking - to stay connected with friends, family, and peers.

We begin this section by describing how social networks and the Internet of Things (IoT) have led to the concept of big data and the field of computational social science (CSS) - the field that this thesis also heavily relies on. After this, an overview of social media and its usage is presented, focusing especially on Twitter usage in Finland.

2.1.1 Digital footprints on social networks

During the past decades, the emergence of social media, mobile phones, and other devices that automatically store digital records have led to the concept of big data. In this era of the Internet of Things, more and more devices are being connected to the internet to gather data. When people interact with these devices and services, they leave digital footprints - traces that can be used to understand individuals' behavioural patterns - both for scientific and commercial purposes. The effective use of big data has been recognized by many organizations as key to gaining a competitive advantage and outperforming peers [37]. Simultaneously, the scientific community has recognized its potential. For these reasons, new computational tools and methods are constantly being created to extract information from these large datasets.

The multi-disciplinary field of computational social science is a growing field which tackles questions within social sciences with the use of computational methods and big data [1]. The used data types vary from natural language such as speech and text to communication timestamps, and sociodemographic data, among others. In this thesis, datasets of tweeting timestamps have been analyzed.

In the past, many studies have focused on the structural properties of social networks, namely the topology and the formation processes of social networks, network sizes, and tie strengths ([21]). Such studies typically consider networks static and end up neglecting the temporal dimension of human activity. Lately, research on the dynamics and evolution of social networks, such as the formation and decay of connections have also been studied ([32]). Some of the research has also focused on the temporal aspect of the communication patterns of individuals, and their relation to human behaviour ([3], [4], [28], [55], [56]). Among others, a strong connection between the chronotypes of people and the structure of social networks that they form has been found [5].

2.1.2 What are social networks used for?

Social media platforms have become a great success, providing users a platform to express their thoughts and feelings, and share their life events and experiences freely. In social media, people actively share their thoughts and opinions on all kinds of matters, providing a valuable source of information for a wide range of fields. Social media data has been utilized for a great variety of purposes varying from stock market predictions to mental disorder detection, and crisis studies. While studying such data, it is worth noticing that such research may come with privacy issues and ethical concerns.

Different social media platforms tend to serve different purposes, and as a consequence, have different kind of users, networks, activities, and contents, varying from LinkedIn's professional networks to Facebook's personal space, and Twitter's broad audiences. Karami et al. [29], have studied the dominant topics of Twitter-based research, and managed to extract 40 topics, including tourism, politics, and disaster analysis.

While all kinds of matters are discussed in social media, not all events are discussed equally. Based on a study on Facebook users by Saha et al. [41], it was found that positive and anticipated events are more likely, whereas significant, recent, and intimate events are less likely to be disclosed on social media. The most common events users shared were personal of type, including vacations and trips. Also health-related events were disclosed online, especially for health gains, while financial, work, and local events were seldom discussed. Work-related events, performance reviews, promotions, heavy work, and job switches were rarely shared, potentially due to concerns of employer surveillance. If discussed, events of such type tended to be positive, including topics such as good work life and work success.

In addition to differences between platforms, information can typically be shared in multiple different ways within the platform. Bevan et al. [7] has found that users preferred to share positive life events in Facebook indirectly, for example through photos with no caption, or relationship status changes without context or explanation, whereas negative life events were more likely to be disclosed directly, for example through status updates.

During the past decades, extensive research on social media has been performed by the scientific community. Common topics among social media based research include sentiment analysis, topic modelling, and content analysis. While most studies tend to focus on the content of the social media, this study concentrates on the timestamps: When people are active online. This approach may provide valuable information of individual users' behaviour as well as the effects of physical events, such as the COVID-19 pandemic and Hurricane Harvey, on the population and individuals.

2.1.3 Twitter usage in Finland

According to Statistics Finland [49], 82 percentages of Finns aged from 16 to 89 used the Internet on daily basis in 2020, showing an increase of 3 percentage points from the previous year. In social media platforms, the increase in usage was even more significant: While 61 percentages of Finns used some form of social media in 2019, the number increased by 8 percentage points to 69 percentages in 2020. To illustrate the magnitude of the increase, during the four previous years from 2015 to 2019, the social media usage had increased only 3 percentage points in total. The relative growth was most significant in the oldest age groups: In 65-74-years-olds, the social media usage increased from below 200 000 to over 300 000, and in 75-89-years-olds from 45 000 to over 75 000. It has been estimated that this development was due to the restrictions imposed by the COVID-19 that shifted socializing more to social networks and motivated older people to learn new social media skills.

In 2020, the most commonly used social media platform in Finland was Facebook with 58 % of the population using it. The following platforms were WhatsApp and Instagram with 50 and 39 % usage, respectively. Twitter was used by 13 percentages of the population, putting it in popularity behind platforms such as Youtube and Snapchat, and on the same level with LinkedIn. [49]

In Figure 1, the distribution of Twitter users to different age groups is presented and compared to the distribution of the Finnish population in 2020. It is worth mentioning that the Statistics Finland has not included the age groups below 16 and over 90 in their statistics of Twitter usage. From the figure, it can be seen how the distribution of Twitter users does not represent the Finnish demographics perfectly, having over representation in young adults, especially within the age group of 16-24-years-old, while the age groups over 65-years-old are notably under represented. The situation is similar for different genders: While the binary gender distribution among Finns is close to equal, 16 percentages of Finnish males use Twitter, while the number for females is only 11 percentages [49].

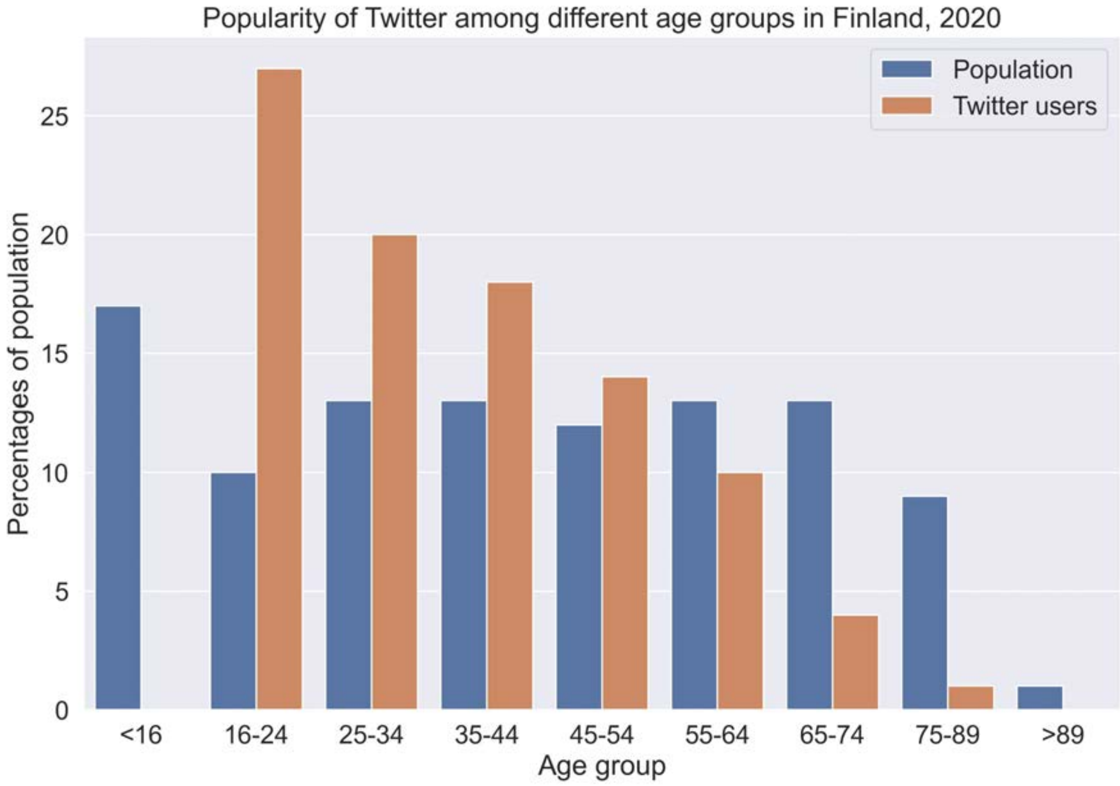

Figure 1: Twitter usage in Finland compared to the country's population distribution for year 2020. [49], [48]

2.2 Temporal patterns of activity

As was mentioned earlier in Section 2.1.1, research on temporal patterns of activity has increased during the past years. In this section, after first defining and describing time series, an overview on the existing research on the topic will be given.

2.2.1 Time series and temporal data

The time series is the most common temporal data structure. It contains information about the evolution of single variable's value over time. While this can be generalized into multi-variate time series where multiple variables are monitored over the same time period, such a case can be consider a collection of individual, 1-dimensional time series. In cases where multiple different events co-occur, an event sequence can be a more suitable data structure than a univariate time series. One of the most important aspects of time series is the natural ordering of the values. [31]

2.2.2 Temporal rhythms of human beings

Human activities follow a set of repeating cycles, namely daily, weekly, and seasonal rhythms. The existence of these rhythms is related to physiology and natural cycles as well as social constructs. The frequency of these rhythms is somewhat similar across people, but its phase is different. Based on the tendency to sleep at different hours of the day, people can be categorized into different groups called chronotypes. Such groups may include e.g. morning-, evening-, and intermediate-types. Traditionally, such categorization is based on questionnaires or manual analysis of timings of people's activity, but recently also more advanced methods have been developed for the task. [3]

One of such methods has been presented by Aledavood et al. [3]. In the study, mobile phone screen usage logs that represent the temporal activity patterns of individuals were collected from 400 university students, and decomposed into components by applying non-negative matrix factorization. In the study, four emergent temporal components were found: morning activity, night activity, evening activity, and activity at noon. Individual behavior was then reduced to weights on these four components. The study did not observe any clear emergent categories of people based on the weights, but individuals were rather placed on a continuous spectrum according to the timings of their activities. The study concludes that such approach allows to categorize people based on their full daily and weekly rhythms of activity.

Understanding chronotypes can be beneficial, because differences in time use among chronotypes have been found. Kauderer and Randler [30] found that morning types spent less time watching television and using the computer, but more time reading and in physical activity. Evening types, on the other hand, spent more time with their friends than intermediate types and morning types. On weekdays, sleep duration was shortest in evening types. Differences between genders were also found in the study: While sleep duration did not differ on weekdays, on the weekends girls tended to sleep longer. It is worth noticing that exogenous factors can play a big role in how an individual behaves and what kind of sleeping patterns they have [3]. Also, differences in cultures, climates, environment, and so on, can make a difference in how these chronotypes are perceived [3].

In [2], Aledavood et al. showed that individuals tend to have persistent daily rhythms for communication across different channels, namely calls and text messages, but the shapes of these rhythms and peak hours differ. This points towards different channels serving a different role in communication. The study also found that the daily patterns of males and females differ both for calls and texts, both in how they communicate with individuals of the same gender vs. opposite gender, and depending on social ties, i.e. kin ties vs. non-kin ties. Therefore, different data streams may provide a bias that should be considered, and perhaps even combined when possible.

The time variation of activity in population and individual level rhythms have been studied extensively using data sources such as Youtube [20], technology-news website Slashdot [28], frequency of edits in Wikipedia [56], OpenStreetMap [55], as well as calls, text messages, and email records [4]. In these studies, periodic patterns of activity have been distinguished on several time scales: The longest scale is that of a calendar year, where special periods such as holidays can typically be distinguished. On a weekly cycle, weekends typically differ from weekdays, and differences between weekdays may also occur. Finally, there is a daily pattern which may significantly differ between different systems.

In Figure 2, an example of an average weekly Twitter user activity pattern in

Finland during 2020 is presented. From the figure, the natural daily cycle can be seen, namely average user sleeping during the nights, and being active during the day time. On daily level, the two clear activity peaks indicate that Twitter is being used more actively in the mornings between 8 and 10 AM, and in the evenings around 8 to 9 PM. While the activity during the weekdays is somewhat similar to each other, the Twitter usage experiences a clear decrease during the weekends. Next, we continue to discuss the COVID-19 pandemic, and what kind of effects it has had on people, including their temporal patterns.

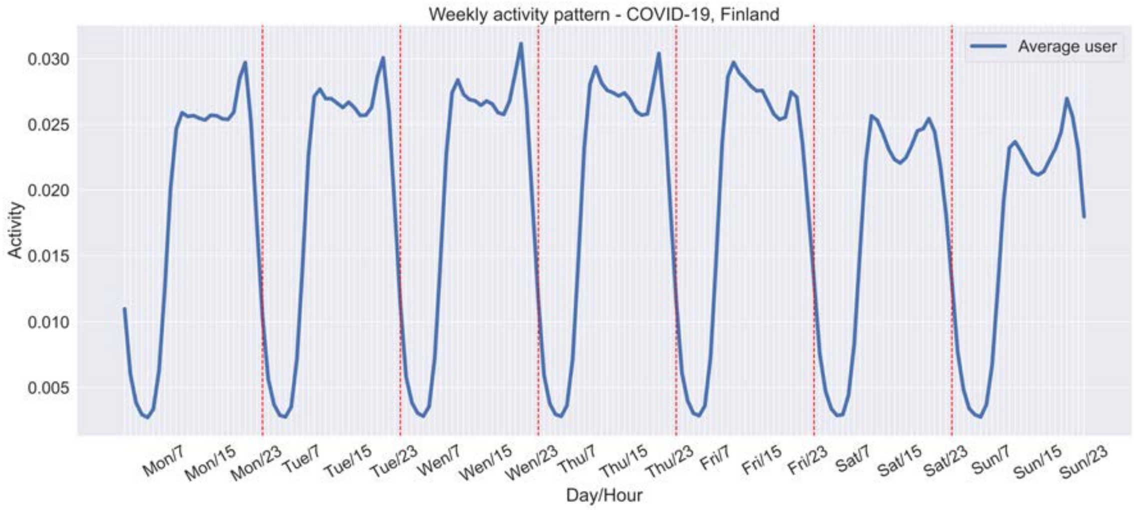

Figure 2: An example of an average weekly Twitter user activity pattern in Finland during 2020.

2.3 The COVID-19 pandemic

The COVID-19 pandemic is a global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The virus was first identified in Wuhan, China, in December 2019, where lockdowns in the city and its surrounding areas were enforced in order to contain the outbreak. However, the containment didn't succeed, and as a result the virus spread to other parts of mainland China and eventually around the world. The World Health Organization (WHO) declared a Public Health Emergency of International Concern on 30 January 2020, and a pandemic on 11 March 2020 [53]. Since 2021, variants of the virus have emerged and become dominant in many countries. By the time of writing this thesis, more than 200 million cases and 4 million deaths have been confirmed, making it one of the deadliest pandemics in history [35].

2.3.1 The pandemic's effects on individuals both globally and in Finland

The COVID-19 pandemic has had a variety of direct and indirect effects on lives all around the globe. First of all, the pandemic resulted in unprecedented changes in the spatial mobility of people across societies due to the restrictions imposed in order to limit the spread of the virus. In Finland, this lead to a significant drop in inter-municipal mobility, with a noticeable shift of population from urban centres to rural areas, which can be explained by secondary housing [54]. At the same time, the closures of national borders limited the movement of transnational people across the Finnish borders [27].

During the pandemic, the time spent at home has increased significantly while the amount of close contacts has decreased as a result of the restrictions. Zhang et al. [59], have found that residents of Hong Kong spent over 30 % more time at home, coming into close contact with 17.6 and 7.1 people per day on average during the normal and pandemic periods, respectively. Students, workers, and older people reduced their daily number of close contacts by 83.0%, 48.1%, and 40.3%, respectively. Similarly, the close contact rates in residences, workplaces, places of study, restaurants, shopping centers, markets, and public transport decreased by 8.3%, 30.8%, 66.0%, 38.5%, 48.6%, 41.0%, and 36.1%, respectively.

In Finland, the time spent at home as a form of remote working has also significantly increased. Before the pandemic in 2019, 14.1% of employed people in Finland usually worked from home, making the country to have the highest proportion of remote working in Europe together with the Netherlands [18]. In May 2020, a national survey of public sector employees showed that about 97% had started or increased the frequency of remote work during the pandemic [10]. In another study, it was found that 36% of local government employees and 72% of state employees worked remotely in 2020 [38].

While the effects of COVID-19 have been somewhat global, also differences in the behaviour of different sub-groups have been detected. By analyzing Austrians' mobile phone data, Reisch et al. [40] managed to conclude that after the lockdown, behavioral gender differences in mobility and communication patterns were increased significantly, while circadian rhythms tended to synchronize. While mobility declined massively for both genders, women tended to restrict their movement stronger than men. After the lockdown, males returned back to normal quicker than women. Similar effect was noticed among young individuals compared to older people. The study found also that the length of a day for both genders was reduced by an hour.

The uncertainties surrounding the pandemic, the precautionary measures such as social distancing and self-quarantining, as well as the societal impacts such as economic downturn and job loss have led to various mental health concerns. Saha et al. [42] studied these effects by performing linguistic analyses and comparing Twitter postings from the COVID period of 2020 to those from 2019, and found that mental health symptomatic expressions had increased by about 14%. However, the study also observed that these expressions gradually lessened over time, suggesting that people adapted to the circumstances. Similarly, Rajkumar [39] found in a review of the existing literature search of the PubMed database that symptoms of anxiety and depression (16-28%), and self-reported stress (8%) were common psychological reactions to the COVID-19 pandemic, and may be associated with disturbed sleep. Also Bueno-Notivol et al. [13] combined community-based studies on depression conducted during the COVID-19, and estimated the pooled prevalence of depression using a random effect model. The study resulted in prevalence of 25% which is 7 times higher than a global estimated prevalence of depression of 3.44% in 2017.

While the themes discussed in social media have experienced a partial shift, the daily patterns of emotions have remained mostly unchanged on population level, thereby suggesting that emotional cycles are resilient to exogenous shocks [15]. However, this may not be the case for the circadian rhythms: A reshaping of these rhythms with an increase of night activity during the lockdown has been observed [15]. Reisch et al. [40] have also found that the length of day had decreased by an hour in Austria during the pandemic.

In addition to the aforementioned consequences, the pandemic has had also various other effects that may have affected the daily activity patterns of individuals. According to Statistics Finland's Labour Force Survey [50], there were 235,000 people without jobs as of January 2021. This is 41,000 more unemployed people than in the same period last year, indicating a steep rise in the country's unemployment rate over the past year. The country's employment rate stood at 69.9 per cent in 2021, compared to 71.9 per cent in 2020.

In a study on the pandemic's effect on alcohol usage in Finland, Oksanen et al. [36] found that one-fourth of Finnish workers reported increased drinking during the pandemic. In the sampling of 1308 Finnish workes, half of the test samples reported no change in alcohol consumption, while around 27% had decreased their alcohol intake. The increased drinking was most common among workers under 30 years of age.

2.3.2 COVID-19 restrictions in Finland

The first COVID-19 case in Europe was confirmed in France, on January 24th, 2020. By mid-March all European countries had a confirmed case. On 16 March 2020, the Finnish Government together with the President of Finland decided on additional measures to address the coronavirus outbreak in Finland, and declared a state of emergency in the country. This included the closure of most schools and government run public facilities, the limitation of group sizes in public, and preparations for the shutdown of borders, among others. [33]

Travel restrictions between Uusimaa region and the rest of the country were activated on 28 March, and lifted on 15 April. The state of emergency lasted until 16 June, 2020. Throughout the summer and autumn, restrictions, e.g. those related to restaurants and sports, were lifted one after another, while towards the winter, new regional restrictions were activated again in regions with high number of infections. [33]

In Figure 3, Oxford COVID-19 Government Response Tracker (OxCGRT) values for Finland in 2020 are presented. The OxCGRT collects publicly available information on which governments have taken which measures, and when, in order to help decision-makers and citizens understand governmental responses in a consistent way.

The information is based on 23 indicators related to four fields: 1) containment and closure policies, such as school closures and restrictions in movement; 2) economic policies, such as income support to citizens or provision of foreign aid; 3) health system policies, such as the COVID-19 testing regime; and 4) vaccine policies, such as cost of vaccination to the individual, and prioritisation lists. [9]

Based on the indicators, the OxCGRT has produced four indices that aggregate the data into a single number. Each of these indices report a number between 0 to 100 that reflects the level of the governments response along certain dimensions. The indices are following: 1) overall government response index, which contains all indicators; 2) containment and health index that combines lockdown restrictions and with measures such as testing policy and contact tracing; 3) stringency index that describes the lockdown strictness ; and 4) economic support index that utilizes the economic indicators and measures income support and debt relief, among others. It is worth noticing that these indices cannot say whether a government's policy has been implemented effectively. [9]

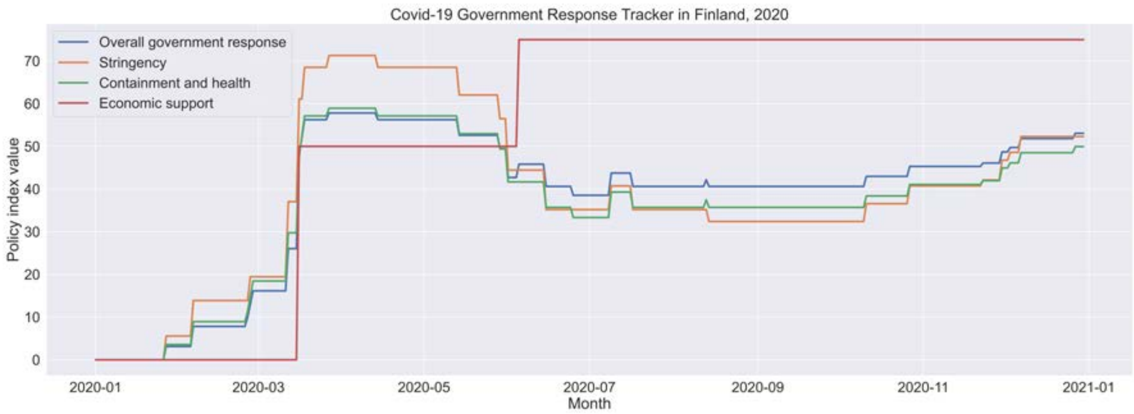

Figure 3: Government response tracker indices in Finland, 2020. Data provided by [22].

2.4 The crisis of Hurricane Harvey

The second case study of this thesis is Hurricane Harvey, which brought unprecedented levels of flooding, property damage, and displacement to the greater Houston area in late summer 2017. Such an event allows us to study pre- and post disaster behaviors of affected individuals. While COVID-19 pandemic has been a slowly developing crisis, Hurricane Harvey was a shock event that lasted only for a few weeks. Although Harvey officially became a hurricane on August 17, its influences, which were not felt in the Houston area until August 25. By 3rd of September, the event was over. [16]

While crossing over Southeast Texas, the hurricane poured a record-breaking 1.5 metres of rain, with the Houston metropolitan statistical area receiving between 0.9 and 1.2 metres, resulting in catastrophic flooding. In total, Harvey directly caused 63 storm-related deaths, over 40 000 displaced individuals, and an estimated $125 billion in damages [8]. Additionally, public health literature has found relationship between exposure to disasters, such as Hurricane Katrina in 2005 and Super Storm Sandy in 2012, and long term mental health outcomes such as anxiety-mood disorders, post-traumatic stress disorder (PTSD) and depression [25][47]. This makes Hurricane Harvey a suitable case for exploring the impact of disaster-induced displacement on human behaviour and their temporal rhythms.

3 Research material and methods

In this chapter, the research material and the used methods are presented. We begin this chapter by describing the data, and its collection and processing in detail. After this, in Section 3.3, the collected data is visualized and analyzed in order to provide understanding of its properties. This contains a comparison of the datasets used in this thesis, with a deeper focus on the users, especially their activities. Additionally, a gender analysis is performed. Lastly, in Section 3.4, the theoretical background of the used methods are presented, and the metrics used for the analysis of the results are described.

3.1 The data and its collection

For this study, a dataset of 228 million tweets from 56 thousand Finnish Twitter accounts between January 2018 and March 2021, were collected and analyzed. In addition to this, a dataset of 18 million tweets from 16 thousand users geo-located in Houston, Texas around the time of Hurricane Harvey (August, 2017) were used.

3.1.1 Data sources

This study focuses on Twitter, a social networking service created and founded in 2006. In Twitter, registered users can post and view messages of a maximum of 140 characters in length called a tweet. Also sharing a tweet originally created by another user, called retweeting, is possible, and counted as a user activity in this thesis.

The data was collected from Twitter utilizing Twitter API version 2, and its full-archive search that Twitter provides for academic research purposes. This endpoint allows the user to programmatically access public Tweets from the complete archive dating back to the first Tweet in March 2006, based on the user's search query.

Three different kind of searches were made in this thesis. First, the generic tweet search was used to gather a list of users that had used specific keywords. These results may include parameters such as user id, tweet content and creation time, and geo-location. Secondly, the user tweet timeline that provides access to tweets published by a specific Twitter account was used for the collection of the user activity. Lastly, user lookup was utilized for gathering the user information such as account creation date, location, and name.

3.1.2 Data collection

The users were selected randomly from a pool of 150 000 usernames. The list of potential usernames were collected using hundreds of popular Finnish hashtags, e.g. #kesä, #jääkiekko, #luonto, #jäätelö, and #bbsuomi (meaning summer, ice hockey, nature, ice cream, and the Finnish Big Brother format, respectively). For example, the ice cream hashtag was used by 1616 unique Twitter accounts during the study period.

As a result of the data collection, the actions of randomly selected 56 thousand users were combined. These users have produced a total number of 227 603 117 tweets during the analysis period of 1.1.2018-31.3.2021.

It is worth mentioning that this study suffers from limitations, and some of these suggest promising directions for future work. First of all, while geo-location could provide a more reliable selection of users located in Finland, it was not utilized due to very few accounts providing that information. Also, the current plan is to focus on Finnish users that communicate in Finnish to ease the potential next phase analysis of the content.

The selection bias of studied individuals has been minimized by collecting a large number of potential candidates and performing random selection. However, the study was completed on a Finnish speaking population, limiting out a variety of users including Swedish speaking Finnish population, non-Finnish speaking immigrants, and Finnish users that prefer to communicate in the medium on other languages, e.g. in English, or prefer not to use hashtags. Secondly, it is worth noticing that Twitter user base does not cover all population groups equally, but is more common among the younger population as was shown in Section 2.1.3. Potentially, further future work can adopt methods such as location-based filtering to better account for geo-cultural and linguistic confounds.

The second dataset focuses on individuals that were potentially displaced by the Harvey floods. The dataset has been collected and provided by Postdoc. Ted Chen, and is described in detail in [16]. Similarly, to the COVID-19 dataset, the data is also based on Twitter users, and has been originally collected using the Twitter research API that was previously described. The target population for the study was specified as those who lived in the Houston area prior to August 2017, and tweeted at least once during the hurricane period. Only users with geo-located tweets are included in the dataset.

3.2 Preprocessing

After collecting the massive amount of data, a rigorous preprocessing is required in order to retrieve a subset of acceptable users in the right format. This preprocessing start by selecting the subjects based on various filtering reasons, after which the data is aggregated for easier analysis. Next, these two processing steps are described in detail.

3.2.1 The selection of subjects

The data cleaning comprises of multiple phases: First, bots, news papers and other such accounts that do not act as personal accounts of individuals, were filtered out. This included the removal of accounts that ended with term "bot", started with term "yle" (the Finnish public service media company), or contained terms "robot", "news", or "uutiset" (news in Finnish). Also, verified accounts were considered unsuitable for this study due to often representing corporations, organizations, or individuals such as politicians, whose accounts may be handled by multiple individuals.

Secondly, inactive users whose activity was on average less than one tweet per week, were removed. The threshold for such minimum average activity level was 170 tweets in total for the COVID-19 dataset, and 26 tweets for the Hurricane Harvey dataset. Additionally, users that weren't active throughout the whole study period, or had significant gaps in their activity were considered unfit for this study. The study period was divided into 3-month bins, i.e. yearly quarters, and it was expected that each bin would contain at least one tweet to show activity.

Lastly, users that could not be tied to the dataset location, i.e. Finland for the COVID-19 and the US for the Hurricane Harvey datasets, in any way were removed. For the COVID-19 dataset, such indication was at least a single message labelled as Finnish or Swedish, or a geolocation or an account location in Finland. For the Hurricane Harvey dataset, this filter was not necessary due to the fact of each user being geo-located in the Houston area as a precondition of becoming selected for the dataset.

While analyzing abnormally active user accounts manually, it was unexpectedly found that such accounts did not only belong to bots, news organizations, and other non-personal usage, but also to individuals that happened to be very active in Twitter. In order for the datasets to contain as realistic variety of individuals as possible, no upper limit was added to the filtering process. However, such users could be considered outliers and have a negative effect on some models and statistics, which shall not be forgotten. The most active individual in the COVID-19 dataset had on average 18 activities per hour. For the Hurricane Harvey dataset, the respective value was 0.7.

In Figure 4, the amount of users being filtered out by each reason are presented. It is worth noticing that each user may have multiple reasons for being removed. The high amount of users that were filtered out due to quarter inactivity indicates that many users were either created only after the starting of the study period, or stopped using the service for long period of time during the study period. While in the COVID-19 dataset one third of the users were very inactive, writing less than once a week on average, in Hurricane Harvey dataset less than three percentages were removed for this reason. The datasets had around two percentages of verified users, indicating an interesting similarity across the datasets that originate from different cultures and time periods. Only a few bot accounts were found in both datasets, 39 in the COVID-19 and one in the Hurricane Harvey dataset. While the COVID-19 dataset had 146 news accounts, only four were found in the other. Lastly, 2 178 users were removed from the COVID-19 dataset due to the lack of any indication of the user being located in Finland during the study period.

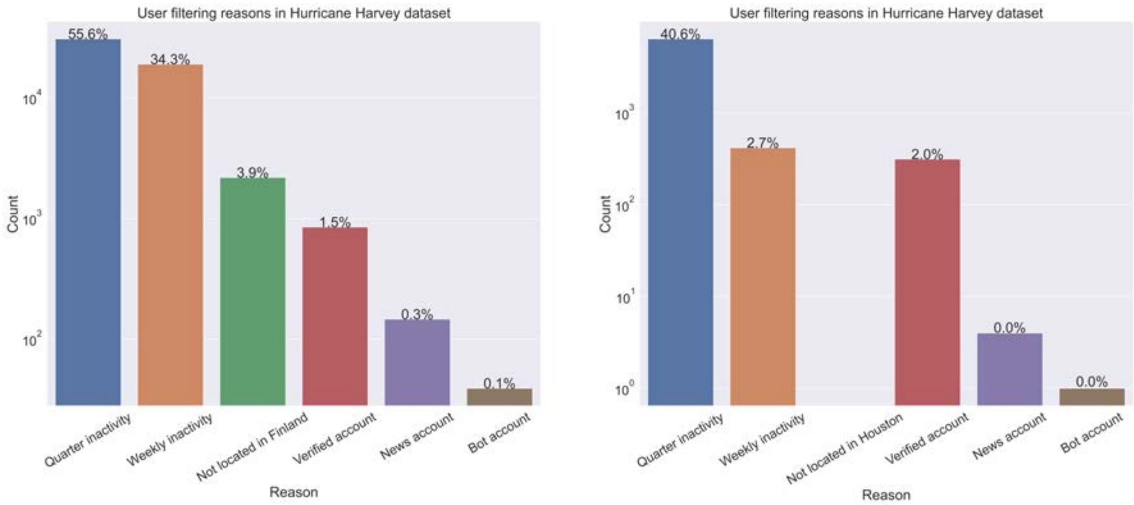

Figure 4: Total number and percentage of users in the datasets being filtered out due to different reasons.

Out of the original 55 222 users in COVID-19 dataset, around 60 percent were filtered out, leaving 22 023 users for this study. For the Hurricane Harvey dataset, 42 % of accounts were removed, resulting in 9 141 accounts out of the original 15 698. More detailed statistics on the datasets are given later in Section 3.3.1.

3.2.2 Aggregation of data

In this research, the weekly data for each individual was aggregated over the course of study period into 168 one-hour bins (7 days × 24 hours), following a calendar week that starts on Monday morning, and ends on Sunday night at midnight.

To build the individuals' daily Twitter activity patterns, we normalize each individuals' time series on weekly levels to unity so that the cumulative activity of each user on each week equals to one - as long as there is at least a single activity during the week. For some daily level analyses, a daily level normalization is used instead of the weekly one. These cases are mentioned separately. The normalization helps us to derive a comparable activity rhythm of the person, equalizing users with different activity levels, simultaneously ensuring that active users do not dominate less active users on population level analyses. The normalization process is also often fundamental for the usage of metrics such as Earth mover's distance. These individuals' normalized activity patterns can then be averaged to gain a population level average user. The result can be seen as a probability distribution of Twitter activity events at different times of the day.

3.3 Diving deeper into the data

In order to gain an overview of the datasets and their nature, an exploratory data analysis is provided in this section. The focus is kept on the users and their tweeting activities, as well as the temporal development of the datasets as a whole in order to detect their potential weaknesses and biases.

3.3.1 Overview of the data

An overview of the datasets is provided in Table 1. As was discussed earlier in Section 3.2.1, around 60 percentages of the users in COVID-19 datasets were filtered out for different reasons. Regarding tweets, this meant that around 32 percentages were removed, leaving 156 million actions for this study. In the Hurricane Harvey dataset, around 42 percentages of accounts, and 51 percentages of tweets were filtered out.

	COVID-19	Hurricane Harvey
Study period dates	1.1.2018-31.3.2021	1.9.2016-31.8.2018
Study period in months	39	24
Unique users	55 222	15 698
Unique users after filtering	22 023	9 141
Tweets	227 603 117	18 250 160
Tweets after filtering	155 908 055	8 997 604

Table 1: An overview of the data.

In igure 5, the total number of tweets that remained after all the data processing steps, are presented in monthly bins for each dataset. For the COVID-19 dataset, the left figure shows the increasing trend that can be explained by the growing popularity and usage of Twitter in Finland that was discussed previously in Section 2.1.3. For the Hurricane Harvey dataset, the situation is very different, decreasing during the winter 2016, stabilizing for the spring to autumn 2017, and decreasing again during winter 2017 before stabilizing for spring 2018. Reasons for this remain unknown. From both graphs, we can see the crisis events as the most active months during the study periods. For the COVID-19 dataset, this value is reach on March 2020, and for Hurricane Harvey during August 2017.

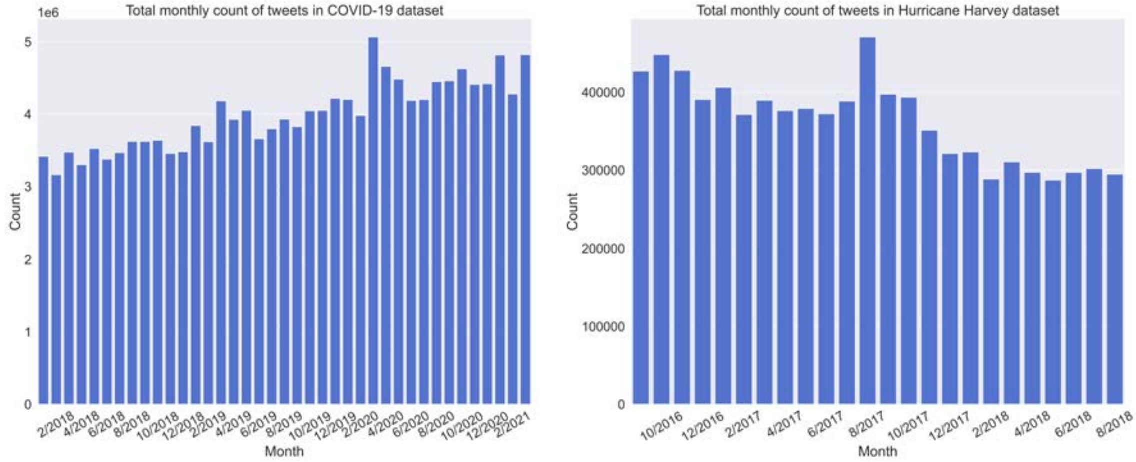

Figure 5: Monthly tweet count for the COVID-19 and Hurricane Harvey datasets.

3.3.2 Users in Finland and the US

As was discussed earlier, after filtering out unsuitable accounts, 22 023 unique users for the COVID-19 dataset and 9 141 users for the Hurricane Harvey dataset remained for this study. The Finnish users had on average of 7 047 actions, and a median of 1 412 actions during the study period, while the corresponding values for users in Houston were 1 162 and 696. Part of the high average in the COVID-19 dataset can be explained by the few hundred hyper active accounts. Interestingly, the median daily activities in Finland and Houston were relatively similar, being 1.17 and 0.97 postings per day. In Table 2, a comparison of the datasets regarding the platform usage is given.

	COVID-19	Hurricane Harvey
Average tweet count per user	7 047	1 162
Median tweet count among users	1 412	696
Median activity / day (number of tweets)	1.17	0.97
Users with at least one geotag (%)	9	100

Table 2: An overview of the users in COVID-19 and Hurricane Harvey datasets.

It is worth noticing that less than one tenth of the Finnish users had at least one geotagged tweet, giving indication on how seldom it is being used. On the other hand, 64 percentages of these users had some location marked in their profile. In this thesis, this information has been utilized for increasing confidence that the user's location has been in Finland during the pandemic. In the future, this information could potentially be utilized for location based sub grouping of users. For the Hurricane Harvey dataset, the 100 % geotagging can be explained by the dataset being collected specifically based on this factor.

In Figure 6, the number of users for different total number of actions in the datasets are presented. The left figure has been cut at 40 thousand for better visualization. However, there are 788 users with more than 40 thousand postings that form a very long and thin tail, the maximum being an enormous 2 444 130 tweets by a single user in the COVID-19 dataset. As can be seen from the left figure, the number of users have somewhat exponential decay as the number of actions increases. For the right figure, a decaying pattern as an increase of the activity can also be seen. However, the graph shows a strange behaviour having a noticeable spike of highly active users.

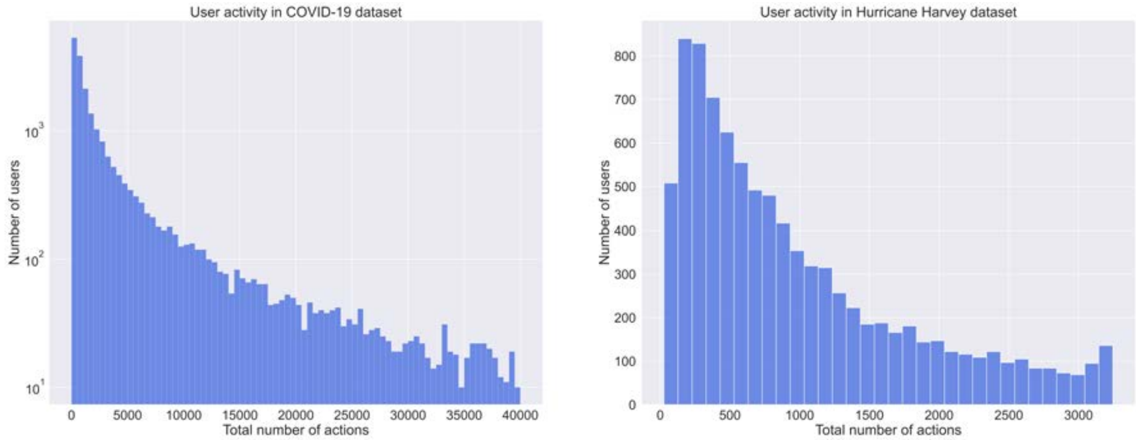

Figure 6: Activity of users in the COVID-19 and Hurricane Harvey datasets. For the COVID-19 dataset, a threshold is set at 40 000 actions due to visualization purposes. Notice different scales in the figures.

A comparison of the average activity rhythms of the selected individuals from the COVID-19 and Hurricane Harvey datasets are shown in Figure 7. The average rhythms show a clear daily periodicity where the lowest level of activity coincides with night times. While the weekly rhythms between the datasets follow similar patterns, it can be seen that the US users are more consistent throughout the week whereas Finnish users tend to be more active during weekdays, and have a lower activity during weekends.

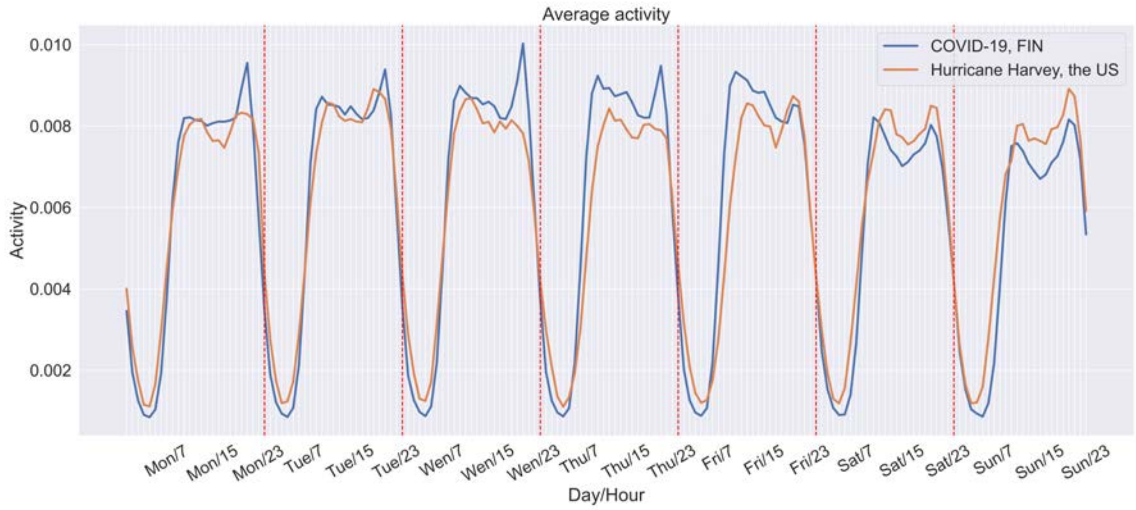

Figure 7: Average user activity between the COVID-19 and Hurricane Harvey datasets.

While looking at these results, we have to keep in mind that these rhythms are a superposition of many individual rhythms. Different people can have very different rhythms and still produce what we see in the figure as the average rhythm. Therefore, we should avoid assuming that the system's aggregate rhythm will be similar to the individuals' rhythms.

3.3.3 Gender analysis

Next, a name analysis was performed on both datasets. It is worth noting that this binary gender analysis is very shallow, and solely provides superficial understanding on the demographics of the datasets, and might warn of potential biases.

In the analysis, the name of each account was first retrieved using Twitter API. After this, the names were split at the first space so that the first part was assumed to represent the first name of the account user. Potential special characters such as commas and dots, as well as numbers were removed sot that only alphabetical characters remained.

For the COVID-19 dataset, the first name was compared to the names on the Finnish name day calendar as well as extensive lists of Finnish men's and women's names combined from multiple sources. The name day calendar contains 482 names for men and 523 names for women, while the more extensive list of names contains 1836 and 2486 names for men and women, respectively. As a result of the analysis, 5097 names were matched to men's names in the name day calendar, and 4019 to women's names, corresponding to 31 % and 25 % , and leaving 44 % without a match. In the extended list, the numbers were 5759 matches for men and 4584 matches for women, corresponding to 35 % and 28 %, leaving 37 % without a match. Based on this analysis, around 56 percentages of Finnish users were men. This result aligns very well with the survey statistics from 2020 [49] that concluded that around 59 percent of Finnish Twitter users were men.

For the Hurricane Houston dataset, an online name-to-gender inference service called genderize.io (https://genderize.io/) was utilized. The service was created in 2013, and attempts to infer the gender of a first name. The underlying data is collected from social networks across 79 countries and 89 languages. The service has been highly popular among publications. Based on Santamaría and Mihaljević [44], the tool works excellently, especially for Western names, gaining accuracy up to 98 % with tight confidence requirements, and inaccuracy rates below 15 % with looser requirements. However, Santamaría and Mihaljević found some gender bias in the tool, wrongly identifying more men as women. Additionally, the study state that the tool does not suit particularly well for inferring the gender of Asian names.

For the prediction, a few optional, configurable parameters exist: confidence parameters count and probability represent the number of data entries used to calculate the response and the proportion of names with the gender returned in the response. Additionally, location and language can be given to improve the inference. In this study, the location was set to the US, and the rest of the parameters kept at default. As a result of the analysis, 66 percentages of names were classified as male and 34 percentages as female. This indicates some bias and overrepresentation of men compared to studies such as [46] that indicate that 54 percentages of American Twitter users are male and 43 percentages female.

3.4 Methods

We start this section by presenting the theoretical background of non-negative matrix factorization, including its algorithms, constraints, and applications. After this, cophenetic correlation coefficient and Pearson correlation coefficient are reviewed. Lastly, the metrics used in this study, namely the Earth mover's distance, cosine similarity, and root mean square error are presented.

3.4.1 Non-negative matrix factorization

A wide variety of matrix factorization (MF) algorithms have been developed over the past decades. Some of these algorithms have also proven useful in statistical data analysis, especially the singular value decomposition (SVD), which underlies principal component analysis (PCA). Recently, studies in machine learning have focused on matrix factorization that directly targets some of the special features of statistical data analysis. One of such methods is non-negative matrix factorization (NMF) that approximates a given data matrix whose elements are non-negative as a product of two non-negative matrices. [17]

The methods ability to automatically extract interpretable decompositions from a set of non-negative data vectors has made it a commonly used tool for the analysis of high-dimensional data. This is common occurrence in datasets derived from text and images, and has been applied to a variety of applications from pattern recognition to text mining ([52], [45]). Other fields of applications have included DNA gene expression analysis ([6]), speech and music separation ([51], [34]), and the analysis of temporal patterns of human behaviour ([3]), among others.

In contrast to many other linear representations such as principal component analysis and independent component analysis (ICA), NMF provides dimensionality reduction that respects the non-negative constraints of the input data, making the representation purely additive. This results in parts-based representation of the data, and can lead to the discovery of data's hidden structures that have meaningful interpretations. [57]

In linear algebra, a matrix factorization is a decomposition of a matrix into a product of matrices. For example, input data matrix $X = (x_1, ..., x_n)$, which contains n data vectors of m dimensions, can be factorized into the product of matrices $W = (w_1, ..., w_r)$, and $H = (h_1, ..., h_n)$, by writing:

$$\mathbf{X} = \mathbf{WH}. \tag{1}$$

Typically in MF, both the input matrix X and the factorized matrices W and H can contain either positive or negative entries, while in NMF these are solely non-negative so that

$$\mathbf{X} \approx \mathbf{WH}. \tag{2}$$

One of the main parameters in NMF is the factorization rank r which defines the number of components used to approximate the target matrix. The choice of this value is typically problem dependent, but a common way of deciding the optimal value is to compute a quality measure on different r values, and to choose the best value according to the measure's criteria. The most common approach is to use the cophenetic correlation coefficient. It is suggested to choose the smallest value of r for which this coefficient starts decreasing. Another approach is based on the variation of the residual sum of squares (RSS) between the target matrix and its estimate where the optimal factorization rank is the value of r for which the plot of the RSS shows an inflection point. [19]

Many numerical algorithms have been developed to solve the NMF problem. In order to measure the discrepancy between the input data and the low-rank approximation, Kullback-Leibler (KL) divergence is often used as an objective function. [43]

While various initialization approaches for NMF exist, random initialization has been commonly used among the literature. In this approach, one needs to run an NMF algorithm several times with different initial matrices and selecting the best solution. However, choosing the initial matrices randomly often gives a poor solution. A more detailed look on the mathematical background of NMF and its various algorithms have been provided among others in the dissertations by Yuan [57] and Zhang [58].

3.4.2 Searching for correlations

Two kinds of correlation coefficients were utilized in this thesis: Cophenetic and Pearson. Next, these are presented.

Cophenetic correlation coefficient

The cophenetic correlation coefficient was proposed by Brunet et al. [12] to measure the stability of the clusters obtained from NMF. The coefficient indicates the dispersion of the consensus matrix. It is computed as the Pearson correlation of two distance matrices: the distance between samples induced by the consensus matrix, and the distance between samples induced by the linkage used in the reordering of the consensus matrix. In a perfect consensus matrix where all entries equal to 0 or 1, the cophenetic correlation coefficient is 1, while when entries are scattered between 0 and 1, the coefficient is below 1. The metric can be used for finding the optimal number of components in NMF by selecting the value of k where the magnitude of the coefficient begins to fall. Various programming language libraries provide a computation for the metric, and for this study Skicit learn was used.

Pearson correlation coefficient

One of the main objectives of this thesis is to find quantifiable changes in Twitter user activities, which can be linked with the ongoing crisis event. For this purposes, Pearson correlation coefficient can be used.

Pearson correlation coefficient is used to measure the statistical relationship between two continuous variables. The coefficient is based on covariance, and it gives information about the magnitude of the association as well as the direction of the relationship. The Pearson correlation coefficient measures the linear correlation between two variables, and takes a value between -1 and 1, so that values close to -1 indicate a negative linear correlation, and values close to 1 indicate a positive correlation. A correlation coefficient close to 0 indicates that there is no linear correlation between the variables. The equation for Pearson's correlation coefficient can be written as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3)$$

3.4.3 Measuring distances, similarities, and errors

In this thesis, a set of distance, similarity, and error metrics have been utilized to quantify changes and differences between temporal patterns and NMF components. For quantifying changes in temporal patterns, the Earth mover's distance (EMD) is used. In the optimization and robustness testing of NMF, on the other hand, we have utilized cosine similarity and root mean square error (RMSE). Next, these three metrics are discussed in more detail.

Earth mover's distance

The Earth mover's distance, also known as the Wasserstein metric, is a measure of the distance between two probability distributions. Intuitively, it can be seen as the minimum amount of work required to transform one distribution into another, where work is measured as the amount of distribution weight that must be moved, multiplied by the distance it has to be moved. This can be computed by solving an instance of transportation problem, using any algorithm for minimum-cost flow problem. [43]

The EMD can be motivated by the following example: The distance between two distributions, in this case the set of bins that form the activity patterns, which have small deformations in neighboring bins should be smaller than that of distribution pairs which differ in non-neighboring bins. Therefore, the metric should sum the changes required to transform one signature into the other with low cost given to local deformations and high cost to non-local ones. [43]

Formally, the EMD distance is formulated as a linear program which aims to minimize the flow f_{ij} between the bins of the source signature (i) and the bins (j) of the target signature for a given inter-bin flow cost d_{ij}. The cost parameter d_{ij} specifies the inter-bin flow cost for each pair of source and target bins. This has been described in equation 4. [43]

$$EMD(P,Q) = \frac{\sum_{i,j} f_{ij} d_{ij}}{\sum_{i,j} f_{ij}} \tag{4}$$

Cosine similarity

Cosine similarity measures the similarity between two vectors of an inner product space by measuring the cosine of the angle between the vectors. The similarity is determined based on whether the vectors are pointing in the same direction or not: A cosine value of 0 means that the vectors are orthogonal to each other and have no match, while values closer to 1 indicate greater similarity. [23]

We can define cosine similarity as

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{y}_i)^2}}, \tag{5}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the two vectors that are being compared for similarity, and $||\boldsymbol{x}||$ and $||\boldsymbol{y}||$ are the Euclidian norms of these vectors, conceptually being their lengths. [23]

The metric is used extensively in a variety of tasks from data mining to information retrieval, and text matching. Typical fields of applications for the metric include document classification and clustering in natural language processing ([24]), and gene expression analysis in bioinformatics ([14], [26]).

Root mean square error

Root mean square error is calculated as the standard deviation of the residuals, i.e. the differences between the observed and predicted values. It is a measure of how spread out these residuals are, and can be defined as follows:

$$RMSE = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}. \tag{6}$$

RMSE is a general-purpose error metric for numerical predictions, and is commonly employed in model evaluation in machine learning. The metric is being used in a great variety of applications from climatology to finance, and in this thesis it is utilized for optimizing the NMF parameters, and evaluating the robustness of the resulting models.

4 Results

This section presents and evaluates the results, while the next section is dedicated to discussing the results and drawing conclusions on them. First, in Section 4.1, the results of the NMF parameter optimization are shown, and the robustness of the method is validated. After this, in Section 4.2, the results on differences between individuals based on the built NMF model, are presented. In Section 4.3, the effects of COVID-19 on the user activities are described. Lastly, in Section 4.4, the results of similar analysis on the Hurricane Harvey -case are shown.

4.1 Optimization of non-negative matrix factorization

In order to utilize NMF effectively, one has to have a good understanding on the suitability of the model on the given dataset, as well as its limitations and optimization. Considering the multiple variables affecting the model fitting such as the number of users and their activity, the selected temporal resolution, as well as the configurable parameters such as the number of components and the method of initialization, optimizing an NMF model fitting is far from trivial. In this section, the NMF model is optimized for the COVID-19 dataset, its limits are sought, and its robustness tested.

4.1.1 Ideal number of components

As was discussed earlier in Section 3.4.1, one of the tools used for finding the optimal number of components in NMF is cophenetic correlation coefficient. This measure is based on the stability of components in different runs of the algorithm. In this thesis, we calculated the cophenetic correlation coefficient using Nimfa, a Python library for NMF [60]. We started with three components which is the number of commonly used chronotypes, and set the maximum number of iterations to 200, and the number of runs to one thousand. According to Brunet et al. [12], the first number of components, where the magnitude of the coefficient begins to fall should be selected. In Figure 8, we are able to see how this point is found at four components. Similar results were found on the Hurricane Harvey dataset.

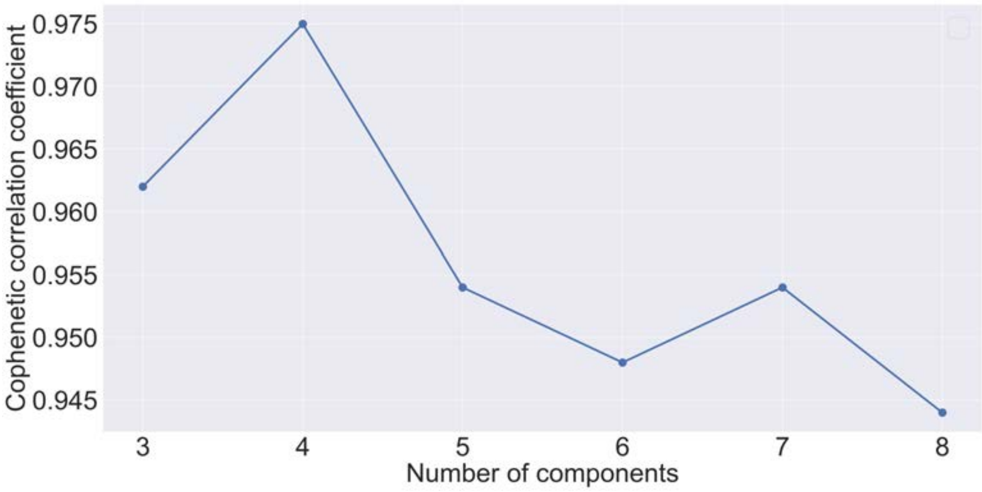

Figure 8: Cophenetic correlation coefficient versus the number of components for the COVID-19 dataset.

The achieved result was investigated further by randomly initializing and fitting ten models for each number of components on the COVID-19 dataset, and calculating the cosine similarities between the resulted components, as well as the RMSEs of the resulted models. In Figure 9, these results are presented as a function of the number of components. On the left plot, we can see how the cosine similarities decrease with the increasing number of components. Similarly, the average RMS-error decreases with the increasing number of components due to the increasing amount of information, although these differences are almost insignificant. Both, for the sake of robustness and interpretability, four components seem to be ideal for this dataset.

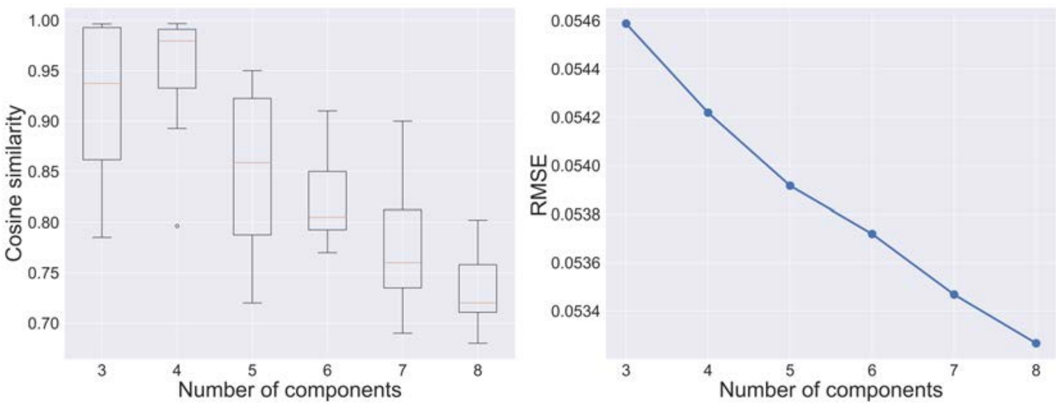

Figure 9: The effect of the number of components on RMSE and the similarity of random initializations.

4.1.2 Number of bins

The number of bins, i.e. the selected resolution that the activity patterns are built on, provides an interesting aspect for the optimization. Typically, in the existing research, 24 hourly bins have been the most popular choice, being easy to build and intuitive to interpret. However, in cases that handle extremely sparse data, such as users that are active on weekly rather than daily level, especially if the activity patterns are built from a very short time interval, such as one day or a week rather than a month or longer, 24 hourly bins might become too ambitious of a resolution.

In this thesis, 4, 6, 12 and 24 bins per day were analyzed. The analysis was performed similarly to the one presented in the previous section, by running ten random initializations for each number of bins. The resulted RMSEs and cosine similarities are presented in Figure 10. As one can expect, the average RMS-error decreases with the increasing number of bins due to the increasing amount of information. Additionally, the cosine similarity between different runs increases for increasing number of bins, indicating more robust performance. For these reasons, 24 bins have been used for the rest of this study. However, as was stated before, one might still want to consider using fewer bins in cases where the amount of data is not sufficient for such a high resolution analysis.

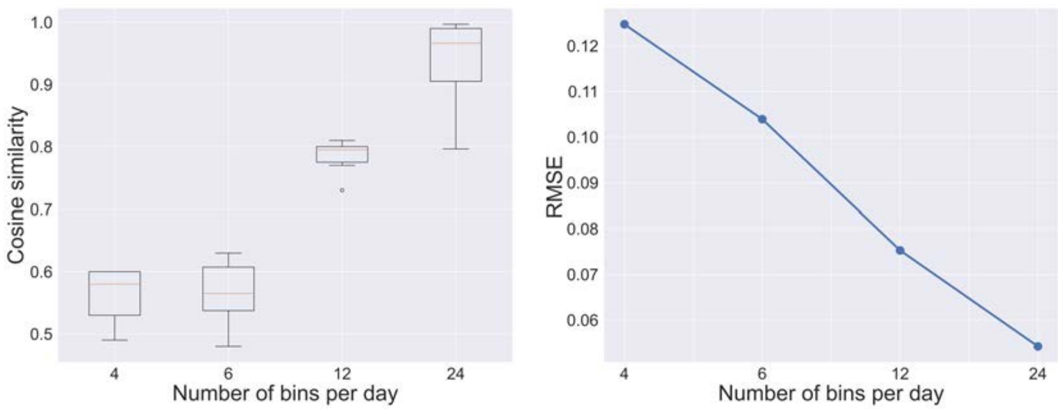

Figure 10: The effect of the number of bins on RMSE and the similarity of random initializations.

4.1.3 Testing the robustness

In the previous sections, we have already seen how the results are sensitive to the parameter configuration and the initialization of the model fitting. We have also shown that when the model configuration is optimized to the dataset in hands, the results remain robust. In Figure 11, models have been fitted on each of the studied datasets, namely COVID-19 and Hurricane Harvey, for the whole length of the study periods. The resulting models are almost identical, having an average cosine similarity of 0.989, indicating strong robustness even across datasets.

Similar results were achieved when fitting the models on subsets of users as long as a sufficient number of users remained. We were able to introduce unsteadiness to the model fitting by decreasing the number of users significantly, by only selecting users with very low activity, and by shortening the temporal period used for building the activity patterns significantly. This was further affected by the choice of the model's initialization method.

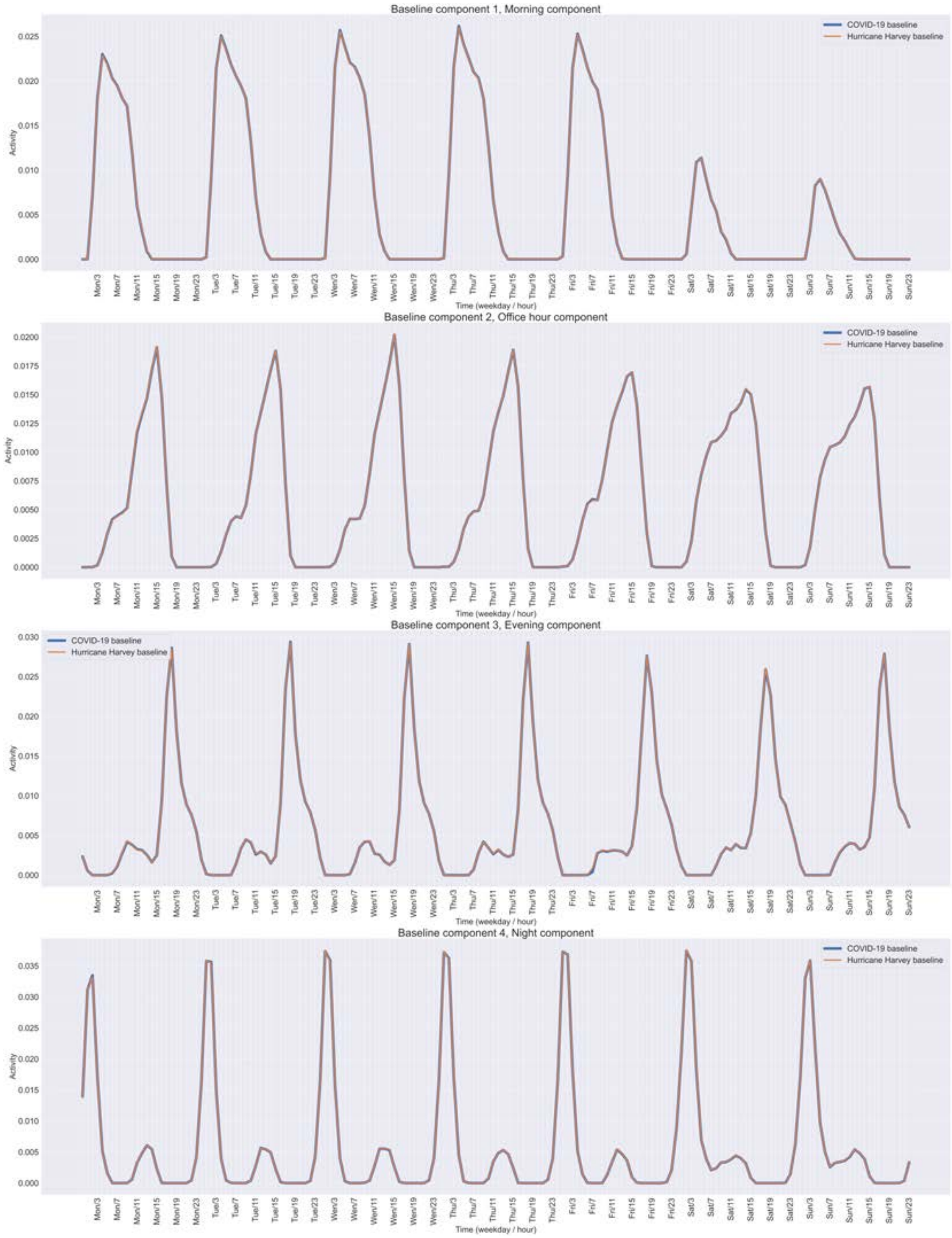

Figure 11: The resulting NMF models fitted with COVID-19 and Hurricane Harvey datasets.

4.2 Differences among individuals

In Section 3.3.2, it was discussed how an average user can be formed from a population for identifying general activity patterns and their changes. It was also stated that these rhythms are a superposition of many individual rhythms, different people can have very different rhythms. In Figure 12, this has been demonstrated by visualizing eight users with the highest value for each activity component. In the figure, it can be clearly seen how some users are more active in the morning, some in the afternoon, and some in the evenings. The figure also shows how there exist individuals whose rhythms seem to be a complete opposite of the traditional circadian rhythm. These individuals are active mainly during the night, and seldom during the daytime.

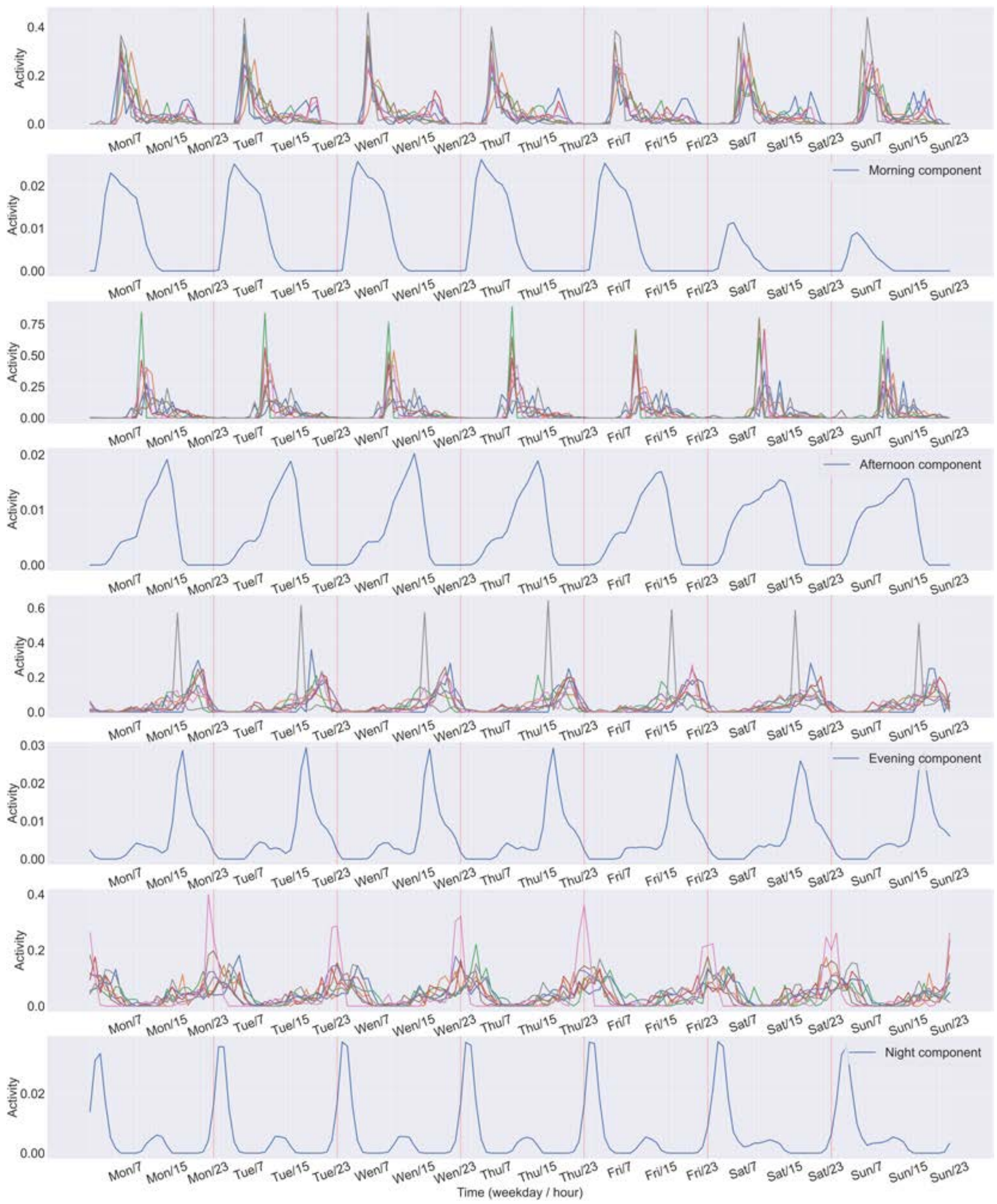

Figure 12: Individuals with the highest values for each activity component for the COVID-19 dataset.

Figure 12 raises questions whether individuals could be categorized into a set of clusters. In Figure 13, a set of 500 randomly chosen users from the COVID-19 dataset are plotted based on their activity component values. As a result of the plotting, no separation into clear clusters is visible, and instead individuals seem to form a continuous spectrum of all kinds of activity rhythms. However, a more comprehensive investigation utilizing modern clustering methods would be needed to confirm the result.

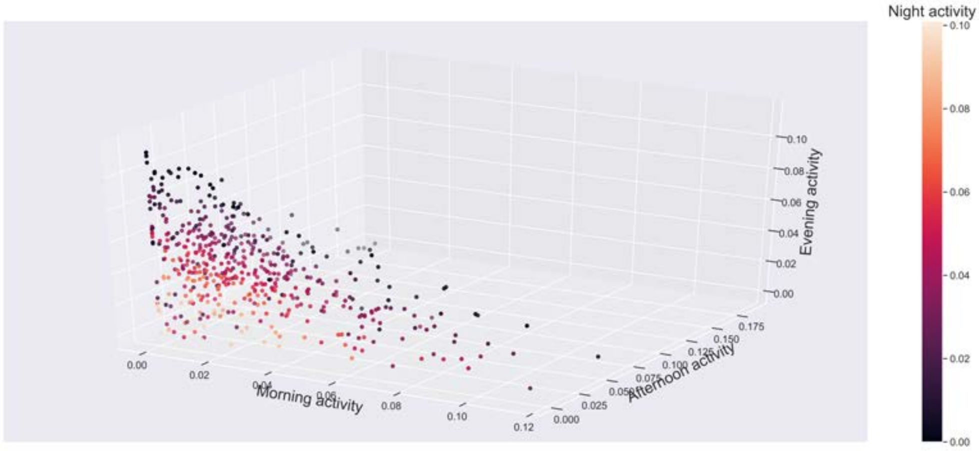

Figure 13: User weights.

4.3 Effects of the COVID-19 pandemic on the activity patterns

While COVID-19 has had a great variety of different effects, in this section the focus is kept on two cases in particular: First, how the COVID-19, especially the lockdown on 16th March has affected the activity patterns of the population, and secondly, whether some relationship can be found between changes in the activity patterns and the COVID-19 government responses.

In Figure 14, the Earth mover's distances for Monday activities to the baseline are plotted. The baseline activity pattern has been formed as an average of 7 different Mondays 6 to 12 weeks before the March 16th lockdown. The graph shows how the lockdown has caused a significant change to the activity patterns that seem to last around four weeks before returning to normal. In Figure 15, the population average Monday activities as well as the baseline are plotted. This figure shows how the week before the event the population level activity is very close to the baseline situation prior the COVID-19 restrictions in Finland. During the event week itself, a significant spike at 5 PM can be seen, indicating the moment the government announced the lockdown. During the weeks following this, the patterns go steadily back towards the baseline.

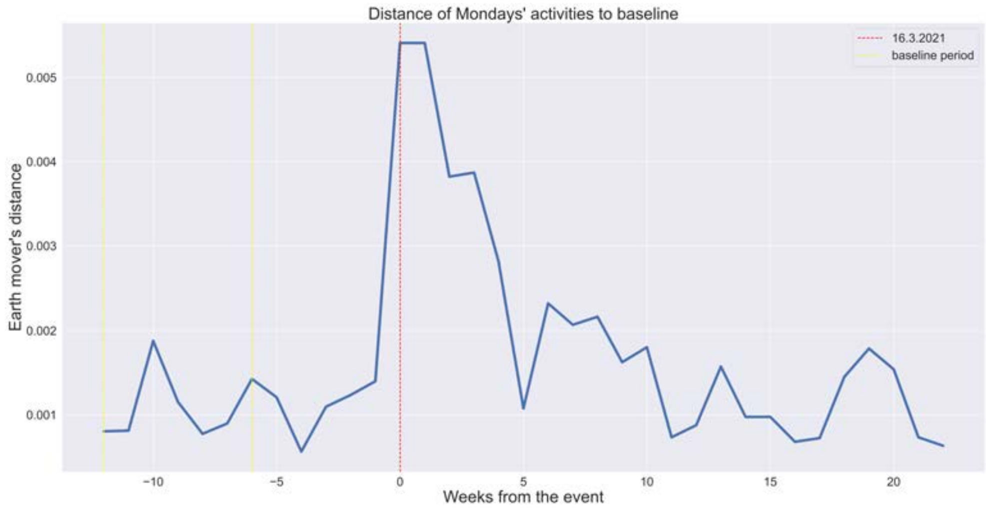

Figure 14: Distance from Monday activities to the baseline.

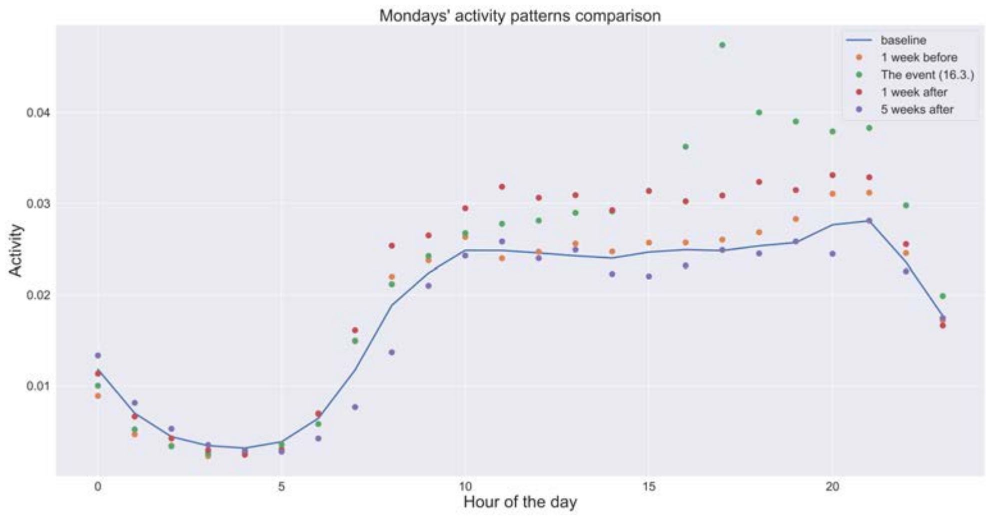

Figure 15: Monday activity patterns for weeks before and after the event.

Figures 16 show the average activity patterns during the first two weeks of lockdown and its comparison to the corresponding time periods the years before and after. In Figure 17, a similar comparison is made regarding the Summers 2018 to 2020. From the graphs, it can be seen how the relative weekend activity decreased significantly during the times of COVID-19, and especially the spikes on Friday, Saturday and Sunday evenings are missing from the 2020 and 2021 plots. Regarding sleeping, all activity graphs demonstrate similar average sleep times around 00-07 AM when calculated as a crossing of an arbitrary chosen activity threshold of 0.0035. A few exceptions to this include the weekends during 2020 and 2021, when activity crosses the threshold at 8 o'clock. Similar phenomenon can be detected for weekdays of 2018. For better understanding of the potential effects of COVID-19 on sleeping habits, a deeper analysis will be required.

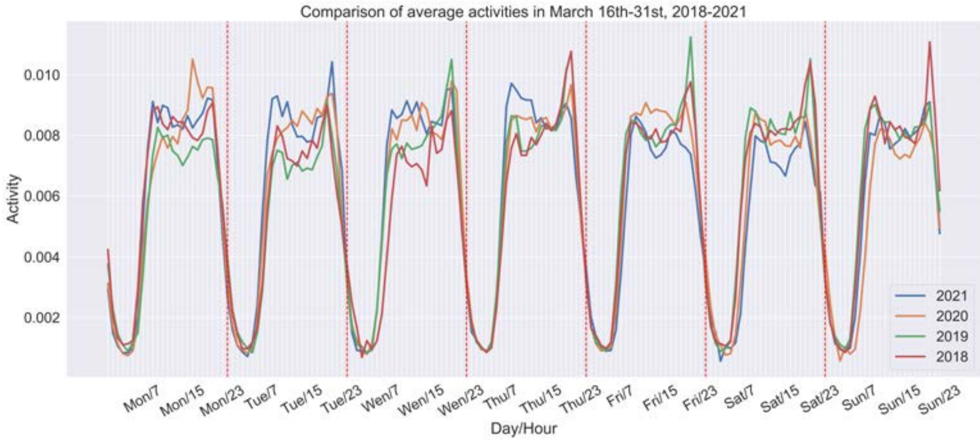

Figure 16: Comparison of average activities during March 16th-31st, 2018-2021.

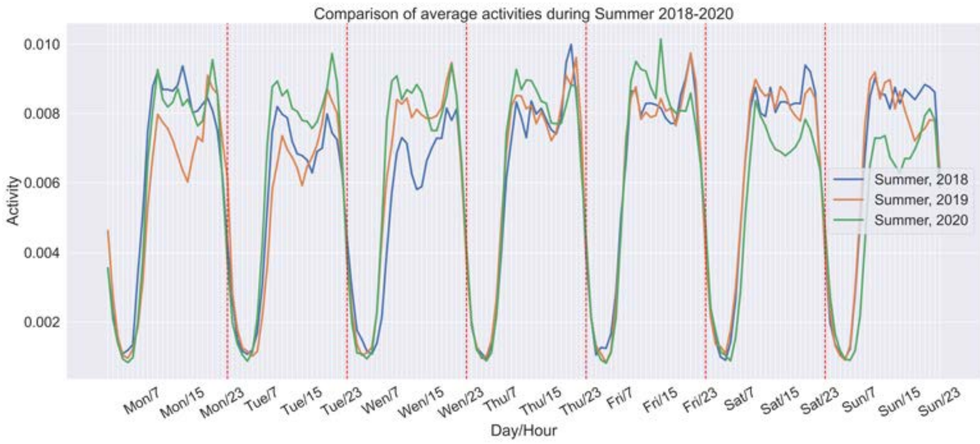

Figure 17: Comparison of average activities during June to August, 2018-2020.

In Section 3.3.2, it was stated that the activity rhythms may vary greatly on individual level. It is, therefore, understandable that also COVID-19 restrictions have varying effects on these individuals. In Figure 18, users with the greatest change in each of the activity component before and after the lockdown have been illustrated. The activity patterns and the components from which the differences have been calculated, have been formed from the three months before and after the lockdown on 16th March, 2020.

The user on the top can be seen to have an increased night time activity after the lockdown. Additionally, his or her Twitter usage has decreased in the afternoon. The second user with the greatest change in the afternoon component has a noticeably increased activity in the mornings. Similar effect can be seen on the third user, who before the COVID-19 tended to use Twitter relatively steadily throughout the day, ended up having high activity in the morning and decreased activity in the evenings after the lockdown. The fourth and last demonstrated user, has an increased activity during night time with the expense of decreased activity for the other hours of the day.

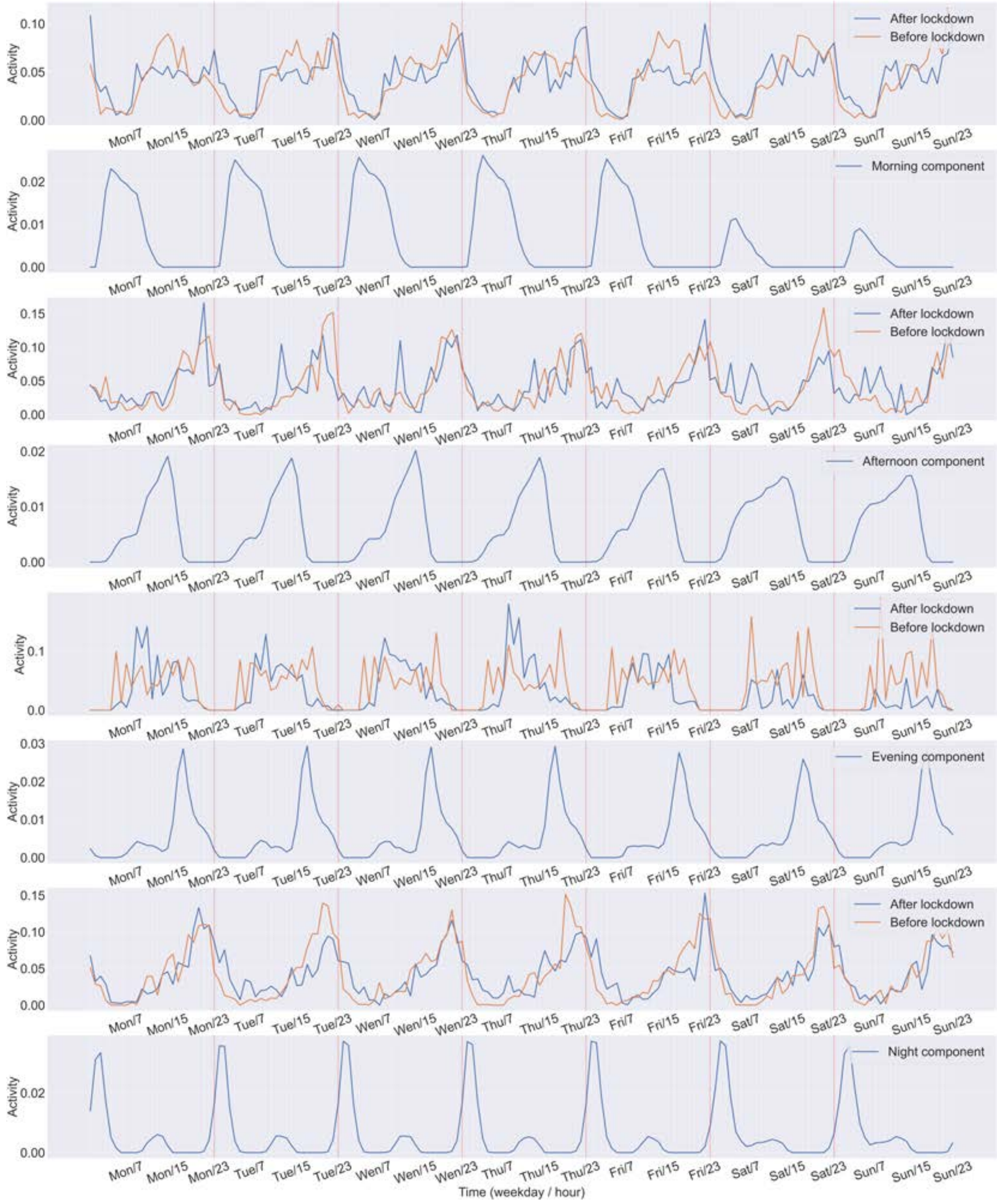

Figure 18: The effect of the COVID-19 lockdown on four individuals. The periods before and after the lockdown comprises of three months from and before 16th March, 2020.

In Figure 19, the average user activity pattern changes between 2020 and 2021 are plotted together with the COVID tracker values. As was discussed in Section 3.2, the average user activity has been calculated by normalizing each user on daily level, after which the average user activity has been formed. The change in activity pattern has then been calculated on weekly level by calculating the Wasserstein distance between each week on 2020 and their corresponding week in 2019. The COVID tracking value has been presented in Section 2.3.2. For this graph, the values have been averaged on weekly level in order to correspond with the resolution of the activity pattern. The red vertical lines in the graphs indicate March 16th, the day when the Finnish Government declared a state of emergency due to the pandemic. The yellow lines label April 15th, the day when travel restrictions between Uusimaa region and the rest of the country were lifted. The green line indicates May 21st, Ascension day.

A noticeable similarity between the graphs is visible. This was further confirmed by calculating the Pearson correlation coefficient between the two data series. As a result, a correlation of 0.61 which indicates moderate positive relationship was found.

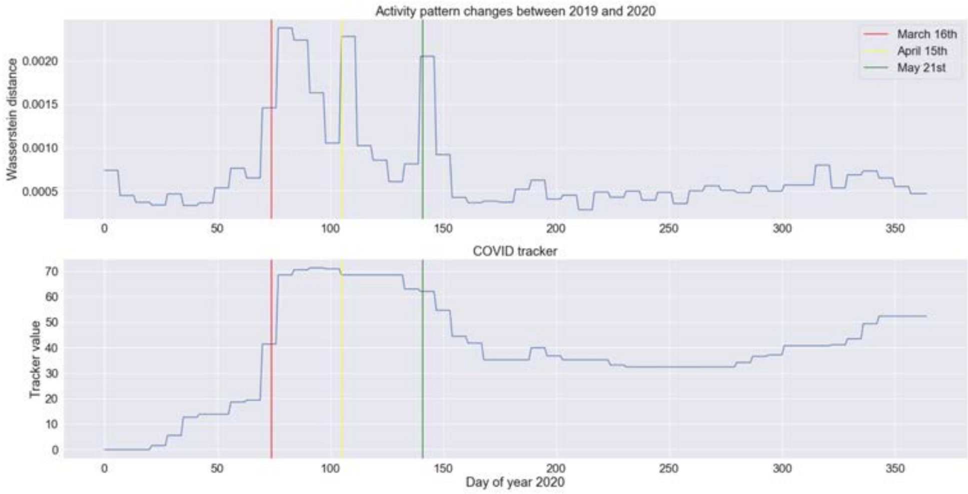

Figure 19: Average user activity pattern changes between 2019 and 2020 compared to a COVID tracker values.

4.4 Effects of Hurricane Harvey on the activity patterns

Hurricane Harvey has had significant effects on individuals in the Greater Houston area. In this study, these effects have been analyzed through the activity patterns of those individuals located in the area. In Figure 20, the total weekly activies are shown. From the figure, the major spike caused by the hurricane is clearly visible. Additinally, around nine weeks after the hurricane, the local baseball team Astros went on to win the team's first World Series championship, which can be seen as the second spike in the graph.

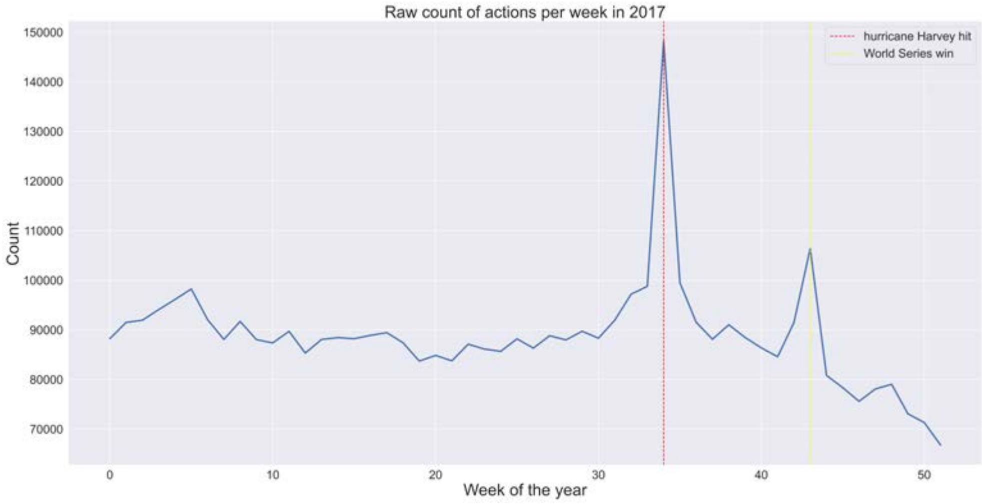

Figure 20: The total number of actions per week in 2017.

In Figure 21, the Earth mover's distances for normalized Thursday activities to the baseline are plotted. The baseline activity pattern has been formed as an average of the Thursdays 6 to 12 weeks prior to the hurricane reaching Houston. Perhaps unexpectedly, and unlike the earlier results seen regarding the effects of COVID-19 lockdown, the hurricane doesn't seem to have had any kind of effect to the population level activity patterns. On the other hand, the World Series championship win is showing an enormous spike, potentially due to celebrations that lasted late to the night.

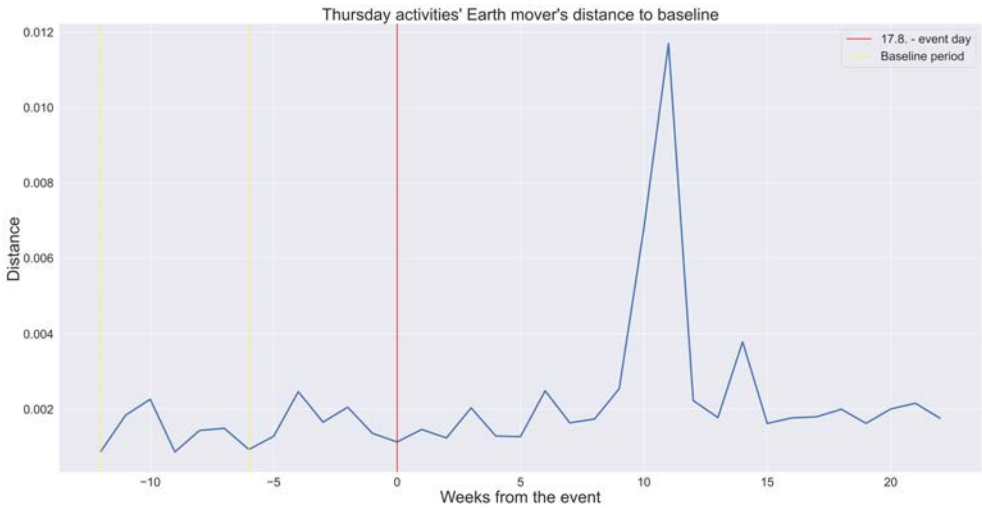

Figure 21: Distances from Thursday activities to the baseline.

5 Discussion

In this section, first the findings of this thesis are summarized, and the impact of the results are addressed. After this, the limitations related to the research are discussed. Various questions still remain open for future research, and these are discussed lastly.

5.1 Conclusions

Human activities follow temporal patterns such as daily, weekly, and seasonal rhythms that are related to physiology, natural cycles, and social constructs. Various studies have shown that these patterns have been affected significantly by the COVID-19 pandemic. The main objective of this thesis was to investigate if these patterns and their changes can be detected using social media data, as well as to better understand how crisis events such as the COVID-19 pandemic and Hurricane Harvey affect these temporal rhythms. This approach could provide a new perspective for better understanding the effects of crises, and the time it takes for the population to adapt to the new normal.

First, in Section 4.1, it was demonstrated how NMF can be optimized to provide robust results for analyzing temporal rhythms. As a result of the optimization, four main components were extracted, namely morning, daytime, evening, and night activities. Interestingly, while the average activity patterns between COVID-19 and Hurricane Harvey datasets differed slightly, their respective resulting NMF components shared high similarity to each other, indicating potential for generalization.

In Section 4.3, it was shown how the population level activity patterns differed from those of the year before during the pandemic, especially during the lockdown period in spring 2020. It was also shown that these changes had a moderate correlation of 0.61 with the overall COVID-19 government response index that follows the level of measures Finnish government had taken regarding the pandemic. It was also shown how during the first five weeks of the lockdown period, a major changes in activity patterns were experienced before returning to the normal. This may potentially give indication on the time it takes to adapt to the changes caused by the pandemic and the resulting lockdown. However, in this case, another explanation could be the reduction of some of the restrictions, namely the traveling to and from Uusimaa, around the same time.

Additionally, in Section 4.3, it was shown how the COVID-19 has had a varying effect on individuals, some experienced significant changes in their temporal patterns. While previous studies have found population level changes in the length of day during the pandemic, similar effects were not clearly visible in this study. However, some potential indication was found on people sleeping an hour longer during the weekends, as well as a significant drop in weekend evening activity during the time of the pandemic.

In Section 2.4, it was shown how the user activity increased significantly during events such as the hurricane, as well as the local baseball team's World Series championship win. However, unlike in the COVID-19 case, no changes in population level temporal patterns of activity were noticed during the crisis. Such changes were, on the other hand, caused by the championship win. Contrary to expectations, the results seem to indicate that the crisis did not end up having direct effects on temporal patterns on population level. This might indicate that comparing a population level average activity pattern of a crisis period to some baseline period in the past seem to work for long term crises that have a direct effect on every individual and their behaviour while other methods may be more suitable for shorter shock events.

Overall, it has been shown that Twitter data provides a fine-grained view of activity patterns on a large scale that can be used to better understand crisis events. This works as a foundation for future research and provides optimism towards the possibility of understanding the effects of crises on population through social media activity patterns.

5.2 Limitations

The limitations of this research are primarily data related, and some of these suggest promising directions for future work.

First, using Twitter as a data source may provide some issues. As has been seen in Section 3.3, the average user activity in Twitter is relatively low. As achieved temporal resolution depends heavily on the user activity, in some cases this may be limiting the possibilities, especially for individual level analyses.

Secondly, social media user demographic does not necessary represent well the whole population. There may also be a risk that the data cleaning that includes the removal of users purely based on their low activity may lead to removing some specific type of users, and ending up creating a bias. This could be investigated further, for example by comparing the demographics of the original user group to the filtered one, as well as to the general demographics in the study area.

The selection bias of studied individuals has been minimized by collecting a large number of data samples. However, the data collection focused on a Finnish speaking population, limiting out a variety of users including Swedish speaking Finnish population, non-Finnish speaking immigrants, and Finnish users that prefer to communicate in the medium on other languages, e.g. in English. Some studies have adopted methods such as location-based filtering to better account for geo-cultural and linguistic confounds. However, as was seen in Section 3.3, very few users provide geo-locations, partially limiting the confidence on the location of the user during the time of the action.

5.3 Future research

This thesis has created the foundation for analyzing crisis events and their effects on individuals and populations using temporal activity data. At the same time, various questions still remain open for future research - perhaps even more than before.

While Hurricane Harvey was record-setting in many ways, similar events, such as the disastrous Hurricane Katrina in 2005, have happened in the recent past, and are expected to happen also in the future, making it both important and suitable event type for further research. Therefore, first of all, a comparative study between areas such as Dallas that were not affected by the hurricane in similar magnitude as Houston was, provides a valuable opportunity for even better understanding on the effects of the crisis.

Secondly, as was discussed in Section 2.3.1, studies have found that behavioral gender differences are reinforced during the COVID-19 crisis. To further investigate this topic, the results on the magnitude of changes in activity patterns due to the pandemic could be deepened from population level into sub group level, in order to better understand how the pandemic has affected different groups of population, e.g. individuals from different genders or living areas.

Thirdly, the used NMF-approach provides a solid foundation for investigating users and their activity patterns on individual level. Potentially, the reasoning behind the main factors explaining different behaviour and their changes can be better understood by combining the activity data with the content of the tweets.

Lastly, it is worth researching how well the findings of this study transfer to other crises, and if they share some similarities to each other. This thesis might also encourage to extend the methodology to other directions and domains, such as other phenomena than natural crises.

References

[1] Aledavood, T. Temporal patterns of human behavior. *Aalto University publication series Doctoral Disserations*, vol. 227, 2017.

[2] Aledavood, T., Eduardo, L., Roberts S.G.B., Reed-Tsochas, F., Moro, E., Dunbar, R.I.M, and Saramäki, J. Channel-specific daily patterns in mobile phone communication. *Proceedings of ECCS 2014*, pp. 209–218, Springer, 2016.

[3] Aledavood, T., Kivimäki, I., Lehmann S., and Saramäki J. A Non-negative Matrix Factorization Based Method for Quantifying Rhythms of Activity and Sleep and Chronotypes Using Mobile Phone Data. *arXiv:2009.09914.* 2020.

[4] Aledavood, T., Lehmann, S., and Saramäki, J. Digital daily cycles of individuals. *Frontiers in Physics*, vol. 3, pp. 73, 2015.

[5] Aledavood, T., Lehmann, S., and Saramäki, J. Social network differences of chronotypes identified from mobile phone data. *EPJ Data Science*, vol. 7, i. 1, 2018.

[6] Attila, F., and Höglund, M. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer informatics*, vol. 6, pp. 275-292, 2020.

[7] Bevan, J., Cummings, M., Kubiniec, A., Mogannam, M., Price, M., and Todd, R. How Are Important Life Events Disclosed On Facebook? Relationships With Likelihood Of Sharing And Privacy *Cyberpsychology, Behavior, and Social Networking*, vol. 18, i. 1, pp. 8-12, 2015.

[8] Blake, E., and Zelinsky, D. National Hurricane Center Tropical Cycline Report: Hurricane Harvey (AL092017). 2018. Referred 1.11.2021. Available at: https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf.

[9] Blavatnik School of Government. COVID-19 government response tracker. Referred 8.11.2021. Available at: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker.

[10] Blomqvist, K. National Remote Work Survey - remote work during the corona crisis. Referred 1.11.2021. Available at: https://cocodigiresearch.files.wordpress.com/2020/05/remote-work-survey-covid-19_en.pdf, 2020.

[11] Borgatti, S. P., Mehra, A., Brass, D. J., and Labianca, G. Network analysis in the social sciences. *Science*, vol. 323, i. 5916, pp. 892–895, 2009.

[12] Brunet, J., Tamayo, P., Golub, T.R., and Mesirov J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, i. 12, pp. 4164-4169, 2004.

[13] Bueno-Notivol, J., Garcia, P., Olaya, B., Lasheras, I., Lopez-Anton, R., and Santabárbara, J. Prevalence of depression during the COVID-19 outbreak: A meta-analysis of community-based studies. *International Journal of Clinical and Health Psychology*, vol. 21. 2020.

[14] Cai, S., Georgakilas, G.K., Johnson, J.L., and Vahedi, G. A Cosine Similarity-Based Method to Infer Variability of Chromatin Accessibility at the Single-Cell Level. *Frontiers in Genetics*, vol. 9. 2018.

[15] Castaldo, M., Venturini, T., Frasca, P., and Gargiulo, F. The rhythms of the night: increase in online night activity and emotional resilience during the spring 2020 Covid-19 lockdown. *EPJ Data Science*, vol. 10, i. 7, 2021.

[16] Chen, T. H., Christopher, J. F. and Xu, X. The Effect of Disaster-induced Displacement on Social Behaviour: The Case of Hurricane Harvey. *Disaster-Induced Displacement*, pp. 1-13, 2019.

[17] Ding, C., Li, T. and Jordan, M. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45-55, 2010.

[18] Eurostat. How usual is it to work from home? Referred: 1.11.2021. Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20200424-1., 2020.

[19] Gaujoux, R. and Seoighe, C. A flexible R package for nonnegative matrix factorization, *BMC bioinformatics*, vol. 11, pp. 367, 2010.

[20] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. Youtube traffic characterization: a view from the edge. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pp. 15–28, 2007.

[21] Grabowicz P., Ramasco J., Moro E., Pujol J., and Eguiluz V. Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE*, vol. 7, i. 1, 2012.

[22] Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., and Tatlow, H. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 2021.

[23] Han, J., Kamber, M., and Pei, J. Data mining: Concepts and techniques, third edition (3rd ed.). *Morgan Kaufmann Publishers*, 2012.

[24] Huang, A.-L. Similarity Measures for Text Document Clustering. *Proceedings of Huang2008SimilarityMF*, 2008.

[25] Ivey, J. Hurricane Harvey: Trauma vs. Resiliency. *Pediatric Nursing*, vol. 43, i. 6, pp. 265-274, 2017.

[26] Jaskowiak, P., Campello, R., and Costa, I. On the selection of appropriate distances for gene expression data clustering. *BMC Bioinformatics*, vol. 15, i. S2, 2014.

[27] Järv, O., Tominga, A., Müürisepp, K., and Silm, S. The impact of COVID-19 on daily lives of transnational people based on smartphone data: Estonians in Finland. *Journal of Location Based Services*, vol. 15, i. 3, pp. 169-197, 2021.

[28] Kaltenbrunner, Gómez V., Moghnieh, A., Meza, R., Blat, J., and Lópen, V. Homogeneous temporal activity patterns in a large online communication space. *IADIS Int J WWW/INTERNET* , vol.6, pp. 61-76, 2008.

[29] Karami, A., Lundy, M., Webb, F. and Dwivedi, Y. K. Twitter and Research: A Systematic Literature Review Through Text Mining. *IEEE Access*, vol. 8, pp. 67698-67717, 2020.

[30] Kaudarer, S., and Randler, C. Differences in time use among chronotypes in adolescents *Biological Rhythm Research*, vol. 44, i. 4, pp. 601-608, 2013.

[31] Kostakis, O. Advances in Analysing Temporal Data. *Aalto University publication series Doctoral Dissertations*, vol. 63, 2017.

[32] Miritello, G. Temporal patterns of communication in social networks. *Springer, Science Business Media*, 2013.

[33] Mörttinen, M. Valtioneuvoston ydin kriisitilanteessa: Covid-19-pandemian paineet suomalaiselle päätöksenteolle. *Sitra memorandum*, 2021.

[34] Nasersharif, B., and Abdali, S. Speech/Music Separation Using Non-negative Matrix Factorization with Combination of Cost Functions. *Proceedings of the International Symposium on Artificial Intelligence and Signal Processing, AISP*, 2015.

[35] NYTimes. World's Coronavirus Infection Total Passes Staggering Figure: 200 Million. 2018. Referred 11.11.2021. Available at: https://www.nytimes.com/2021/08/04/world/europe/covid-global-cases-200-million.html.

[36] Oksanen, A., Savolainen, I., Savela, N., and Oksa, R. Psychological Stressors Predicting Increased Drinking During the COVID-19 Crisis: A Longitudinal National Survey Study of Workers in Finland. *Alcohol and Alcoholism*, vol. 56, i. 3, pp. 299–306, 2021.

[37] Oracle. An Enterprise Architect's Guide to Big Data: Reference Architecture Overview. Referred: 2.11.2021. Available at: https://www.oracle.com/topics/technologies/enterprise-architecture/entarch.html, 2016.

[38] Pekkarinen, L. Julkisen alan työhyvinvointi vuonna 2020. Referred: 15.11.2021. Available at: https://www.keva.fi/globalassets/2-tiedostot/ta-tiedostot/tyoelamapalvelut/julkisen-alantyohyvinvointi-2020-raportti.pdf, 2021.

[39] Rajkumar, R.P. COVID-19 and mental health: A review of the existing literature. *Asian J Psychiatr*, vol. 52, 2020.

[40] Reisch, T., Heiler, G., Hurt, J. et al. Behavioral gender differences are reinforced during the COVID-19 crisis. *Scientific Repoprts*, vol. 11, article 19241, 2021.

[41] Saha, K., Seybolt, J., Mattingly, S., Aledavood, T., Konjeti, C., Martinez, G., Grover, T., Mark, G., and Choudhury, M. What Life Events are Disclosed on Social Media, How, When, and By Whom. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[42] Saha, K., Torous, J., Caine, E., and De Choudhury, M. Psychosocial Effects of the COVID-19 Pandemic: Large-scale Quasi-Experimental Study on Social Media. *Journal of Medical Internet Resaerch*, vol. 22, i. 11, 2020.

[43] Sandler, R., and Lindenbaum, M. Nonnegative Matrix Factorization with Earth Mover's Distance metric. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1873-1880, 2009.

[44] Santamaría, L., and Mihaljević, H. Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, vol. 4, 2018.

[45] Shahbazi, Z., and Byun, Y.-C. Topic Modeling in Short-text Using Non-negative Matrix Factorization Based on Deep Reinforcement Learning. *Journal of Intelligent Fuzzy Systems*, vol. 39, no. 1, pp. 753-770, 2020.

[46] Shearer, E., and Mitchell, A. News Use Across Social Media Platforms in 2020. *Pew Research Center.* Referred: 15.11.2021. Available at: https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/pj_2021-01-12_news-social-media_0-04/.

[47] Shultz, J., and Galea, S. Mitigating the mental and physical health consequences of Hurricane Harvey. *Jama*, vol. 318, i. 15, pp.1437–1438, 2017.

[48] Statistics Finland. Official Statistics of Finland (OSF): Population structure. Helsinki. 2020. Referred: 10.11.2021. Available at: http://www.stat.fi/til/vaerak/2020/02/vaerak_2020_02_2021-05-28_tie_001_en.html.

[49] Statistics Finland. Suomen virallinen tilasto (SVT): Väestön tieto- ja viestintätekniikan käyttö. Helsinki. 2020. Referred: 10.11.2021. Available at: http://www.stat.fi/til/sutivi/2020/sutivi_2020_2020-11-10_tie_001_fi.html.

[50] Statistics Finland. Number of employed persons decreased from the year before. Published: 2nd March, 2021. Available at: https://www.tilastokeskus.fi/til/tyti/2021/01/tyti_2021_01_2021-03-02_tie_001_en.html.

[51] Wang, M., and Plumbley, M. Musical audio stream separation by non-negative matrix factorization. *Proceedings of the DMRN Summer Conference*, 2005.

[52] Wei, X., Xin, L., and Yihong, G. Document Clustering Based on Non-Negative Matrix Factorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 267–273, 2003.

[53] WHO. WHO Timeline - COVID-19. 2020. Referred 11.11.2021. Available at: https://www.who.int/news/item/27-04-2020-who-timeline—covid-19.

[54] Willberg, E., Järv, O., Väisänen, T., and Toivonen T. Escaping from Cities during the COVID-19 Crisis: Using Mobile Phone Data to Trace Mobility in Finland. *ISPRS International Journal of Geo-Information*, vol. 10, i. 2, pp. 103, 2003.

[55] Yasseri, T., Quattrone, G., and Mashhadi A. Temporal analysis of activity patterns of editors in collaborative mapping project of OpenStreetMap. *Proceedings of the 9th International Symposium on Wikis and Open Collaboration*, pp. 13, 2013.

[56] Yasseri, T., Sumi, R., and Kertész, J. Circadian patterns of wikipedia editorial activity: a demographic analysis. *PLoS ONE*, 2012.

[57] Yuan, Z. Advances in independent component analysis and nonnegative matrix factorization. *Dissertations in information and computer science, Helsinki University of Technology*, vol. 13, 2009.

[58] Zhang, H. Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis. *Aalto University publication series Doctoral Dissertations*, vol. 127, 2014.

[59] Zhang, N., Jia, W., Lei, H., Wang, P., Zhao, P., Guo, Y., Dung, C., Bu, Z., Xue, P., Xie, J., Zhang, Y., Cheng, R. and Li, Y. Effects of Human Behavior Changes During the Coronavirus Disease 2019 (COVID-19) Pandemic on Influenza Spread in Hong Kong, Clinical Infectious Diseases. *Clinical Infectious Diseases, ciaa1818.* 2020.

[60] Zitnik, M., and Zupan, B. Nimfa: A python library for nonnegative matrix factorization.. *Journal of Machine Learning Research*, vol. 13, pp. 849-853, 2012.