KU LEUVEN

FACULTY OF SCIENCE

Using Digital Trace Data to Generate Representative Estimates of Disease Prevalence [COVID-19 Infections] in Belgian Municipalities

Masters Thesis

Dishani Sen

Supervisor: *Roberto Cerina*
Assistant Professor,Maastricht University.
Co-supervisor: Emmanuel Lesaffre
Leuven Biostatistics and Statistical Bioinformatics Centre.

Thesis presented in
fulfillment of the requirements
for the degree of Master of Science
in Statistics and Data Science

List of Figures

List of Tables

Acknowledgements

This thesis is written in collaboration with Sciensano, Belgium.
I would like to thank Program Director Ingrid Van Keilegom and Program Coordinator An Carbonez for providing me with this opportunity.

I particularly extend my sincere thanks to Prof. Dr. Sofie De Broe, Scientific Coordinator of the Strategy and External Positioning unit at Sciensano.

Many thanks to my supervisor and mentor R. Cerina Assistant Professor in Computational Statistics, Maastricht University for his time and advise during the writing of this master thesis. I could not ask for a better guide.

I also want to express my gratitude to my co-supervisor Emmanuel Lesaffre and Professor Geert Molengerghs for the insightful guidance during the mid-term presentation and suggestions on refining the analysis.

I thank scientists from Sciensano, Toon Braeye and others present during the Modelling Series Webinar for useful comments and critical reflections on the analysis strategy.

Finally, my gratitude goes to my family and friends for their continuous support and encouragement.

Dishani Sen,
Leuven, Belgium.

Abstract

Is it possible to predict the area-level prevalence of COVID-19 infections in Belgium by analyzing self-reported symptoms on Twitter? This research project is about generating estimates of the incidence of COVID-19 infections, at the municipality level, by using Multilevel Regression Post-Stratification (MrP) to account for sampling biases in the social media sample. At first, tweets are obtained from users based on keywords derived from previous research, e.g., tweets mentioning fever, cough, loss of taste, fatigue, etc. Then, key demographic and geographical features of interest are extracted using the M3 deep learning pipeline, as well as simple self-reported characteristics, effectively transforming the unstructured twitter sample into a survey-like object. Finally, based on these demographic features and census characteristics, a mixed effects logistic regression model with post-stratification according to the Belgian census is proposed to forecast the number of infected individuals on a particular day. This study intends to contribute to the proof of concept of a complete end to end pipeline to perform real time predictions of disease prevalence at a granular level in a population using social media data.

Keywords: COVID-19, Digital-Trace, Feature Extraction, Twitter, Bayesian Hierarchical Modelling, MrP

Contents

Chapter 1

Introduction

COVID-19 pandemic has changed our lives forever. The outbreak of 2019's disease coronavirus (COVID-19) is one of the worst recorded in history [33]. 510,270,667 confirmed positive cases had been reported globally as of April 29, 2022, resulting in 6,233,526 confirmed deaths [3] . With the spread of the pandemic over two years now, several research projects have taken place and are underway, ranging from the testing of potential vaccines to forecasting the outbreak's progression to assess the virus's features by studying infected patients.

Early research studies designed to determine the symptoms experienced by patients infected with the virus largely comprised mostly hospitalized or clinically treated patients[14]. Many infected persons have moderate symptoms or are asymptomatic and do not seek medical attention, while the exact number of asymptomatic carriers is still unclear[7].

According to the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, the true global death toll is more than double the reported figures [11]. Countless people that die while contaminated with SARS-CoV-2 are never tested for it, so their counts are not included in the official totals [37]. Therefore, it is nearly impossible to investigate all the symptoms of the infection by relying only on health records.

There are more people who have been infected with COVID-19 virus than it is ever recorded. Many people who are also infected by the virus, do not get officially recorded because of many reasons; they could be asymptomatic, or they never get tested formally. To better understand the full spectrum of prevalence of COVID-19 and the symptoms experienced by infected people and make further inferences regarding the spread of the infection among people, there is a need to look beyond hospital- or clinic-focused studies.

One of the appropriate means to gain insights into the proliferation and spread of diseases within a population is through surveys. Despite popular belief that surveys are simple to perform, they require comprehensive preparation, time, and effort to produce reliable results. Even with a modest sample size, offline questionnaire surveys are expensive, and the expenses of a typical large-scale survey employing postal questionnaires may be tremendous [41] [20] . Moreover, given the global pandemic situation, it is practically impossible for enumerators to go door to door and collect information regarding every cases of COVID-19 infection. The usage of online surveys avoids this issue of excessive cost by removing the requirement for paper and other expenditures but there are several drawbacks that has to be considered before

adopting online survey methods such as difficulty getting back response, misinterpreted survey questions, survey fatigue, long wait time before receiving response and many more.

Therefore, conducting surveys, either offline or online, is not the most ideal way to gather information about a population regarding whether it's infected by COVID-19 or not. In this research study, an alternate way is proposed to create survey-like objects from digital-trace data. Digital trace data is defined as the untapped resource of mass scaled contextual information that is available on social media platforms, like Twitter. On Twitter, online users post tweets, about their experiences, beliefs, attitudes regarding any topic of their choice, which is digital-trace data. The everyday opinions expressed in social media provide promising opportunities for measuring population-level statistics for health metrics, political outcomes, or general attitudes. In this research, the application of social media data for inferring about health metric, more specifically, disease prevalence is chosen.

However, social media data is often unorganized, and a non-representative sample of the population due to demographic skew in usage frequencies and access rates. As such, any direct estimate from a platform like Twitter is likely biased towards certain demographics. For example, it is commonly believed that the younger generation is more active on social media platforms, whereas the older population as less to no presence online. To account for the same, the most suitable methodology has been adapted in this study:

- The digital trace data is converted into survey- like objects after integrating with key demographic features and location attributes.
- The survey like objects is then combined with a stratification frame of the Belgian census to fit a mixed effects multi-level regression model with post-stratification and utilize it for generating estimates that are as representative as possible.

For this study, Twitter is used as the primary source of information to gather data about cases of infection, reported by users online. This data is then converted to survey-like objects to then utilize for generating representative estimates about disease prevalence in the population. The main contribution of this research comes from two key parts: a) The methodology by which mass-scaled data from Twitter can be converted into survey-like objects, that looks exactly similar to data gathered from a real survey, and b) using a mixed effects multi-level regression model with post-stratification to make predictions about the COVID-19 prevalence made from these survey-like objects. The population of choice is residents of the country of Belgium. This is a step towards generating automated estimates of COVID-19 prevalence across Belgium, where data is gathered beyond public health institutes, from social media, mainly Twitter. In this study, the goal is to present a complete end-to-end pipeline for the proof of concept of how, digital trace data (data from social media), that is available beyond official health records, can be used as an inexpensive source of valuable information to monitor, and determine progression of disease over a population.

Sciensano, the Belgian health institution, is in charge of the epidemiological follow-up of the COVID-19 epidemic in collaboration with its partners and other healthcare entities. The data gathered can give insights into the development of the epidemic, assist in forecasting alternative situations, and aid to develop potential methods to halt the virus's spread. [2].However, to the best of knowledge, this is the first study to focus on extracting COVID-19

symptoms from public social media in Belgium, at the granular level of its municipalities, and attempting to convert the unstructured data from Twitter into a structured survey-like object that can then be further utilized to make inferences about the municipality level spread of COVID-19 infections in Belgium. Furthermore, there hasn't been much work done with Belgian data using MrP, so there's ample opportunity for innovative and exciting work. This study aims to bridge some of that gap and encourage other scholars to pursue similar groundbreaking research.

The progression of the research is further outlined in the following sections - Chapter 3 overviews the Literature Review. Chapter 4 discusses the problem statement and the research questions for the study. In Chapter 5, the methodology is elucidated. It has the following sub-sections, data collection, feature extraction and survey like objects. In the next Chapter, 5, modelling strategy is covered. In Chapter 6, the study's findings are presented and finally in Chapter 7 discussions are expressed.

Chapter 2

Literature Review

Demographers, data scientists, and public-health professionals are working to reduce the uncertainty in pandemic estimates globally. Both academics and journalists are contributing to this endeavor, which use methods ranging from satellite photos of cemeteries to door-to-door surveys and machine-learning computer models that attempt to extrapolate worldwide figures from existing data [4]. New approaches such as digital epidemiology [32] or info-demiology [12] offer another option for tracking patterns of health and disease through digital data, including data from keyword search engines (e.g., Google) [35].

Indeed, academics have lately begun to utilize the content of Twitter tweets to assess and forecast real-world events such as movie box office returns [6], elections [28], and stock market returns [9]. These studies reveal that tweets posted by individuals online have a great deal of potential and may be used for a range of applications.

Therefore, one alternative methodology to attempt to predict the course of a pandemic can be to look at social media, namely Twitter, where symptoms are self-reported by users who tested positive for COVID-19. It has been found that there were several self-reporting of COVID-19 confirmed cases on Twitter, despite the fact that such reports were buried in a sea of noise. A prevalent tendency among Twitter users in which they described their day-to-day illness progression since the commencement of their symptoms. It was also confirmed from many people who claimed to have tested positive but had no symptoms [34]. According to one study, the frequency of Google Trends on anosmia, or loss of smell, is associated to the emergence of COVID-19 cases in different nations [38]. Another study discovered that COVID-19 keywords peaked 10–14 days before the peak in occurrence in social media and internet searches. It also confirmed that early and low-cost access to reliable data is possible through social media about the outbreak and progression of COVID-19 infections [24]. Social networking sites seem to have been widely used by the general public in keeping contact with friends and family in order to minimize loneliness and boredom during the pandemic [10]. One of the key advantages of using social media to investigate symptom patterns and clusters for COVID-19 is that it is a prompt and economical data source. [38].

Plentiful digital trace data that is freely available to academics via the Twitter API [8], can be utilized to make statistical inferences about disease prevalence. In this research, it is attempted to account for self-reported symptoms from users on the social media platform, solely Twitter. Users report symptoms online via their tweets which suggests that cases of

infection can be detected via identifying the tweets that involve certain keywords that are closely related to COVID-19 symptoms.

The Centers for Disease Control and Prevention (CDC) recorded just three major symptoms of COVID-19 at the start of the pandemic in March 2020 : cough, fever, and shortness of breath. As the pandemic progressed, new COVID-19 symptoms were detected and added to the CDC's list of COVID-19 symptoms. The CDC added chills, recurrent shaking with chills, muscular discomfort, headache, sore throat, and new loss of taste or smell to the list of COVID-19 symptoms in late April 2020. In May 2020, the Centers for Disease Control and Prevention added numerous new symptoms, including fatigue, congestion or runny nose, nausea, vomiting, and diarrhea. During the early phases of the pandemic, tweets cited all of the symptoms advised by the CDC for COVID-19 screening in March, April, and May of 2020. It was also discovered that numerous COVID-19-related symptoms were noted in Twitter tweets prior to the CDC's release, raising the possibility that monitoring social media data as a viable strategy to public health surveillance [15].

But there are two key challenges to making use of the data collected from Twitter in the production of representative statistics:

- The data is relatively unstructured (high in 'Variety' [31]), meaning that information about one's age, for instance, can be found in a users' profile image, but also in the language and topics used in their comments, and in their social network, as well as in other public observable fields - but it is almost never explicit, i.e. rarely will users typically state their age in a way that is systematically retrievable.
- The data suffers from dramatic selection effects [26], which means it needs to go through significant model-based adjustments [22] to be useful. The data is highly non-representative of the real population. In research from [36], it is found that Twitter users in the United Kingdom, for example, are often more likely to be male and younger than the general population. Men and inhabitants of highly populated regions are over represented in the United States [17], whereas there is a combination of over- and under sampling on Twitter for people with certain ethnic origins [27].

The challenge with respect to non-representativeness, i.e., adjustments for, for example, age and gender is tackled by proposing to use the deep learning model called M3 inference pipeline to extract individual-level demographic characteristics of users from social media, to generate survey like-objects. Advances in Deep Learning have enabled relatively accurate feature extraction from social media data. A novel method is proposed by [40]) for assigning users to demographic strata and exploiting it in a regression framework for de-biasing that allows direct estimation of the probability of an individual with given demographics to be on the given social media platform. M3inference attempts to determine the age, gender, and if an account is a business account or not using only four pieces of information from your public Twitter profile: The username, the display name, bio (which is a short descriptive text for each user profile), and avatar (profile) image. The resultant stack of data collected will be replication of data collected from a survey, whereas, in reality, it is originated from the massive, untapped contextual information that is available on social media (Twitter). These survey-like objects are then finally amenable for statistical inference.

One of the primary goals of statistical technique is to make inferences about the population. Multilevel regression and post-stratification (MRP) have been found to be an effective approach of adjusting the sample to be more representative of the general population [25] [23] [29]. Recent advances in statistical tools, namely the introduction and wide-spread adoption of Multilevel Regression and Post-Stratification (MrP) [30], have enabled to account for highly complex selection effects into non-representative convenience samples. Survey analysis researchers have dealt with non-representative polls through sample re-weighting; with post stratification being a well-known technique [13] [19].

However, post-stratification is a useful strategy if demographic information is available [40]. MrP makes use of basic demographics attributes: most prominently age, gender and location [29] that are derived from the M3 pipeline. Where demographic details are provided, post-stratification turns out to be valuable technique [39] [42].Therefore, the M3 inference pipeline solves two problems:

- It helps transform the unstructured social media data into structured survey like objects, that resembles original data collected from a survey.
- Because statistical techniques like MrP can only utilize structured data for further analysis, M3 pipeline ensures that the social media data can further be utilized for modelling and predictions.

The aim of this project is to put the two methodologies of M3 inference and MrP together and attempt to make representative inference from highly unstructured non-representative samples of Twitter digital trace data. The application of choice is disease prevalence of the COVID-19 pandemic in Belgium. Because the data from Twitter is relatively unstructured, meaning that information about one's age and gender is not clearly evident, the deep learning model, M3 pipeline have been used to systematically retrieve the demographic features of users, whose tweets have been collected. Since by using the M3 pipeline, demographic information is made available, therefore, relevant adjustments for selection bias can be provided via classic MrP, essentially treating the extracted features as data coming from a survey.

The primary goals are to transform unstructured digital trace data into structured survey like objects amenable to statistical analysis:

- collect mass-scaled digital trace data (tweets) where the users report their experiences with COVID-19—including the symptoms experienced—from Twitter,
- assign a location value to tweets for a granular level analysis, in this case municipalities of Belgium?,
- propose a deep learning model, M3 pipeline, to extract demographic features of the users which are important for the outcomes for COVID-19 pandemic (e.g., Age, Gender), and
- propose a fixed effects multilevel regression model with post-stratification to make real time predictions.

In the context of Belgium, thus, an end-to-end pipeline with multi-level regression and post-stratification modeling technique to solve the challenges of obtaining representative estimates of disease prevalence at the area-level using social media data is suggested. The objective is to precisely forecast the number of COVID-19 infections for a particular day by using social media data turned into survey-like objects along with the previous day's COVID-19 cases count.

Chapter 3

Problem Statement and Research Questions

Social media data have a potential application in the early identification novel virus symptoms in digital epidemiology. It has been verified previously that many COVID-19 infection related symptoms were noted in tweets on Twitter even before the official release by CDC. Therefore, there's an optimistic future in the possibility that monitoring social media data as a viable strategy to public health surveillance. With that in mind, in this thesis, an attempt is made to utilize social media data, i.e., tweets (digital trace data) to make inferences about the granular level prevalence of COVID-19 infections in Belgium.

Due to social media, (Twitter, within the scope of this research), we have access to archival data from Twitter. Thus, in this study, the digital trace data is in the form of tweets, that are posted by users online on the social media platform Twitter. The objective of this research study is how can this digital trace data, which by its characteristic is unstructured, non-representative and biased, be utilized to make inferences about the COVID-19 pandemic in Belgium. To refine the research further, a granular analysis is presented, at the municipality level in Belgium.

Thus, the broad research question is: Is it possible to make representative predictions about the granular level prevalence of the COVID-19 pandemic (in Belgium) by aggregating tweets from different municipalities of Belgium?

Several research questions emerge from this overarching purpose to be addressed in this study:
RQ 1: How is the mass scaled digital trace data (tweets) collected?

The digital trace data, in this case the tweets need to be collected in an orderly manner to enrich the primary data source for the rest of the research, by considering all the relevant keywords that are associated to users indicating being infected by the COVID-19 virus. This has to be done with the idea of maximizing the count of tweets collected for a greater degree of inclusivity of all cases, where users report online about experiencing COVID-19 symptoms in the research on an appropriate time-frame. To consider for Belgium, which is a multi-lingual country, three languages are taken into account: French, Dutch, and English, to maximize the data collection.

After the tweets are collected systematically, the next step is to prepare the data in such a way that further statistical methods can be performed.

RQ 2: How is the unstructured digital trace data transformed into structured survey like objects amenable to statistical analysis?

Because the biggest drawback of the data collected from Twitter is that it is unstructured and non-representative, i.e., demographic information like the users' 'age' and 'gender' are not distinctively handy, further statistical analysis on them is constrained. So, the first and foremost step is to counter that. Further this research question is broken into sub-parts:

RQ 2.1 How can we assign a location value to tweets for a granular level analysis?

In this study, the attempt is to not just make inference about Belgium as a country, but to be able to speak in a much deeper level. Thus, the chosen bifurcation is municipalities in Belgium. Against every tweet collected from an user online, a corresponding location value of municipality is assigned, depending on where the user has tweeted from. For simplicity, it is considered that the location from which the user has tweeted is his/her natural location of habitat, i.e., he/she stays in that particular municipality of Belgium.

RQ 2.2 How can we extract demographic features of the users which we know are important for the outcomes for COVID-19 pandemic (e.g., Age, Gender)?

In the next step, for every user, demographic information has been extracted using a deep learning model, M3 inference which is a robust technique to infer information like age and gender accounting for multiple languages from users' profiles' information on Twitter.

By doing so, the data gets a structured form and exactly looks like being collected from a survey. Thus, it is now called survey-like objects.

RQ 3: How can these biased survey-like objects be utilized for generating representative real-time estimates of COVID-19 cases at municipality level in Belgium?

The final goal of the study is to propose modelling strategy to make real-time predictions of prevalence of the COVID-19 pandemic at municipality level in Belgium. The survey-like objects are structured, but they still suffer from strong selection bias and are non-representative of the actual population. The concepts of multi-level regression with post-stratification are adopted on the survey-like objects as it a well-known technique when demographic details like location, age and gender are available to speak for the population.

The overall hypothesis for this research is as follows:

Hypothesis: The area level prevalence of the COVID-19 pandemic (in Belgium) at its granular level (municipalities) can be modelled by Multilevel Regression and Post-Stratification (MrP) on features extracted (like age and gender) from aggregated tweets of users from different municipalities of Belgium to make real-time predictions and generate representative

estimates. The results of Multilevel Regression and Post-Stratification (MrP) is similar to actual data of prevalence of COVID-19 infections in Belgium.

In this study, an attempt is made to build a complete pipeline that includes collection of social media data, converting it into organized survey like objects and finally fitting a mixed effects multi-level regression model with post-stratification to forecast COVID-19 cases on a particular day at municipality level in Belgium. Contribution to the three core elements of are being made: the collection of mass scaled tweets, the extraction of demographic features and assigning a location value to convert unstructured digital data to survey like objects, and, using a multi-level regression model with post-stratification to make real-time predictions on the population using the digital trace data.

Chapter 4

Methodology

This study is exploratory in nature. The purpose of this research work is to investigate how digital-trace data can be utilized in order to make inferences about, a disease (in this case, COVID-19 infections) and make it amenable to the statistical analysis to further make estimates about a population of choice.

More specifically, it is aimed to provide a complete proof of concept implementation to demonstrate the feasibility of using social media data for representative predictions of spread of disease in a population of choice. That is achieved through exploring how digital trace data, i.e., social media data can be collected for this purpose, and how this type of data is processed to further statistically model it to generate real time representative estimates to contribute to understanding of the spread of the disease and provide valuable insights into understanding its impact on the population.

The digital-trace data in context of this study is tweets from the social media platform Twitter. The phenomenon of choice is disease prevalence as it has been recently discovered that social media has a huge potential application in health monitoring. The disease of choice is COVID-19 infections, as it has been proven before by previous research studies that users on Twitter have posted about being infected by the infection and also professed experiencing symptoms. The population of choice is Belgium. To get a granular level prediction, the study makes inferences at the municipality level of Belgium.

In an attempt to achieve the overall goal, there are three core components in this study. The first component is about collecting the digital trace data, which are in this case tweets. The second component is where the collected data is prepared for further statistical analysis. In this step, the unstructured digital data is converted into survey like objects. And the final component is where the survey-like objects are modelled to make real time, representative estimates about the COVID-19 pandemic's prevalence in Belgium, at a granular level of its municipalities.

4.1 Data Collection

For this study, the source of data is primary. The data in concern is digital-trace data. In the context of this study, digital trace data are tweets that are posted by users online, on the social

media platform, Twitter. Twitter is one of the most used platforms for social science research because of its ease for obtaining data. In this research, Twitter is the primary source of data.

Tweets posted in three languages: English, Dutch and French, related to COVID-19 infection and its' symptoms between March 01, 2020 and February 29, 2022 using search terms of COVID-19 synonyms and common COVID-19 symptoms recommended by the CDC were collected. The time frame chosen is appropriate as it takes into account tweets from the initiation of the pandemic's spread in Belgium, i.e., early March 2020.

For the purpose of this research, the keywords of choice are based of research and exhaustive list of symptoms provided by the CDC. Tweets are collected from the beginning of the global pandemic of COVID-19, from users, who are based in Belgium whose tweets have evidence of reporting having the infection. To account for the second and third wave of the pandemic, the keywords are inclusive of the two later variants of COVID-19, the delta variant and the omicron variant.

Tweets including important COVID-19 phrases along with phrases for the different variants of the virus (e.g., corona, covid, delta and omicron) as well as at least one of the symptoms are collected. This is done to ensure that only tweets that indicate the user is distinctively mentioning that he/she or someone they know is experiencing the symptom and has been infected. The tweets that have been collected using the infection-related keyword and the self-reported symptoms are treated as "cases" where there is evidence of its corresponding user being infected by the COVID-19 virus. Each keyword is carefully designed to corresponding to one symptoms of the infection. The following table shows the different keywords that has been used in the four different languages:

Some remarks regarding the keywords used are as follows:

- The keyword "covid" accounts not only for the exact word covid, but it also covers keywords like COVID-19 or covid-virus. Similarly, the keyword corona accounts for related words like coronavirus.
- In the style users write tweets online, it is extremely difficult to completely replicate the way the symptoms experienced by the user is expressed in his/her tweets. Therefore, certain tweaks in the keywords are done to capture tweets related to a particular symptom of the infection. For example, the keyword "taste" is used to capture tweets where the users maybe talking about the symptoms of the infection "loss of taste" and the keyword "smell" is used to entail cases where the users are experiencing the symptom "loss of smell". Pain/Ache covers symptoms like headache, stomach pain.

Some examples of tweets that have been collected are as follows:
"I got a **Covid** test today because of a slight **fever**. I think it will be something else, which I have an appointment for tomorrow but it still feels weird. I prefer working from home with the kids to being ill with them by a long stretch. I can't postpone being ill until they sleep."

Symptom	Keyword in English	Keyword in French	Keyword in Dutch
Fever	fever	fièvre	koorts
Fatigue	fatigue	fatigue	vermoeidheid
Cough	cough	toux	hoest
Sneeze	sneeze	éternuer	niezen
Body- ache	hurt	blessé	zeer
Body pain	pain	douleur	pijn
Loss of smell	smell	odour	geur
Loss of taste	taste	goût	smaak
Sore throat	sore	irrité	gevoelig
Runny nose	nose	Nez	neus
Chills	chill	froid	koude
Sweating	sweat	transpiration	zweet
Loss of smell	smell	odour	geur
Insomnia	insomnia	insomnie	slapeloosheid
Dizziness	dizziness	vertige	duizelig
Vomit	vomit	vomir	overgeven
Weakness	weak	faible	zwak

Table 4.1: Description of symptoms of infections and the respective keywords used for data collection

"3 days ago, we found out that our son's baby-sitter tested positive for **corona**. As awful as it may seem, my son is also teething, he had a **fever, diarrhea, colds,** and lots of **irritability**"

Tweets that met the above mentioned criteria of time-frame and keywords were retrieved using R software (version 3.6.3) and the Twitter Application Programming Interface (API) , via the R library **"academictwitteR"**. The tweets made by users online have been collected with the help of Twitter API for Academic Research.

The Twitter Application Programming Interface, or API, debuted in 2006. It was created primarily with business goals in mind. However, over time, scholars began to rework the Twitter API for scholarly purposes. Twitter unveiled the "Academic Research Product Track" in January 2021, which allows for significantly improved data access.

Twitter launched the "Academic Research Product Track" in January 2021. This dramatically expands academics' access to Twitter data. Existing R packages for accessing the Twitter API, such as the renowned **rtweet** package [21], have yet to include capabilities that would let users to access to the new v2 API endpoints using Academic Research Product Track credentials.

Twitter officially gives access to its new v2 Twitter API for academic research using which one can use additional tools and capability to acquire more exact, full, and impartial data-sets from Twitter's real-time and historical public data. The maximum limit of collection of tweets is 10 million Tweets per month. It is a free service, with an extensive verification process by Twitter to ensure that the access is provided only to legitimated researchers who are associated with

an academic cause. The access to this API is only for non-commercial use cases [1].

Christopher Barrie, Lecturer in Computational Sociology at University of Edinburgh and former PhD student Justin Chun-ting Ho have released the R package **"academictwitteR"** to collect data from the Twitter API for academic research. The **"academictwitteR"** package [8] is designed for academic research. It gives R users access to the entire set of parameters provided by the new v2 Twitter API, allowing researchers to search tweets by user, tweet content, and several more factors (e.g., excluding retweets, searching by tweets containing media or images, and searching by the location of the user or tweet).

The **"academictwitteR"** package has been designed to facilitate ease of collection of tweets providing access to full-archive search without any time constraints [8]. It's better than any of the other pre-existing R package because they only provided limited access to the Twitter API.

For the purpose of this study, **"academictwitteR"** is the most suited library for a number of reasons. First, it is essential to have access to archived tweets to study the prevalence of COVID-19 infections in Belgium. This library allows to access tweets dating back to March 2020 without any constraints. Secondly, there is no limit on the number of tweets that is permissible to be collected. This ensures that all the tweets matching the criteria of keywords and time-frame can be utilized for the research study. Thirdly, promoted tweets can be removed to ensure that the research study is considering only cases where users self-report their symptoms, rather than any commercially fueled tweets. And the final reason, is the greater flexibility that "academictwitteR" offers in terms of pinpointing the location from which tweets are collected. This is discussed in detail in the section under Feature Extraction.

No promoted tweets have been included in the research by discretely using the function remove_promoted=T in **academictwitteR**. Retweets have been removed to eliminate double counting of the same tweet, using is_retweet=F in **academictwitteR**. Only unique tweets were extracted for the analysis.

Alongside tweets that indicate cases of COVID-19 infections, a reference group of controls is also collected. In order to correctly estimate the regression coefficients, i.e., to estimate the effect of gender on the chances on being infected, controls are needed. The way the controls are obtained here are by sampling tweets based on keywords that are completely unrelated to being infected.

The calibrated number of controls are to exactly the same as the number of cases (of being infected). However, having the same number of cases and controls biases the intercept of our model (likely upwards), because in reality the probability of being infected is very less, almost close to none, it is still useful to estimate the regression coefficients on age, gender and contextual information (area-level predictors) appropriately. By comparing the cases where people experienced the infection with another random sample of people, who have not, inferences and predictions are made by statistical techniques followed in this research.

For the reference group of controls, a set of neutral keywords have been chosen, such that it has no intentional correlation with keywords related to the COVID-19 infections and its symptoms. The chosen keyword is "happy" in English in the same time frame as mentioned

above. Tweets are also collected in dutch and French language, with keywords “vrolijk” for dutch and “heureuse”, and “heureux” and for French.

It is essential to note that the data collected from Twitter is still unstructured, and non-representative.

4.2 Feature Extraction

To make the primary digital trace data amenable to further statistical analysis like post-stratification, important features like location, age and gender are prerequisite. Thus, the next step is to extract those features and convert the cases into structured survey-like objects. The resultant data is a replication of results from a survey, containing, the age, gender, location and the date on which the particular individual has experienced symptoms.

In feature extraction, a location value is assigned to each of the cases, based on where the user has tweeted from. This is the first step in converting the unstructured data of cases where users have reported online that they are infected by COVID-19 into structured survey like objects. The next step is to assign demographic information to each of these cases. At first, a location value is assigned to each of the cases. Followed by it, demographic information like age and gender are assigned to each of the cases. Then finally, the unstructured mass digital trace data, implying cases of COVID-19 infections get converted into survey like objects, where each case, has a location, age, gender and time variable to corresponding case. Because this information replicates coming from a real survey, but is derived from social media data, they are called survey-like objects.

4.2.1 Location

In the context of this study, the location of interest is municipalities of Belgium. Belgium is a country with an area of 30,688 km^2. It is primarily divided into three regions namely: the Flemish Region, the Brussels-Capital Region and the Walloon Region. Additionally, the regions are sub-divided into provinces. There are in total 10 provinces in Belgium. Further, Belgium has 581 municipalities, 300 of which are split into five provinces in Flanders and 262 of which are divided into five provinces in Wallonia, with the remaining 19 located in the Brussels Capital Region, which is not separated into provinces. Municipalities are, in most situations, Belgium’s smallest administrative units [5].

For the purpose of this study, where the aim is to provide a proof of concept at a granular level, inferences are chosen to be made at the municipality level. In Belgium,municipalities are the smallest administrative geographic levels and thus are of interest in this study. MrP makes high-quality predictions by leveraging correlation between area level divisions. In other words, MrP re-weights information available for other regions based on crucial socio-demographic data to allow estimation even when very little is known about other regions.

There is evidence in the MrP literature of the relative benefits of applying this approach in the context of opinion estimation where at least 50 minimum small area units are present, verified by research on 50 states in the US [30] [23]. Findings by [16] confirm that, in settings with a

large number of areas, researchers can possibly generate reliable estimates even if detailed post-stratification information are unavailable. Indeed, many nations have a significant number of tiny administrative districts, and their corresponding information gathered via area-level predictors may benefit to get reasonably better estimates when using MrP. Thus, for the Belgian study, municipality level is chosen.

The **"academictwitteR"** package on R provides great flexibility in terms of control over the location from which tweets can be retrieved, among other things. The "point radius" argument of **"academictwitteR"** package has been utilized to regulate collection of data from every municipality of Belgium. Every municipality is conceptualized to be a circle in shape and tweets have been accumulated from every municipality, using the "point radius" argument, which takes, the latitude, longitude and radius as inputs. These information about the latitude, longitude and radius corresponds to the center point of each municipality. The R libraries, **"BelgiumMaps.StatBel"** which is contributed by StatBel Belgium and "sf" have been utilized to find the centers of each municipality in Belgium. The radius for each municipality is calculated by the formula radius= sq root (area of each municipality/pi) units.

It is essential to make a remark here, that in this attempt, there can be some edge cases, where the assignment of municipality to each tweet (which are cases in the context of this study) is not similar to the municipality in which that case actually belongs in reality because of the strict assumptions that each municipality is circular in shape. However, the idea is to, consider as many tweets, and cases as possible, and corresponding to it, add a location (municipality) value. This mis-classification is expected to not severely affect the analysis because of the similarities of people living in nearby municipalities.

4.2.2 Demographic features like age, gender

Because most social media sites do not publish demographic information about their users, it is difficult to identify and remedy for such biases. Despite the fact that interesting strategies for extracting demographic information from social media have been developed, no comprehensive multilingual approach to this job exists. This problem is solved by employing M3 inference pipeline [40] to extract fundamental demographic information. On a large sample of multilingual profile data from Twitter, the M3 inference pipeline model is adjusted for selection biases, making it the best available approach for gathering demographic data to turn digital trace data into survey-like objects.

M3 inference pipeline, which is a multilingual, multimodal, multi-attribute deep learning system for inferring demographics of users has been utilised to extract demographics information like age and gender of the cases. M3 inference pipeline is a deep learning system trained on a big Twitter data-set for demographic estimation. It has three primary characteristics: first, it is multimodal which means it accepts both visual and text input. The inputs are a profile image, a name (e.g., a natural language first and last name), a user-name (e.g., the Twitter screen name), and a brief self-descriptive text (e.g., a Twitter biography). Secondly, M3 is multilingual, operating in 32 main languages. And finally, M3 is Multi-attribute which means that the model can predict three demographic attributes at the same time (gender, age, and human-vs-organization status).

M3 inference pipeline works by retrieving a user's Twitter username, display name, bio, and avatar image and feeding them into a neural network. It processes the avatar using DenseNet and the text with three LSTMs and then adds two additional neural network layers to integrate the predictions of all four networks into a single forecast.

A DenseNet is a style of convolutional neural network that uses dense connections between layers via a module called Dense Blocks, in which all layers are directly connected with one another. To maintain the feed-forward nature, each layer receives extra inputs from all previous layers and passes on its own feature-maps to all following layers. Dense Connections, also known as Fully Connected Connections, are a form of layer in a deep neural that employs a linear operation to connect every input to every output through a weight.

Long Short Term Memory networks (LSTMs) are artificial neural networks that are employed in the disciplines of artificial intelligence and deep learning. LSTMs are specifically designed by [18], to avoid the problem of long-term dependency. It is basically their default behavior to remember information for lengthy periods of time.

M3 inference pipeline has been used on Python 3.6.6 via the package "m3inference" to gather the demographic information. The output from M3 inference is in probabilities (with 0 to 1) corresponding to four classes of age; age less than equal to 18, age 19-29, age 30-39 and age more than equal to 40, two classes of gender; male and female, and two classes of organisation; non-organisation and is organisation. The class with the highest probability for each attribute is the predicted class.

Despite the fact that interesting methods for extracting demographic information from social media have been developed, no comprehensive multilingual approach for this job exists. Moreover, in a large-scale analysis of Europe, M3 inference pipeline model can reliably infer regional population counts and give demographic adjustments to selection biases. This research is based on a European country, Belgium and the related tweets have been collected in three languages, so a multi-lingual as well as robust technique like M3 inference is best suited for the task of feature extraction.

During validation of M3 inference pipeline it was found that the Positive Predicative Value is 96.7% and the false positive rate is 4.2 %.This means that for the 1296 records in our survey-like objects, 96.7 % individual are classified into correct categories of age, gender and whether organization or not.

4.3 Survey-like objects

The final survey like results from Twitter data and M3 inference output looks exactly like data collected from a survey. In the results from M3 inference, only non-organizations, i.e., human cases are taken into account.

There are total 1296 records in the survey like object, out of which 521 cases (40.2%) has infection and the rest 775 controls, who do not have infection (has_infection $= 0$).
Even though, essentially the goal was to maintain a 50-50 proportion of cases and controls,

this mismatch is a result of the intermediate data-cleaning stage where two issues were mainly dealt with.

- Records with missing value were omitted. Missing value appeared from the classification result of the M3 Inference pipeline.For certain records, the same could not classify the resultant Gender, Age-group or Organisation status of the corresponding entities. Thus, such records were removed from while aggregating the survey-like objects.
- Duplicate records were omitted. There were instances, particularly for "cases", where multiple symptoms have been experienced by an individual which reflected in their tweets. Thus, the same tweet was captured numerous times because of the different keywords. In such situation, only one such record has been considered in the survey-like object.

In the following table the attributes of the survey like objects are described:

Attributes	Values	Details
Gender	F or M	F stands for female, M stands for Male
Age	Age <= 18, Age 19-29, Age 30 -39, Age>=40	The class of Age group
has$_{i}nfection$	0 or 1	0 means “does not have infection”, 1 denotes “has infection”
Municipality	Name of municipality	The municipality to which the case/control belongs
CD_MUNTY_REFNIS	Municipality Code	A unique code associated with each municipality
Date	Between March 1 2020 to February 29,2022	The date on which the case/control (tweet) was posted

Table 4.2: Description of Attributes of the Survey-like Objects

A snapshot of how the survey-like objects in R is presented below:

	CD_MUNTY_REFNIS	Gender	Age	has_infection	date	Municipality
2	24104	F	age_>=40	0	18735.00	Tervuren
3	35013	F	age_30-39	0	18732.00	Oostende
4	44021	M	age_30-39	0	18629.00	Gent
5	21016	F	age_<=18	0	18959.00	Ukkel
6	21015	M	age_30-39	0	18922.00	Schaarbeek
7	23096	M	age_>=40	0	18424.00	Zemst

Table 4.3: R output snippet of the survey like objects

Chapter 5

Modelling Strategy

In the next step it is discussed how these biased survey-like objects are modelled for generating real time estimates about COVID-19 infections at municipality level in Belgium. The survey like objects presented above are now structured and have key demographics like age, gender, and a geo-location associated to municipalities in Belgium, and are thus amenable to statistical analysis. But they are still non-representative.

The technique of choice and relevance here is a multi-level regression model with post stratification. MrP is a prominent method for adjusting non-representative samples in order to analyze opinion and other survey results more effectively. It uses a regression model to link individual-level survey results to multiple factors before rebuilding the sample to better reflect the population. MrP can thus not only provide a better understanding of behaviors, but also enable to analyze data that would normally be illegible for statistical inference.

5.1 Multi-level regression with post-stratification (MrP)

In this research MrP as a statistical technique is utilized to produce estimates of disease prevalence for small defined geographic areas. In the scope of this research, the defined geographic unit for which inferences are to be made at municipalities of Belgium.

In general, MrP is an excellent technique to achieve certain goals, but it is not without drawbacks. If there is a biased survey, then it's a great place to start with MrP but it's not, a miracle of course . In this case, the digital trace data are biased as it is not a probability sample. They are not an actual representation of the real population for a number of reasons: people on social media do not represent the actual population. Firstly, there's a lack of representation of all age groups, social classes, on social media. And secondly, even if people use social media platforms, not all of them tweet about them getting infected and self-report their symptoms. Nevertheless, it is still attempted to make real time estimates about the disease prevalence of COVID-19 infections among Belgian population from digital trace data with MrP.

When dealing with survey data, MrP is a useful strategy. Hanretty explains how we utilize MrP because the alternatives are either inadequate or exorbitantly expensive[16] . In practice, collecting non-probability samples is less expensive, thus there are advantages to being able to use this form of data. In the context of this research, the non-probability samples are derived

from social media data, which are absolutely free of cost. However, MrP is not a magical cure, and the rules of statistics continue to apply. The estimates from such a model will be still be uncertain, and they will still be prone to all of the standard biases. MrP fundamentally trains a model on the survey data and then applies that learned model to another data-set. There is a major benefit of using MrP, that it allows to 're-weight' in a way that keeps uncertainty in mind and isn't majorly hampered by small samples. In this case, small samples are in the form of cases pertaining to each municipality.

The survey- like objects are modelled on the basis of their demographic characteristics and what is known about their area (i.e., which municipality of Belgium they belong to). This is the "multilevel regression" part.

There are elements of a person's life, like age, gender and the municipality of Belgium where he/she lives, and these may provide an indication as to their likelihood of being infected by the disease in a certain way. The MrP technique allows us to model, based on demographic characteristics, the disease prevalence. So, this technique will assess the likelihood of an individual with a particular set of variables to be infected. "Multi-level regression" examines to what extent each of these elements has an effect on having the infection.

For example, the following statement can be made: a female in the age group 19-29 living in Antwerp has a higher probability of being infected than a male in the age group 30-39 living in the municipality Etterbeek.

In the subsequent "post-stratification" stage, the census data is used to calculate how many people of each demographic type live in each area and combine this with additional relevant contextual information to predict how many of these people will be infected.

5.2 Stages in Generating Estimates Using MrP

When generating estimates using MRP, there are five important phases that must be completed:

5.2.1 Collection of survey data

The first step is to collect survey data (which are structured in form) that containing information on whether the respondent is infected or not, as well as information on respondents' demographic characteristics and which area the respondent resides in. This step is already achieved and are presented in the form of survey like objects discussed earlier.

5.2.2 Data on the area level key features

The area bifurcation in this study is at the municipality level. For each municipality, area level features regarding education level of the population, the income level by population, population density by population, the proportion of people by age group by population and the count of officially reported infections on the previous day is curated. These information for the area-level predictors for this study.

The following are the area level predictors that have been included : Net income/population, Population density/population, Proportion of population with Low level of education, Proportion of population with High level of education, Proportion of population with Middle level of education, Proportion of population with NAP level of education (Not applicable),Proportion of population with UNK level of education (Unknown), Proportion of population aged less and equal to 18, Proportion of population aged between and 19 to 29, Proportion of population aged between 30 to 39, Proportion of population aged equal and more 40 and Number of cases per Municipality the previous day. The source of all the socio-demographic information is https://statbel.fgov.be , i.e, the Statistics department of Belgium, Statbel. The data regrading COVID-19 cases is obtained from Sciensano.

By merging the survey like objects and the area level features, the training set for the model is obtained.

5.2.3 Estimation of a multilevel regression model on the training set

Now that the training set is prepared,in this step,the mixed effects multilevel logistic regression model model can be fit on the training set, i.e., the data from the first and second steps.

Regression models are models that attempts to use an equation to connect measures of a dependent variable, or outcome, to values of one or more independent variables, or predictor variables. In this case, the outcomes are categorical, "has infection" can take only two values, 0 denoting the person is not infected, 1 indicating that the person is infected. Therefore, a logistic regression model is utilised in this study.

Multilevel regression models are those in which the model's parameters apply to various levels. These layers are often organized hierarchically in MrP models. In other words, the model includes information about respondents from the survey like objects (level 1) as well as information about the area level features in which respondents are situated (level 2 information). Therefore, here a multilevel logistic regression model is used. In general, as modelling whether an individual is infected or not is possible using a regression model, it may also be estimated as part of MrP.

In addition, multilevel regression for the purposes of MrP is often estimated through Bayesian methods, and such is the decision in this research study, therefore **Stan** and **"rstan"** library on R is used to carry out the modelling. There are a few reasons why a Bayesian approach is chosen over the classical approach for regression analysis in this case. They are:

- The flexibility in specifying hierarchical priors is greater in Bayesian probabilistic programming languages such as Stan.
- The propagation and interpretation of uncertainty throughout a hierarchical model is more easily implemented and understood in Bayesian methods.
- The Literature has developed packages and tools to deal with MrP which are related to Bayesian framework.

The information on being infected (yes or no) for 1296 individuals of 575 municipalities is present and is the outcome variable.A mixed effect multi-level logistic regression model is fit

on the training set; where fixed effects are used for intercept, individual level regressors (age, gender) and area level regressors to leverage correlation across areas over certain variables like income and demographic profile, whereas random effects are used for area level identifiers, in this case a nested index for municipalities.

The benefit of using a random effects model is that it can account for not only the unique qualities of each municipality, but also the uncertainty that cannot be explained by such unique model features. The random effects model has frequently been employed to account for both 'between group variance' and 'within group variation'. In this case, the between cluster variation reflects the variation caused by individual municipality characteristics, whereas the within cluster variation reflects the random variation caused by uncertainty that cannot be explained by such municipality characteristics. In this study, the practical reason behind using random effects for municipalities is to also be able to make predictions for the municipalities that were not present in our sample, and thus account for the unobserved variations. Thus, uncertainty is preserved about unexplained information of the municipalities. And lastly, with an unrepresentative sample, like the one in this study, random effects model also allows for shrinkage towards 0 which reduces the bias in the model.

For the fixed effects parameters, α ,β, and θ, the assumption of weekly informative normal priors is considered. Stan describes normal distributions using standard deviations rather than variance or precision, like BUGS does. γ is the municipality-specific random effects and the hyper-prior for σ is chosen to be normal truncated (0,1). Finally, the outcomes are Bernoulli with probability provided by the linear predictor's inverse logit.

The model is presented as follows:

$$\text{y} \quad Bernoulli(\pi[\text{i}])$$

$$\text{logit}(\pi[\text{i}]) = \mu[\text{i}]$$

$$\mu[\text{i}] = \alpha + \gamma[\text{area\_id[i]}] + \beta\text{X[i]} + \theta\text{Z[i]}$$

$$\begin{aligned} \alpha &\quad N(0,1) \\ \beta &\quad N(0,1) \\ \theta &\quad N(0,1) \\ \gamma &\quad N(0,\sigma) \\ \sigma &\quad N^{+}(0,1) \end{aligned}$$

where X is matrix of covariates with predictors Age (4 categories) and gender (two categories), area_id is an index vector with a numeric identifier for each municipality and this nested index ensures each record is assigned to its correct municipality effect (its correct gamma) and finally, Z is a matrix of covariates of all the area-level predictors.

5.2.4 Obtain the test set

In order to obtain the test set, at first, a stratification frame has been constructed here.

A stratification frame is a count vector with one element per cell. Gender, age, educational attainment, marital status, and so on are examples of mutually exclusive sets of human traits that uniquely designate a type of interest. Stratification frames give counts that enable researchers to generate unbiased estimates of the true chances of each cell being included in the population as a whole. These stratification frames are typically thought of being precise; they estimate cell counts with little uncertainty. In general, a country's census may offer extremely exact frames. In this study, census data has been utilized to access the stratification frame for Belgium. Access to the whole census micro-file allowed for an incredibly precise count of the number of persons in each cell, subject to the census's demographic profiles as of January 2021.

The stratification frame used this study contains information on the proportion of municipality residents according to gender (male/female) and age group (four different groupings). There are therefore 8 (2 X 4) individual types for each municipality, and 4600 (3072 X 575) rows in the data.

The test set is obtained by combining the area level predictors along with the stratification frame. For the proof of concept, the predictions are made for January 24 2022. In Belgium, the same day had the highest number of consolidated daily new cases (76,034) according to official reports by Sciensano. Thus, it has been chosen as a dummy to test the efficacy of using the methodology presented in this study for generating real time estimates of COVID-19 infections. Therefore, for the area level predictor cases of previous day, data of January 23 2022 is used.

5.2.5 Making predictions from the multilevel regression model and post-stratification

The final step is to make predictions from the mixed effects multilevel logistic regression model estimated in stage 4 for each row of the test set and then post stratify to obtain the estimates of infection cases.

Now that the mixed effects multilevel logistic regression model is formulated and a test set is constructed, using the estimated regression model parameter values, the expected probabilities of being infected is calculated for each row of the test set. These anticipated values are then post-stratified, to obtain the count of total infections at the municipality level. In post stratification the weights of under-sampled and over-sampled sub-populations(here municipalities) are adjusted such that the total sample is more reflective of real municipalities.

The model for the test set is presented below:

$$\mu_t[i] = \alpha + \gamma[\text{area\_id}_t[i]] + \beta \text{X}_t[i] + \theta \text{Z}_t[i]$$

$$\pi_t = ilogit(\mu_t)$$

$$\pi[a]^* = \sum_{j \in a} \pi_t[j[a]] \times N[j[a]]$$

where π^* is the area level estimates (counts) and j[a] is a cell within area/municipality a and N[j[a] is number of people in cell j, living in municipality a.

This formula is different from what Lauderdale shows for post-stratification because in this study, the area level counts are of interest and not area level proportions.

Chapter 6

Results

Now the findings of the research are discussed. At first, mixed effects multilevel logistic regression model is validated such that inferences can be made from the model. Then the posterior distribution of coefficients of regression and their interpretations are presented. Finally, the performance of the model is clarified using two metrics: the correlation between estimated and genuine COVID-19 cases across all municipalities on a dummy date for prediction, January 24, 2022, and the root mean squared error (RMSE).

6.1 Validating the Regression Coefficients

The Rhat shows that the chains are in agreement, while the trace plat is a visual description of the same process.

6.1.1 Rhat

The Rhat values are all under 1.05 which suggests that the model has converged and is validated. Therefore, the posterior distributions are reliable to make inferences from.

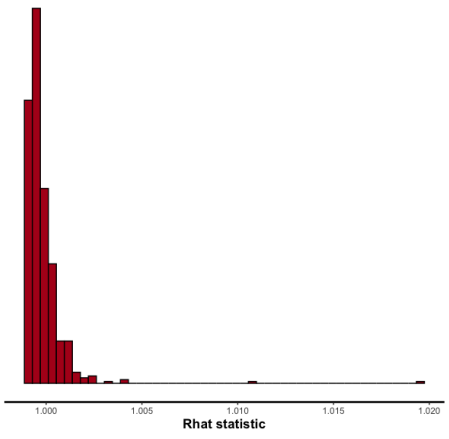

Figure 6.1: Histogram of Rhat Statistic

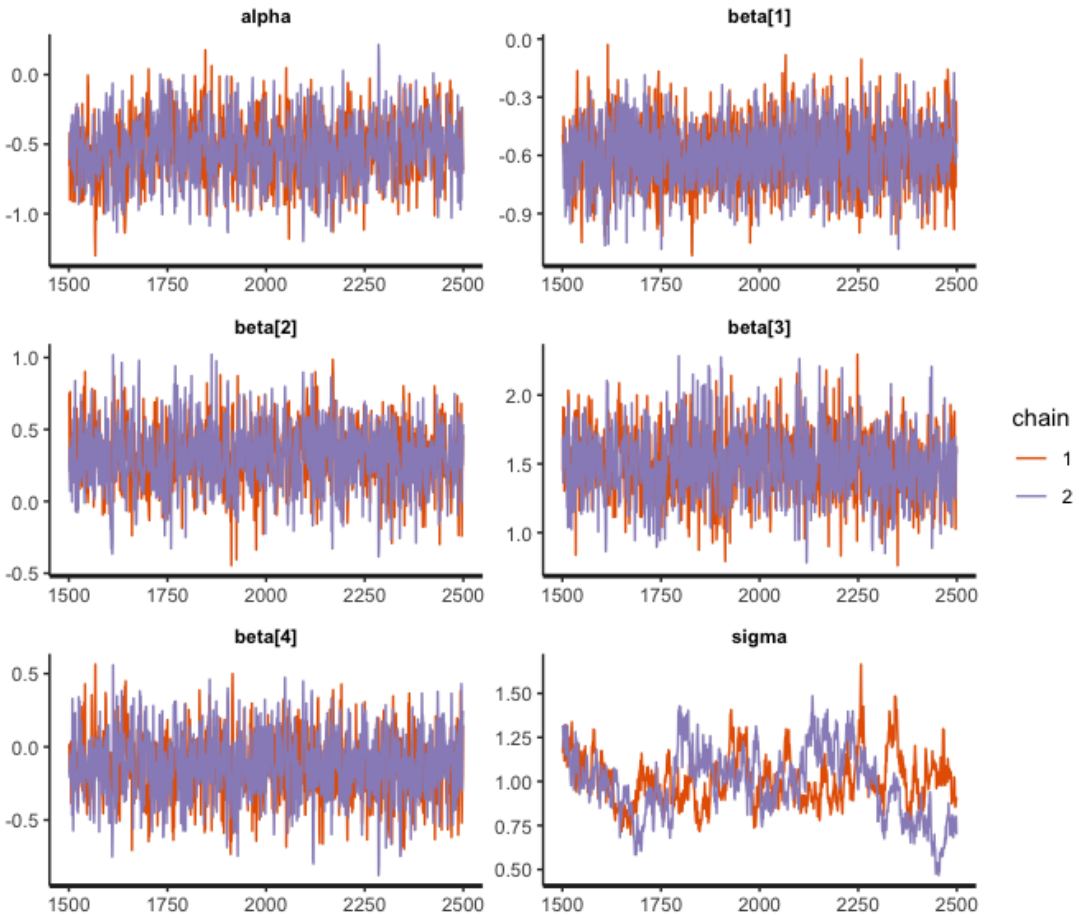

Figure 6.2: Trace plots of alpha, beta and sigma

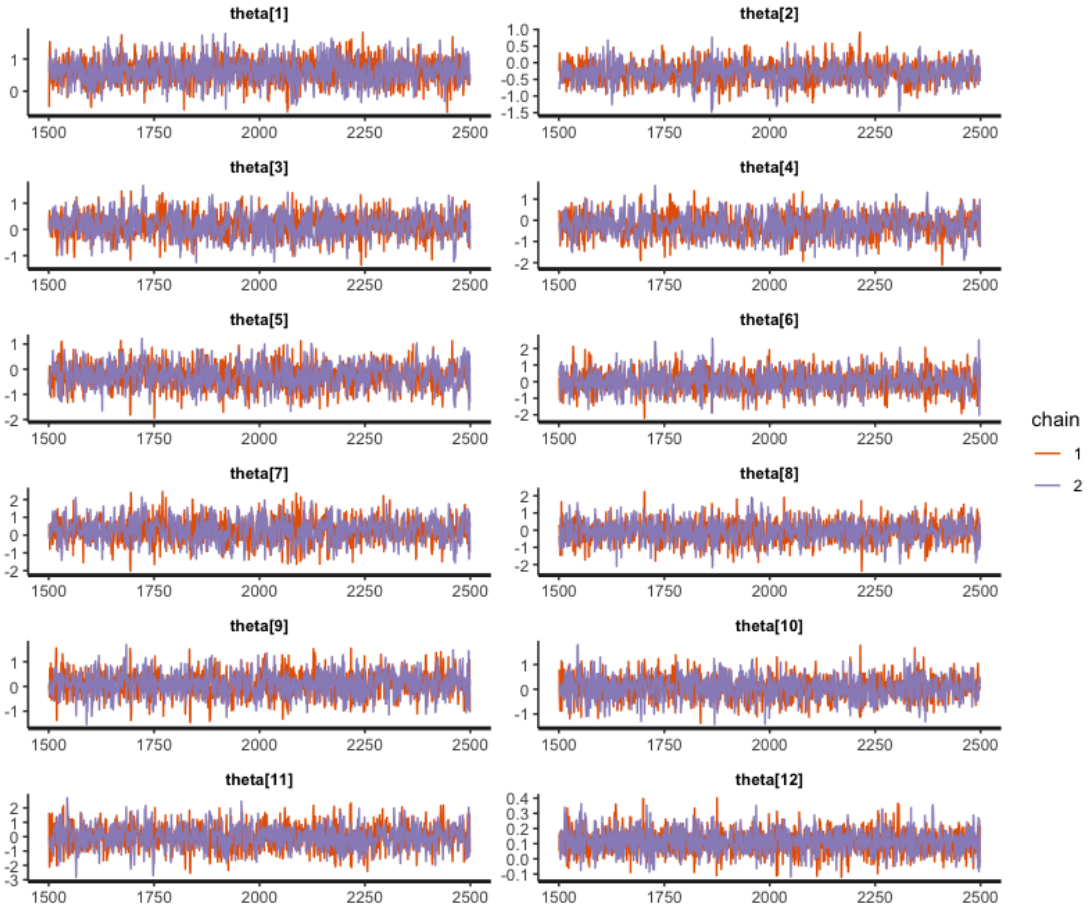

Figure 6.3: Trace plots of theta

6.1.2 Trace Plots

A trace plot displays the values that the parameter took during the chain's execution.

There is acceptable mixing of chains for alpha, betas and thetas across the iterations; which is a positive indicator.The fixed effects parameters have converged which is found from the fact that the two chains mix, and without the colouring, it will not be possible to distinguish between the chains. Thus it can be mentioned that the model has converged. The two chains in the case of sigma have not mixed well.However, it is normal for hierarchical parameters to struggle with convergence, but that the fact that all other parameters have converged, and those parameters are the quantities of interest in this study, is suggesting the model is fine. It is not possible to make inference on sigma by itself.

6.2 Examining the Posterior distributions of the Regression Coefficients

In this section, the posterior distributions of the fixed effects parameters are displayed and inferred.

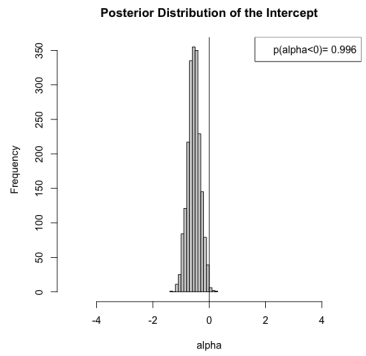

Figure 6.4: Posterior distribution of the intercept

The following are the inferences from the posterior distributions of the regression coefficients:

- The probability of the baseline rate of infection (intercept) to be negative is 0.996 and it's highly significant.What it means is, on average, an individual from the twitter population is extremely unlikely to report having the infection.
- The probability of the individual level regressor Gender F, to be negative is 1 and is highly significant.
- The probability of the individual level regressor Age $>=$ 40, to be negative is 0.076 and is not significant.
- The probability of the individual level regressor Age 19-29, to be negative is 0.
- The probability of the individual level regressor Age 30-39, to be negative is 0.029 and is significant.

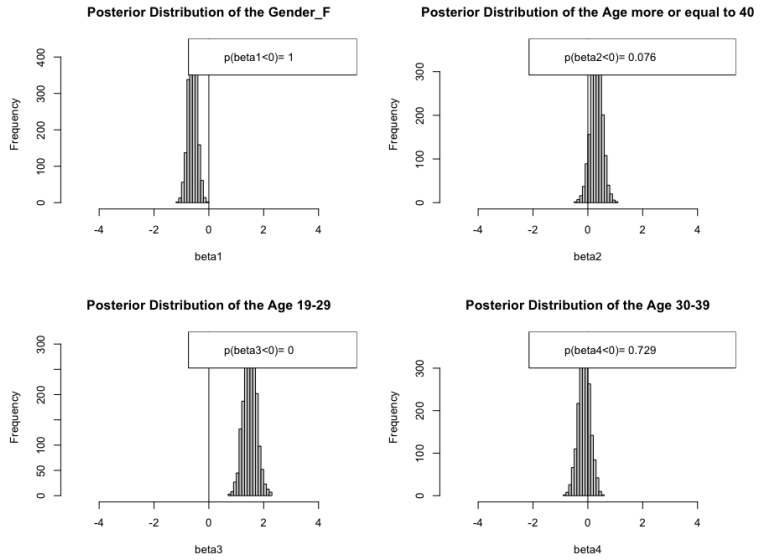

Figure 6.5: Posterior distribution of the individual level regressors

- The practical understanding from these interpretations are that female and population from age group 30-39 on an average are unlikely to report having the infection, whereas, people in the age group 40 and above have higher chances of reporting the infection.
- From Figure 6.6 and 6.7 inferences can be made about the posterior distribution of the area level regressors as well.

Likewise, the regression coefficients can be interpreted to learn in details about the log odds and odds of being infected based on individual level as well as area level attributes. Some inferences from the regression coefficients are that the baseline rate of infection is -0.54. Gender Female has a negative correlation with the probability of being infected. Being female decrease the log odds of getting infected by -0.59 as compared to male. Which in plain words mean than being a woman implies a decrease in probability of being infected relative to the baseline population, of male. Being in the age group of 40 or more years, increases the log odds of getting infected by 0.314, while being in the age group of 30-39 years of age decreases the log odds of getting infected by -0.13, both as compared to the baseline group for reference, age group less than 18 years.

6.3 Performance of the Model

In this case, the expectation from the model is not to make accurate predictions but to show that there is at-least some correlation with the model estimates and the actual observed number of cases per municipalities. Through that, it is established that there is feasibility in the approach used in the study, that is, building a pipeline from scratch to collect, organise and analyse social media data to predict COVID-19 infection prevalence. The results can be refined by improving the model and in future similar pipelines can be put in production for real time estimation of infection.

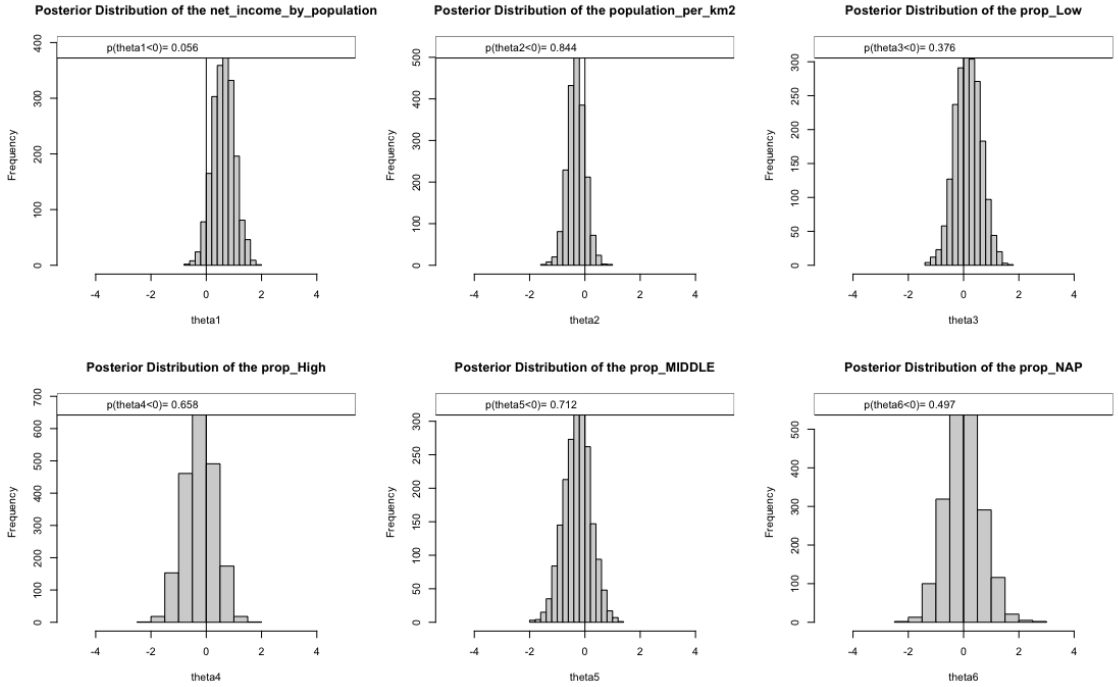

Figure 6.6: Posterior distribution of the area level regressors[Part 1]

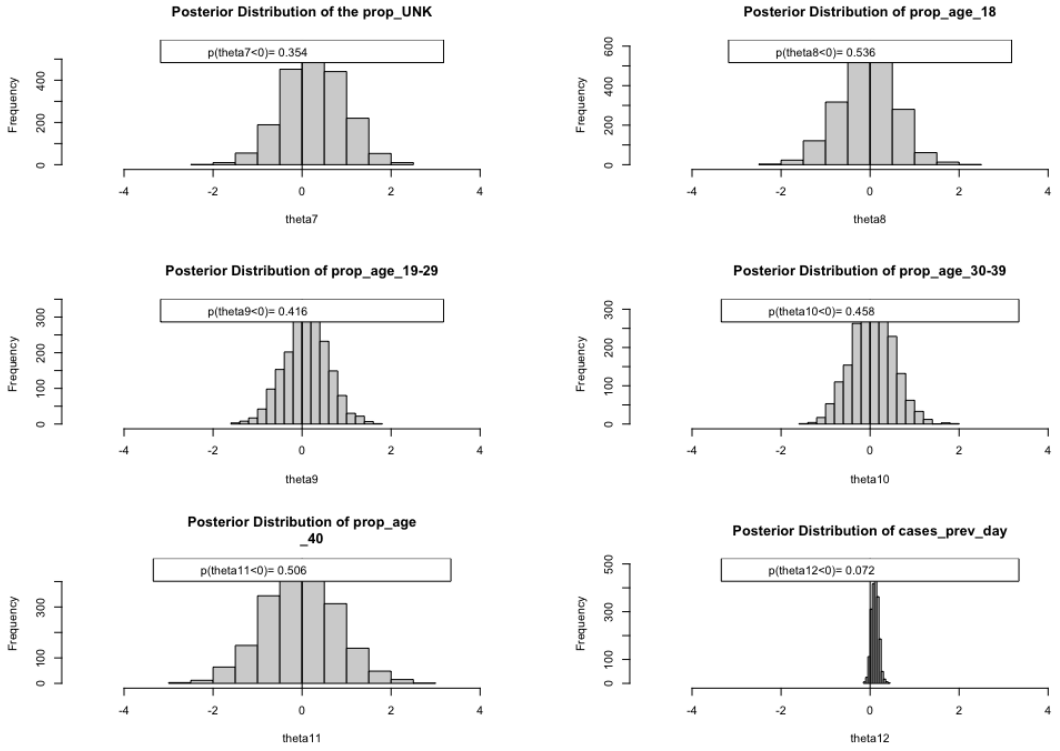

Figure 6.7: Posterior distribution of the area level regressors[Part 2]

The Pearson's Correlation Coefficient between the predictions based on cases count of 23 January 2022 from the model for the next day and the actual number of cases reported on 24 January 2022 is 0.938. This strong positive correlation is a very promising indication that there is enormous signal in the twitter data.

On the other sand, the RMSE is found to be very high at 16241.3. The predictions are massively inflated, as can be seen from the plot below, because the artificial split of the cases and controls is 50:50. That has increased the baseline rate dramatically.

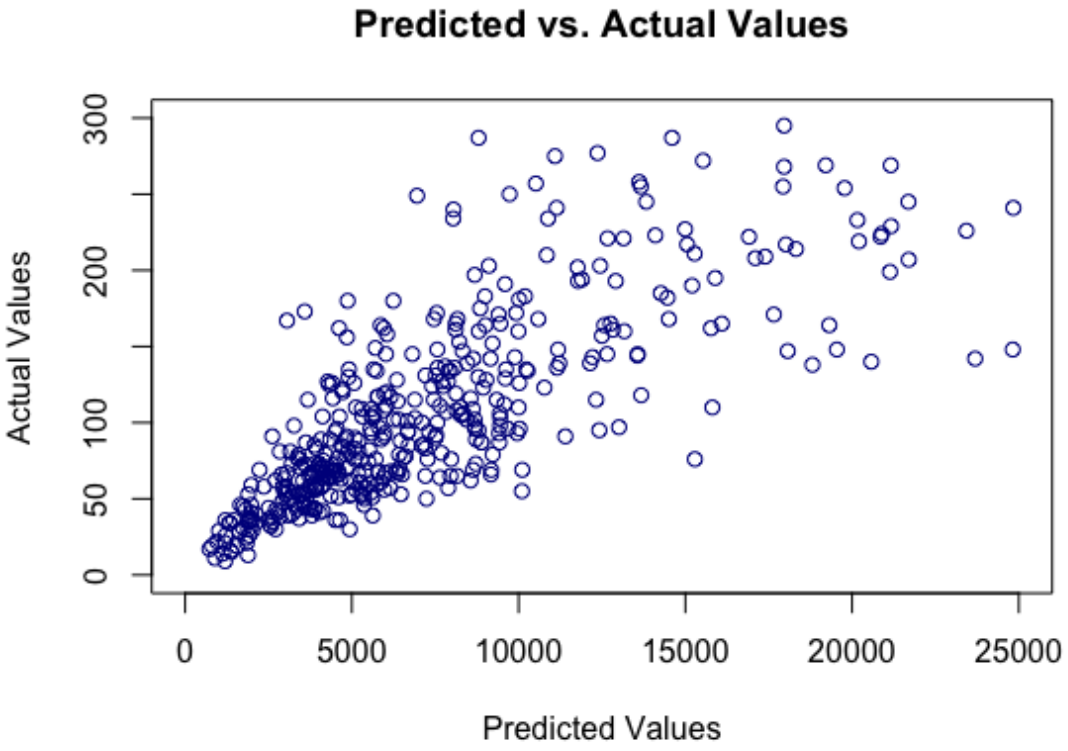

Figure 6.8: Predicted vs Actual Values

Chapter 7

Discussions

The recent outbreak of the coronavirus, COVID-19 (formerly SARS-CoV), has caused global concern. As part of the response to this pandemic, a number of studies have been conducted in an effort to improve our understanding and control the spread of the virus. In this study, a particularly innovative approach has been undertaken, that is, using publicly available social media data to track the spread of the disease in real time. This research project involves collecting tweets written in English, French and Dutch between March 1, 2020 and February 29, 2022 by Belgian citizens based on certain keywords related to symptoms of COVID-19 and with geo-locations of the municipality that these users belong to.

This method has several advantages over traditional surveillance methods. Firstly, it is free for the academia (however, business and industry professional might have to incur fees for using Twitter API), and the data is easily accessible - there are no specific hardware or software requirements to use it (beyond the standard academic toolkit of a laptop, R, Python and Stan) and the method does not require any additional infrastructure. This serves as an inexpensive way to get hands on important data as compared to physical surveys.

Secondly, it is in principle able to track the spread in real-time; meaning it is possible to monitor new cases as soon as they emerge. Finally, it is based on completely anonymized data of users online who have self-reported being infected by coronavirus and does not rely on physical self-reported information from the population. After the data processing step, i.e., turning Twitter data into survey like objects, the information becomes anonymous; there is no personally identifying information. The subjects of the data, whether infected or not, is no longer recognizable.

These advantages make social media monitoring an attractive field of research for tracking epidemics and monitoring the effectiveness of public health interventions. It can also be used to identify areas at high risk of the disease as well as people with potentially serious symptoms who may require further medical attention.

There are two main drawbacks of using this approach to monitor the spread of COVID-19. The first issue is that this method is subject to misclassification. Misclassification refers to the inclusion of tweets that are not actually related to the disease in the analysis. For example, if an individual is complaining about a headache in their tweet then this should be included in the analysis because it indicates that the individual may be suffering from some symptom of

COVID-19. However, if the individual is complaining because they have had too much to drink the night before then this would be excluded from the analysis because it does not indicate that they are suffering from an infectious disease and is therefore not an indicator of the spread of the virus.

To ensure such misclassification is reduced as much as possible, a combination of keywords is used, such that tweets containing both the symptom headache and the term COVID-19 are collected for the study. However, that is not completely full proof. A tweet mentioning how it has become a headache that the COVID-19 cases are rising everyday might still be taken into account. There is also the possibility of deliberate misreporting in the analysis of tweets which are deliberately designed to mislead the analysis, for example, bots spreading misinformation on Twitter.

The second issue is the reliance on historical data to determine whether certain individuals should be considered positive for the disease. This relies on the assumption that previous cases are linked to the current ones. The problem with this approach is that it does not take into account the possibility that the two cases are unrelated in any way other than the fact that they occurred at the same time. This is particularly interesting because of the different variants of the coronavirus. In the approach presented in the study, temporal dynamics are not accounted for in the modelling strategy, which can be made possible by including time series elements in the multilevel regression model. This remains unexplored in this research.

And finally, the major drawback of using data from Twitter is that although social media appear to be a goldmine of free data, they frequently contain significant demographic biases. Due to demographic imbalance in usage frequencies and accessibility rates, social media is a non-representative sample of the population. As a result, any direct inference or analysis derived from such a platform is likely to be biased towards specific demographics. It's mostly argued that a greater proportion of people on social media are younger population. Older population are generally considered to less tech savvy and have less inclination towards the usage of social media. It's worth mentioning that for a study concerned with prediction of coronavirus cases, an infection that affects the older population more than the younger ones, usage of social media data skews the analysis to a great extent. The second degree of bias comes from the fact that, among the people on social media, the percentage of people who consider self-expressing about being infected by a disease is very sparse. Furthermore, many of the "users" on social networking sites aren't genuine individuals at all - they're fake bots tweeting on behalf of advertising firms, or even false profiles for people who don't exist at all. These false accounts are frequently formed by unethical businesses in order to increase follower numbers and deliver misinformation.

M3 inference is a comprehensive approach to tackle the both the problems. It employs a unique method for allocating users to demographic strata that permits direct calculation of an individual's likelihood to belong to a particular age group, gender and whether that user is an organization or not. To ensure that in this study information from such fake profiles, are eliminated, only the classification of non-organization from the M3 inference has been utilized. This has significantly reduced the total number of cases of COVID-19 cases that have been gathered across all municipalities in Belgium.

So, to make these social media data compliant for statistical analysis, M3 Inference pipeline has been used to predict the key demographic features based on the twitter data and provide a gender and age classification to each case. As a result, the unstructured data is converted to survey like estimates. But there remains huge scope among social researchers to find effective tools to adjust for biases in social media. When it concerns non-probability samples, such as this one, the shortcomings of MrP as a statistical technique to make real time predictions to generate representative estimates, are yet unknown. As a result, it offers an interesting field of research in both academia and industry.

The main limitation of the methodology incorporated in this study, is the availability of appropriate quantity of data (from social media; Twitter) for training and testing the model. Although the study was able to identify a significant correlation between the predictions from the proposed model and the real COVID-19 prevalence as reported by the official records, the performance of the model can be improved manifold. The reason for the dramatic inflated predictions is a direct result of a 50:50 split of cases and controls in the methodology. But,the fact that the estimates from the model are still highly correlated with actual number of reported cases, suggests there is enormous signal in the twitter data and this paper can be viewed in the light of optimism. If the problem with the inflated intercepts can be fixed, potentially using an advanced case-control design, the RMSE would decrease drastically.

Moreover, it is unknown what proportion of the population from each municipality uses Twitter. Therefore, we could not adjust for this bias. With additional data that could eliminate the selection bias of social media and be used to train the algorithm.

This study has been ambitious in its goal to be able to use social media data to generate representative real time predictions of COVID-19 infection cases at the municipality level in Belgium, and the pipeline presented has ample opportunities of improvement. However, to the best of my knowledge, this is the first ever such attempt to provide a complete and comprehensive strategy, from data collection to making predictions made in Belgium. This is a back-bone proof of concept study that tries to show that, it is indeed possible to use statistical techniques, to utilize the untapped resource of social media data into digital epidemiology and study the prevalence of any disease at a granular level.

Even though there are certain un-ignorable pitfalls, such as high incidence of sampling bias and inefficient geo-coding of the collected data, it is believed that these shortcomings can be overcome with further advanced research in the same. The main strength of the approach taken by this study is that it confirms and possibility of using historical data from social media platforms to make predictions of possible new cases at a municipality level. This offers tremendous potential for real-time disease spread monitoring and control at a granular level, here the municipal level in Belgium. More importantly, it provides the first opportunity to test the use of social media data for disease surveillance purposes in a real-world setting (as opposed to controlled laboratory experiments). It is also believed that this approach can easily be generalized to other countries and regions and will be of tremendous help in the domain of public health monitoring.

The proposed pipeline consists of several steps: first, data on the number of infections for each Belgian municipalities are obtained using the Twitter API, using a custom R script. Second,

using deep learning methods, key demographic features like age and gender were added to convert the unstructured social media data into structured survey like objects. Third; these survey-like objects were then amenable for statistical analysis, and the method of MrP was utilized to make predictions and generate real time representative estimates at the municipality level in Belgium.

It is hoped that this demonstration will encourage other researchers and scholars within the academic community to develop similar approaches that can be used to automate the process of detecting and tracking novel outbreaks in real time from digital trace data.

Given that the feasibility of research is this domain is well established now through the discussion, it definitely is an indication of the lucrative scope in near future for advanced techniques and sophisticated modelling. As an outlook for the future, it would be interesting to extend the scope of utilizing digital trace data from only Twitter, to other social media platforms like Reddit, Facebook, etc. Reddit is a community news compilation, content rating, and discussion site based in the United States. Registered users post content to the site, such as links, text entries, photos, and videos, which are then rated positively or negatively by other users. Reddit features a robust API that permits users to access most of the information on the site. Further, an online dashboard available publicly that will enable users to access results of real-time predictions of progression of disease outbreak and track symptoms at their own province or municipality level is another goal that can be achieved in the future. This would allow users to compare the results of their municipality with those of other municipalities as well as healthcare officials to better manage emergencies by proper resource management and planning to tackle any novel outbreak.

In conclusion, it is suggested that the end to end pipeline presented in this research represents a very valuable source of information for the scientific community and public health authorities and this approach has the potential to make a major contribution to the fight against any infectious disease like COVID-19 by accelerating the rate at which new cases are detected, via sources of information other than the official health records and carrying out informed targeted interventions to mitigate the spread of the disease.

Bibliography

[1] Twitter API for Academic Research | Products.

[2] Epistat – COVID-19 Monitoring, 2021.

[3] Coronavirus disease (COVID-19), 04 2022.

[4] The pandemic's true death toll: millions more than official counts, 01 2022.

[5] The Regions | Belgium.be, 04 2022.

[6] Sitaram Asur and Bernardo A. Huberman. Predicting the Future with Social Media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

[7] Yan Bai, Lingsheng Yao, Tao Wei, Fei Tian, Dong-Yan Jin, Lijuan Chen, and Meiyun Wang. Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA*, 323(14):1406, 2020.

[8] Christopher Barrie and Justin Ho. academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software*, 6(62):3272, 2021.

[9] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[10] Samantha K Brooks, Rebecca K Webster, Louise E Smith, Lisa Woodland, Simon Wessely, Neil Greenberg, and Gideon James Rubin. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. *The Lancet*, 395(10227):912–920, 2020.

[11] Owen Dyer. Covid-19: Study claims real global deaths are twice official figures. *BMJ*, page n1188, 2021.

[12] Gunther Eysenbach. Infodemiology and Infoveillance. *American Journal of Preventive Medicine*, 40(5):S154–S158, 2011.

[13] M. J. Greenacre. Influential analysis and presentation of survey data. *Journal of Applied Statistics*, 14(2):153–164, 1987.

[14] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David S.C. Hui, Bin Du, Lan-juan Li, Guang Zeng, Kwok-Yung Yuen, Ru-chong Chen, Chun-li Tang, Tao Wang, Ping-yan Chen, Jie Xiang,

Shi-yue Li, Jin-lin Wang, Zi-jing Liang, Yi-xiang Peng, Li Wei, Yong Liu, Ya-hua Hu, Peng Peng, Jian-ming Wang, Ji-yang Liu, Zhong Chen, Gang Li, Zhi-jian Zheng, Shao-qin Qiu, Jie Luo, Chang-jiang Ye, Shao-yong Zhu, and Nan-shan Zhong. Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, 382(18):1708–1720, 2020.

[15] Jia-Wen Guo, Christina L. Radloff, Sarah E. Wawrzynski, and Kristin G. Cloyes. Mining twitter to explore the emergence of COVID-19 symptoms. *Public Health Nursing*, 37(6):934–940, 2020.

[16] Chris Hanretty, Benjamin E. Lauderdale, and Nick Vivyan. Comparing Strategies for Estimating Constituency Opinion from National Survey Samples. *Political Science Research and Methods*, 6(3):571–591, 2016.

[17] Brent Hecht and Monica Stephens. A tale of cities: Urban biases in volunteered geographic information. In *proceedings of the international AAAI conference on web and social media*, volume 8, pages 197–205, 2014.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[19] D. Holt and T. M. F. Smith. Post Stratification. *Journal of the Royal Statistical Society. Series A (General)*, 142(1):33, 1979.

[20] Janet Ilieva, Steve Baron, and Nigel M. Healey. Online Surveys in Marketing Research. *International Journal of Market Research*, 44(3):1–14, 2002.

[21] Michael Kearney. rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, 4(42):1829, 2019.

[22] Benjamin E. Lauderdale, Delia Bailey, Jack Blumenau, and Douglas Rivers. Model-based pre-election polling for national and sub-national outcomes in the US and UK. *International Journal of Forecasting*, 36(2):399–413, 2020.

[23] Jeffrey R. Lax and Justin H. Phillips. How Should We Estimate Public Opinion in The States? *American Journal of Political Science*, 53(1):107–121, 2009.

[24] Cuilian Li, Li Jia Chen, Xueyu Chen, Mingzhi Zhang, Chi Pui Pang, and Haoyu Chen. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance*, 25(10), 2020.

[25] R. J. A. Little. Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association*, 88(423):1001–1012, 1993.

[26] Jonathan Mellon and Christopher Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research amp; Politics*, 4(3):205316801772000, 2017.

[27] Alan Mislove, Sune Lehmann, Yong yeol Ahn, Jukka pekka Onnela, and J. Niels Rosenquist. Understanding the demographics of twitter users. In *In Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM*, 2011.

[28] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[29] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, 12(4):375–385, 2004.

[30] David K. Park, Andrew Gelman, and Joseph Bafumi. Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls. *Political Analysis*, 12(4):375–385, 2004.

[31] Seref Sagiroglu and Duygu Sinanc. Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013.

[32] Marcel Salathé. Digital epidemiology: what is it, and where is it going? *Life Sciences, Society and Policy*, 14(1), 2018.

[33] Nuhu A. Sansa. Effects of the COVID-19 Pandemic on the World Population: Lessons to Adopt from Past Years Global Pandemics. *SSRN Electronic Journal*, 2020.

[34] Abeed Sarker, Sahithi Lakamana, Whitney Hogg-Bremer, Angel Xie, Mohammed Ali Al-Garadi, and Yuan-Chi Yang. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *Journal of the American Medical Informatics Association*, 27(8):1310–1315, 2020.

[35] Soo-Yong Shin, Dong-Woo Seo, Jisun An, Haewoon Kwak, Sung-Han Kim, Jin Gwack, and Min-Woo Jo. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Scientific Reports*, 6(1), 2016.

[36] Luke Sloan. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society*, 3(1):205630511769898, 2017.

[37] The Economist. The pandemic's true death toll, 02 2022.

[38] Abigail Walker, Claire Hopkins, and Pavol Surda. Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak. *International Forum of Allergy amp; Rhinology*, 10(7):839–847, 2020.

[39] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[40] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *The World Wide Web Conference*, 2019.

[41] James H. Watt. Internet systems for evaluation research. *New Directions for Evaluation*, 1999(84):23–43, 1999.

[42] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, 43(4):721–734, 2017.

AFDEL
Straat nr bus
3000 LEUVEN, BE
tel. + 32 16 00 0
fax + 32 16 00 0
www.kuleuve