Following the Spread of Zika with Social Media: The Potential of Using Twitter to Track Epidemic Disease

Mo Wang

A Thesis
in
The Department
of
Geography, Planning and Environment

Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science (Geography, Urban and Environmental Studies) at

Concordia University

Montreal, Quebec, Canada

August, 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mo Wang

Entitled: Following the Spread of Zika with Social Media: The Potential of Using Twitter to Track Epidemic Disease

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Geography, Urban and Environmental Studies)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Kevin A. Gould	Chair
Perez Liliana	Examiner
Zachary Patterson	Examiner
Sébastien Caquard	Supervisor

Approved by _____

Chair of Department or Graduate Program Director

_____

Dean of Faculty

Date _____

Abstract

Following the Spread of Zika with Social Media: The Potential of Using Twitter to Track Epidemic Disease

Mo Wang

Epidemic outbreaks detection and monitoring is an important but challenging task in epidemic control strategies. In recent years social media has been seen as a promising data source to track epidemic disease. Epidemic detection approaches often rely on data mined from Twitter and Facebook. These data can be geolocated in two ways: either based on geographic coordinates of the location from where a tweet or a post has been submitted if available, or based on place names mentioned in the text posted. In this thesis I propose to further explore the potential of place names in tweets to track a specific disease outbreak: The 2016 Zika outbreak. To explore this potential I have first collected about 1 million of tweets mentioning "Zika" during a period of 15 weeks. I have then geoparsed this database using different approaches to identify Twitter activity related to Zika for 13 selected countries. I have systematically compared these results with the official number of new cases of Zika recorded by official organizations for each country, every week. The results of this first set of analysis show that the degree of correlation between the volume of tweets and the number of official new cases is overall pretty low. Throughout this analysis, I was able to identify that the volume of Zika related tweets was largely affected by events not directly related to this disease (e.g. *Venezuela struggles to contain Zika outbreak amid economic crisis*). In order to better understand the nature and the impact of these events, I have done an in-depth qualitative analysis focusing on one case study: Venezuela. Although for Venezuela the quantitative analysis showed a strong correlation between the number of tweets and the number of new official Zika cases per week, the qualitative content analysis

confirmed that very few of these Zika related tweets talk about new cases. In fact I was not able to identify even one Twitter account that would consistently provide information about new Zika cases while the disease was spreading out throughout the country. Based on these results I was able to emphasize the inappropriateness of using Twitter alone to try to track the spread of a disease, as well as the extensive use of Zika as a keyword for a large number of individuals and organizations to push other political and economic agendas.

ACKNOWLEDGEMENT

I would first like to thank my supervisor Sébastien Caquard of the department of Geography, Planning and Environment at Concordia University. He was always willing to support me whenever I ran into trouble or had difficulties with writing. Through the process of accomplishing my thesis, professor Caquard allowed me to think independently but also steered me towards the goal of this project with respect. His dedication and diligence not only motivated me to work hard on my thesis but also stimulated a higher pursuit of my life.

I would also thank Bailin Zuo who helped me develop data process machine to classify and clean up the enormous amount of data in this project. Having dedicated hours day and night to design the machine, his incredible work enabled an intelligent and efficient data process method that greatly alleviated the onerous manual tasks that would be required.

Finally, I would like to express my gratitude to my family, friends and lab mates for being so supportive and considerate. I would not finish this research without their encouragements.

Table of Content

LIST OF FIGURES

LIST OF GRAPHS

LIST OF TABLES

I. Introduction

Epidemic diseases cause great loss to our economy, public health and results in social panic (Philipson, 2000). This is further exasperated by lack of early detection and monitoring of epidemic outbreaks and communicating these to the general public. The 2003 (SARS) outbreak in China let to mass panic and hysteria by the general population, and a large part of this population bought drugs hoping that these drugs will help them control the spread of SARS virus (Zhong & Zeng, 2006). Over the years, a considerable amount of economic and human resources have been invested to better understand and control the spread of epidemic diseases. However the traditional epidemic control strategy involves several steps such as case management, monitoring and contact tracking as well as laboratory service, which makes this control process complicated and inefficient (Infection control strategies, 2008). Recently researchers noticed the potential offered by social media to detect epidemic outbreaks. These approaches often rely on the geolocation of people posting relevant comments on social media regarding an epidemic outbreak. However, a small proportion of social media users are willing to share their geolocation to the public. In fact, only about 1% of Twitter users share their geolocation when posting their tweets (Cheng et al, 2010). Furthermore, the disease mentioned in the post is not necessarily related with the user's location. For example, a Twitter user in Canada could post a tweet mentioning Ebola in Africa. In light of these considerations, another approach of mobilizing the geographic potential of social media data is developed based on the place names mentioned in disease-related social media texts. Although this method seems promising as illustrated by its use in some online platforms designed to monitor the outbreak of epidemic disease and to alert the public, such as HealthMap and SickWeather, its real potential to track the spatial spread of an epidemic disease still requires overcoming a set of challenges. In this project I propose to

study the potential and limits of geoparsing social media to track the spread of an epidemic disease. More specifically, I propose to address this issue through the study of the spread of the 2015 – 2016 Zika outbreak using data from Twitter.

The Zika outbreak was first detected in Brazil in 2015 and rapidly transmitted across South America before reaching North America (HealthMap, 2016). Pregnant women who are infected with the Zika virus have a high probability of developing severe birth defects such as microcephaly (Microcephaly & Other Birth Defects, 2016). Given the spatial nature of this outbreak and its increased exposure on media and social media, it appears to be a very relevant case to assess the potential of geoparsing social media to track its spread.

To assess this potential, I first review the literature on epidemic disease, epidemic control strategies and the potential of social media for epidemic detection. In the following section I describe the methodology, elaborating on collecting and geoparsing Twitter data. Then I compare the results obtained with the official new cases for a selection of 13 countries for a period of 15 weeks. This systematic statistical analysis is followed by a more qualitative analysis of the tweets mentioning both Zika and Venezuela. Finally I conclude this thesis by emphasizing the limits of this research and by highlighting new research directions in the domain of geoparsing social media for studying epidemic disease.

II. Literature review

2.1 Epidemic disease

An epidemic is the rapid spread of infectious disease to a large number of people in a given population within a short period of time, usually two weeks or less (Teutsch, 2000). Epidemic diseases can have huge human and economic consequences and lead to social disaster if there are no limited or therapeutic interventions (Coker, 2011). Usually, infectious diseases spread through populations by contact between infective individuals (those carrying the disease) and susceptible individuals (those without the disease, but can catch it) (Newman, 2002). Human infectious diseases can be detected when individuals notice infected individuals (neighbors) around them. Those 'neighbors' initiate a 'small-world' effect, which means that two people can still infect each other even without physical contact (Amaral et al., 2000). Experiments performed by Travers (1967) suggest that there are only about six intermediate acquaintances separating any two individuals on the planet, which in turn suggests that theoretically, a highly infectious disease could spread to all six billion people on the planet in only about six incubation periods of the disease. Furthermore, with current advanced and highly developed modern transportation networks, the outbreak of epidemic diseases which used to emerge endemically and periodically in isolated populations then die out without spreading to a larger region can now become worldwide crises (Miller, 2007).

2.2 Current controlling measures and spreading patterns of epidemic disease

Five components are required to control epidemic disease: (1) case management, (2) monitoring and contact tracking, (3) laboratory service, (4) safe burials, and (5) social mobilization. Among the five components, the most challenging ones are case management as well as monitoring

and contact tracking (Nicholson, 2016). People with minor symptoms of disease are suspected of being infected. In cases of highly infectious epidemic diseases caused by viruses, those infected are treated in specially designed isolation rooms. This process requires a good estimation of the number of people that may potentially get affected. For instance, to control the SARS outbreak in Taiwan, 764 specially designed negative-pressure rooms equipped with air filter devices were built for infected individuals and for suspected cases of infection. Furthermore, a lot of professional healthcare workers were sent to the infected area from organizations such as World Health Organization and Centers for Disease Control (Twu et al., 2003).

The monitoring and tracking of disease consists of two parts: (1) microscopically observing changes in the disease causing virus and (2) macroscopically monitoring all people that had contact with suspected cases (Ecker, 2005). Microscopically observing mutations in the virus is relatively easy to do, while tracking its potential spread is much more challenging. For example, to monitor the spread of the Ebola virus, medical workers tried to identify all individuals that had direct or indirect contact with infected people. However, because the outbreak happened in remote areas, it was extremely difficult to get in touch with people. As a result, the potential virus carriers were sometimes difficult to identify. This difficulty can sometimes become the most challenging part of disease surveillance.

Although epidemic disease can spread very quickly, it has a certain spatial distribution pattern that can be calculated. Christophe Fraser and his research group argue that the spread of infectious disease such as the 2009 H1N1 outbreak, has a geographic pattern that can be determined by human

interactions and mobility across multiple spatial scales (Fraser et al., 2009). In light of this, obtaining real-world data about human interaction and mobility becomes a fundamental and critical issue in assessing the spread of disease (Balcan et al., 2009). For instance, the spread of influenza has been studied with a gravity model for epidemics originating in two different locations: California and Wyoming (Figure 1).

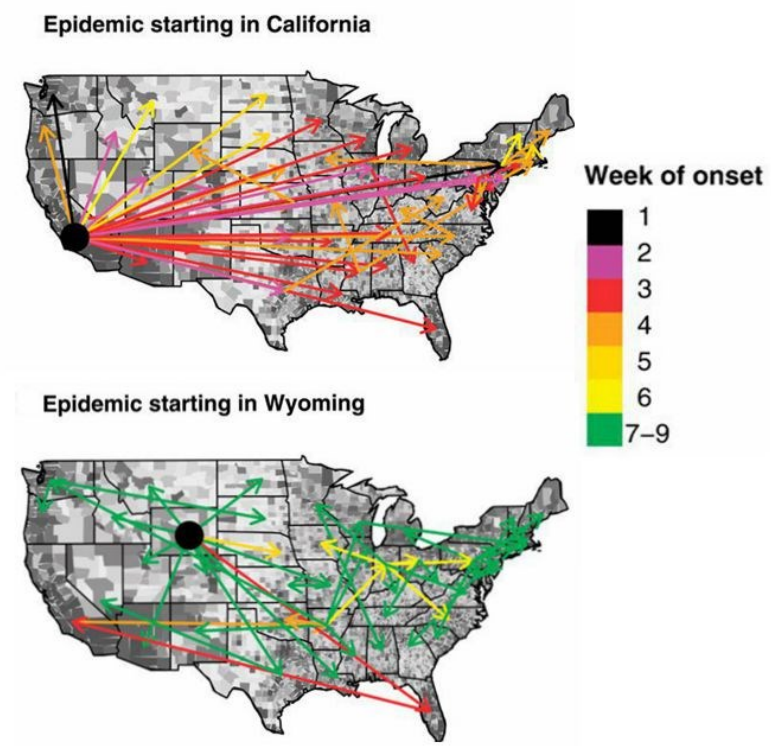


Figure 1: This shows the spatial-temporal distribution of the epidemic starting in two different locations (California and Wyoming). The two maps show the time required for a disease originating in California / Wyoming to spread across the United States. Filled black circles represent the location of initial cases. Arrows indicate the spread of the infection in each individual state. Arrows are color coded, based on the date of the epidemic's onset in individual states, from black = early onset, to green = late onset; see color bar (Viboud et al., 2006).

Data related to the geographical distribution of epidemic diseases are usually produced in text forms describing certain cases instead of being systematically structured in tables. Although some of these data can be mined manually from these texts (e.g. origin, age, sex), automated process are required to mine a large corpus of texts, such as texts posted on social media.

2.3 Geoparsing

Geoparsing refers to the extraction of place names from unstructured text. It is also known as Geographic Information Retrieval (GIR), which is an extension of Information Retrieval (IR) (Yates & Neto, 2011). Specifically, GIR focuses on identifying and extracting geographic-associated entity names or place names. Although GIR is often considered a relatively new approach, it was already mentioned in 1900s as a subtopic of IR (Plewe, 1997). Information Retrieval is an activity to obtain information from a document based on a given content detecting engine (Frakes & Baeza-Yates, 1992). Geoparsing process is often developed by connecting the IR machine with a gazetteer to specifically retrieve geographic names by matching extracted items in the document with the toponyms in the gazetteer.

In the last few years Geoparsing attracted much attention given the growing volume of unstructured text in the GIS database. It has been helping people on decision making and first-response during emergency operations, in finding shelters with special facilities (Hariharan, Hore, Li, & Mehrotra, 2007). More specifically, many studies have started to utilize geoparsing on social media data to achieve a broader application. For example, it was applied to predict the cholera spread pattern during the 2010 Haiti Earthquake by identifying infected areas mentioned in social media posts

generated by the local people (Chunara, Andrews, & Brownstein, 2012). It can also be applied to estimate the damage of rapid flood through places reached by the flood as mentioned by social media users (Poser & Dransch, 2010) as well as to study the urban activity (Cranshaw, Schwartz, Hong, & Sadeh, 2012).

There are a few geoparsers available that have been developed by private companies and universities. These geoparsers can be categorized based on the openness of their source code. The geoparsers with the programming code not available online are called proprietary geoparser, while others that share their developing code are open source geoparsers. Four proprietary geoparsers are available: (1) Metacarta is a private corporation that produces products related to geographic information including a geoparser named Geographic Text Search (GTS) (Frank, 2007). (2) Yahoo PlaceSpotter is one of the geotools developed by Yahoo. According to Gritta (2017) Yahoo PlaceSpotter is supposed to offer good performance on identifying and disambiguating place names in unstructured text (e.g., differentiating "London" UK from "London" Ontario). It is also able to record the number of times each place name is mentioned and where it is mentioned in the text (https://developer.yahoo.com/boss/). (3) NetOwl Geotagging was developed in France by SRA international. It is an advanced geoparsing tool that includes all classic geoparsing functions as well as latitude/longitude information for "relative location phrases" such as "a town 50 km northwest of Paris". Furthermore, NetOwl Geotagging can also identify where a person has been or where a certain event happened. (4) GeoCLEF is known as a Cross-Language information retrieval tool focusing on geographic information extraction developed by Mandl and his research group (2006). The advantage of GeoCLEF is that it is able to extract place names in English, German and Portuguese, but the reliability of the results has been challenged (Martins et al, 2010).

There is also a range of open-source geoparsers available to the public. Most of them can be found on GitHub with detailed installation instructions and packages. One of the most popular open-source geoparser is CLAVIN. It was developed by Berico technologies. It extracts place names from unstructured text documents by employing Natural Languages Processing (NLP) techniques, and then resolves these place names against a free version gazetteer downloaded from GeoNames that includes countries and major cities around the world. It also has the capacity to disambiguate and recognize incorrectly spelled words through a fuzzy search (https://clavin.bericotechnologies.com/). DIGMAP is another state-of-the-art geoparser developed by EDINA, a UK-based data centre. It offers a range of on-line geographic mapping and data related extraction tools, but its use is restricted for users within of the United Kingdom (http://www.webarchive.org.uk/wayback/archive/20140614011421/http://www.jisc.ac.uk/edina). Other geoparsers such as TextGrounder, Geodict, GeoDoc and Geotext are available on Github, but they offer limited description and no performance evaluation.

In order to select the relevant geoparser for this project, I reviewed the performance information provided by the developers. Generally, the geoparsers are tested by their developers to assess their performance in different categories. These categories are usually set by the developers according to the goals of the project. For example, the evaluation of GeoCLEF focuses on the machine's ability to detect languages and retrieve information among different languages interactively. Although other tests assess the capacity of the geoparser to extract location based on context, there is no standard test for GIR machine performance estimation.

2.4 Social media

Social media can be defined as electronic communication (such as Web sites for social networking and blogging) through which users create online communities to share information, ideas, personal messages, and other content such as videos (Edosomwan, 2011). The predecessors of social media are social networks and blogs. Social media can serve different functions, such as sharing information, feelings and ideas with others who can be located on the other side of the world (Sutton, 2008). It has the capacity to disseminate images and messages to a large number of people within an extremely short period of time, regardless of distance (Scanfeld et al., 2010). For instance, in March 2014, Ellen Lee DeGeneres posted a photo, on Twitter, of an all-star selfie at the Oscars and within one hour, it was retweeted more than 900,000 times. The photo was reposted by users all around the world, spreading from the hall of the Oscars in Los Angeles to a laptop in an apartment in Tokyo. With these distinct advantages, social media has become a potential collector of fresh and large-scale data. Social media include a broad range of applications such as Twitter, Facebook, YouTube, Google+ and Instagram (Kaplan & Haenlein, 2010).

Although these different applications could be used to study social interactions, Twitter appears to be the most relevant social media application for tracking and analyzing epidemic disease. Twitter is a free social network microblogging service that allows registered members to broadcast short posts that are called tweets. Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices. Tweets and replies to tweets can be sent by cell phone text messages, desktop clients or by posting at the Twitter.com website. A large number of users are active on Twitter. As of the first quarter of 2015, the microblogging service averaged at 236 million monthly active

users (Number of monthly active Twitter users, 2017).

The main reason why Twitter is the most relevant social media application for my research is that data from Twitter is much more accessible than data generated via other types of social media such as Google+ and YouTube. In fact, among the five social media platforms mentioned above, Twitter is the only one that allows public access to data posted and to the users' accounts. Although Twitter gives free access to only a fraction of the data it generates (Noordhuis et al, 2010), given the huge amount of tweets generated on a daily basis, this fraction is often considered representative of the overall data generated on Twitter (Chew, et al, 2010). Tweets are organized by topic so that users are able to follow accounts that always post messages that interest them, whereas Facebook users have to add each other as friends or group members in order to see or forward others' posts. Twitter users are able to follow whomever they are interested in, and are able to view, comment on and retweet posts from the people they follow. Therefore, Twitter relationship model allows users to keep up with the latest happenings posted by any other Twitter users (Russell, 2013). Additionally, Twitter posts are brief and specific which is in contrast with other platforms such as Instagram, where the posts can be very informative and descriptive and consequently more complex to mine and extract. Beyond these advantages, there are several issues associated with the use of Twitter to track epidemic disease (Schmidt, 2012), as discussed later in this thesis.

2.5 Social media and epidemic disease

Studying the distribution pattern of infectious disease requires collecting spatial data about infected people, including where and when they were infected. Several methods have been developed

over the years to estimate the actual number of patients affected by a given illness, from school and workforce absenteeism figures to phone calls and visits to doctors and hospitals (Neuzil, 2002). Other methods include randomized telephone calls, or even sensor networks to detect pathogens in the atmosphere or sewage (Ivnitski et al., 1999). All of these methodologies require an investment in infrastructure and have various drawbacks, such as delays due to information aggregation and processing times.

The use of social media has become a new trend of tracking infectious disease. Health Map (http://healthmap.org/) was the first social media tool to monitor infectious disease. It provides a more informative report on the status of disease than official monitoring dose such as Centers for Disease Control (CDC). This is because Health Map aggregates the outbreak news from various sources of authoritative websites or alerts, thus allowing for very timely updates on new cases displayed in a good visual map on HealthMap. This can be considered a huge improvement in disease tracking. A major limitation of Health Map is that it provides misleading information on the outbreak locations based on wrongly interpreted data from the official websites (Schmidt, 2012). It detects new infectious disease cases based on keywords in the webpages, which may wrongly report some "outbreak locations" where no new cases have actually occurred. For instance, a conference or an event that included the key word disease could be considered an outbreak location. Crowdbreaks (http://www.crowdbreaks.com/), developed by Salathe research group in the Center for Infectious Disease Dynamics and the Health Map team at Boston Children's Hospital, is a collaborative effort to crowd-source disease monitoring using aggregated Twitter feeds, user-driven data refinement, and machine-learning algorithms. Compared to Health Map, it has a lower detection sensitivity. GermTrax

is another online infectious disease monitoring application. It is an interactive disease alert tool with which users could share their sickness condition using a computer or mobile phone, and when a certain place becomes an aggregation of a group of sick people with similar symptoms, it may be an indication of a potential infection outbreak. At the local level, this information might point out a particularly problematic location; for example, if 50 people who were sick with food poisoning all visited the same restaurant, the restaurant would have been uncovered as potentially problematic. At the global level, GermTrax can help individuals and health agencies discover large-scale sickness trends (GermTrax, 2015). A drawback of this application is that the group of users is too small to help researchers detect outbreaks. Furthermore, similar to Twitter and Facebook, people are less habituated to posting their disease condition, especially on a specific disease alert application. Sickweather has the same basic function as many disease monitoring systems. It extracts disease related data from social media such as Twitter and Facebook, but it also warns people when there is an outbreak of infectious disease around the user's area. It, however, faces the same problem as GermTrax: the number of users is too small and could not provide good service outside the U.S.A. ProMED-mail is similar to Sickweather, however, it receives disease outbreak information from subscribers via e-mail or other forms of messages, then it verifies this information with official reports before alerting the public of the verified outbreaks. Besides alerting the public, it also provides necessary information to help people avoid getting infected. In light of its data analysis process, the alerts usually will not be ahead of the real outbreak. The last two online disease detecting applications mentioned are more relevant to observe the number of sick people around a user's area, rather than track an epidemic disease by mining disease-infected information on social media. In summary, there is still a lot of improvement required on the sensitivity and accuracy of tracking infectious disease by mining data

from social media.

Apart from the online maps, some very advanced health organizations are mining data from all the accessible social media platforms and using these modern communication technologies to monitor many initial outbreak reports. This kind of informal data sources provides the organization with a more complete picture of the epidemic threat to global or national health security. For example, The Global Public Health Intelligence Network (GPHIN), developed by Health Canada in collaboration with WHO, is a secure Internet-based multilingual early-warning tool that continuously searches global media sources such as news wires and web sites to identify information about disease outbreaks and other events of potential international public health concern (Epidemic intelligence, 2017). Proved by previous disease monitoring experiences with GPHIN, more than 60% of the initial outbreak reports are detected by unofficial informal monitors. Nowadays, GPHIN has become one of the most important sources for disease control (About GPHIN, 2017).

Overall, using social media as a data source offers an alternative source of information to observe the real world. The advantage of using Twitter to monitor the diffusion of disease is that it can help reveal an experienced and evolving situation based on a stream of data (Tweets) created within a few hours by a large number of people (Lampos, 2010). Engaging with and using emerging social media may well place the emergency management community, including medical and public health professionals, in a better position to respond to disasters. Controlling the geolocated information of mobile phones and the timeline of posts, applications such as Twitter allow people to check in at specific locations and specific time to share information about their immediate surroundings (Yang,

2009). These data can then be retrieved to study the spread of particular phenomena such as a disease outbreak.

These advantages have been further explored in recent studies focusing on monitoring infectious disease via spatial information hidden in social media data. Research on the Haitian cholera outbreak shows that analysis of place names extracted from Twitter data got timely estimates of the spread of the cholera virus two weeks earlier than traditional predicting approaches (Chunara et al., 2012). By studying the toponyms geoparsed from H1N1-related Tweets, which contained opinions and experiences surrounding the bird flu, Cynthia Chew and Gunther Eysenbach found that Tweets can reflect the spread of influenza in the real world (Chew et al., 2010). The key element of using social media on disease tracking is geoparsing the toponyms such as county names, city names or street names from those disease related posts, then simulating the disease spread using this information collected from social media data. It is these potentials that I want to further explore to study the 2014 Zika outbreak.

2.6 A case study: The 2014 Zika outbreak

Zika virus, as well as some other well-known viruses such as dengue, yellow fever, Japanese encephalitis and West Nile, belong to the virus family Flaviviridae and the genus Flavivirus (Faye et al, 2014). Zika virus (ZIKV) was first identified in Uganda in 1947 in rhesus monkeys and is mainly transmitted by daytime-active Aedes mosquitoes. Five years later, it was subsequently reported in humans in Uganda and the United Republic of Tanzania (The history of Zika virus, 2016). By 2016, cases of Zika virus disease had been identified in Africa, the Americas, Asia and the Pacific. Zika

infection often causes mild symptoms such as mild fever and joint pain (Malone et al, 2016), but the ZIKV can last in blood and continue infecting other people. The effects of this disease do not seem severe, but it can potentially impact people's lives as well as that of their descendants. For instance, during the large outbreaks in French Polynesia and Brazil in 2013 and 2015 respectively, national health authorities warned the public about the potential neurological and auto-immune complications of Zika virus disease. Recently in the northeast of Brazil, the local health authorities have noticed an increase in Guillain-Barré syndrome as well as a number of babies born with microcephaly. The increase of these pathologies coincided with the development of the Zika virus in the general public (WHO, 2016).

There are different vectors that can transmit the Zika virus to human: (1) Bites from infected mosquitoes from the Aedes genus, mainly Aedes aegypti in tropical regions (Transmission, 2017); (2) Blood transmission; Recently, in a Zika outbreak area, 2.8% of blood donors who were asymptomatic at the time of donation, tested positive for acute ZIKV infection after donating blood (Cao-Lormeau et al, 2014); (3) Sexual transmission as identified throughout a survey conducted by Mark R. Duffy and his research group (2016). These different transmission vectors contribute to the spread of the disease around the world (See figure 2).

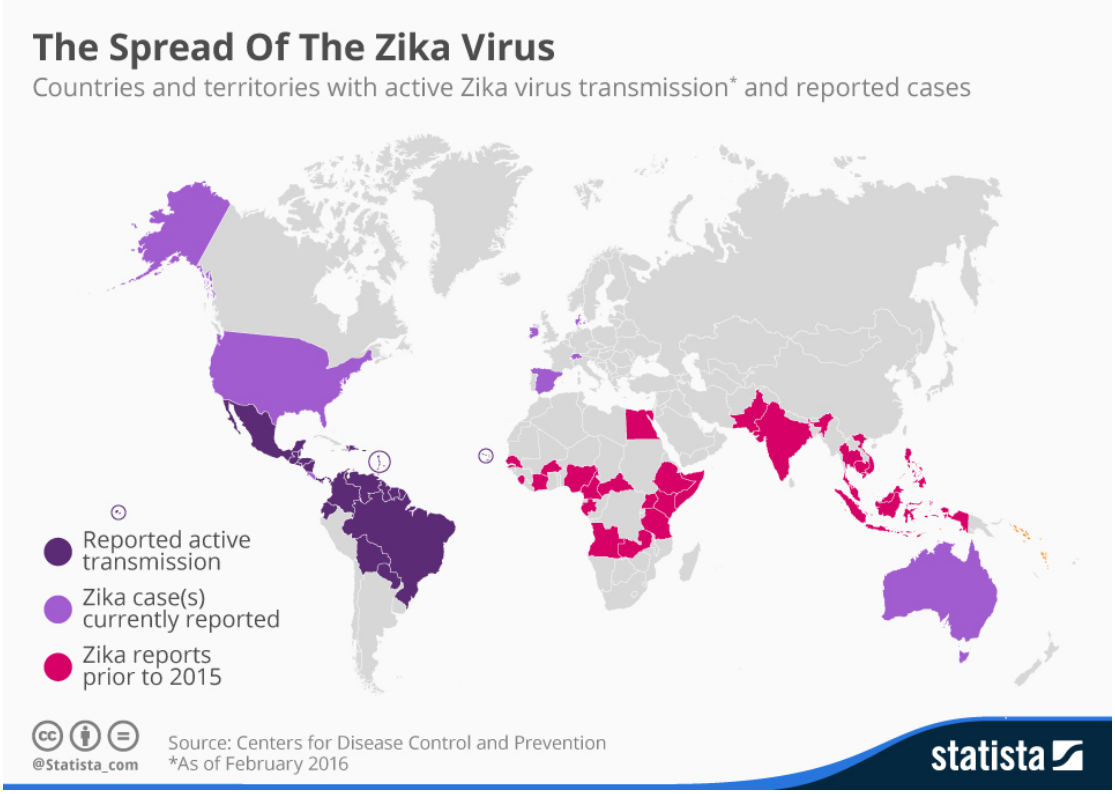


Figure 2: The spread of the Zika virus around the world. Data source: https://www.statista.com/chart/4322/the-spread-of-the-zika-virus/

To prevent the Zika virus from infecting people, tips and warnings of how to self-protect and self-diagnose are made available in public space as well as on the Internet. However, making the public aware of the risks of getting infected by ZIKV is effective but not enough to control the spread of the disease. One major problem is that there is no vaccine or drug available today to prevent or treat the disease, and developing a safe and efficient vaccine requires 10 to 15 years and would cost approximately US$1.8 billion (Paul et al, 2010). Since the effective treatment has yet to be developed, tracking the spread of ZIKV and taking action before outbreaks can be critical. Using social media to track the disease is definitely one way of addressing this issue. According to my literature review, there was no research on this topic using this approach at the time this thesis was written. As discussed earlier, in this project I aim to explore the potential of geoparsing tweets to study the spread

of the Zika virus. In the following chapter, I will introduce the methodology I have developed to track the Zika outbreak using data from Twitter.

III. Methodology

3.1 Data Collection

3.1.1 Official data collection

The first step in this project was to compile a reliable database of Zika cases that will be used as a reference to assess the accuracy of Zika data mined from Twitter. There are several reliable health associations that constantly monitor and report global or regional Zika epidemic situations on weekly bases: the Pan American Health Organization (PAHO), the European Center for Disease Prevention and Control (ECDC), the World Health Organization (WHO) and the Center for Disease Control and Prevention (CDC). Among them, PAHO provides the most comprehensive statistics about Zika outbreaks in South America. ECDC and WHO provide a better view at a global scale. CDC focuses more on the situation within the United States and in parts of South America. PAHO and WHO provide data about Zika outbreaks more often and more regularly: data is mostly updated every week. They provide information about the number of new cases and the nationalities of the patients, and sometimes provide extra details such as age, gender and where people had been prior to getting infected. The reports also include the accumulated numbers of Zika cases in each affected country, which are presented in tables and maps. Based on the review of these different sources, I compiled data from PAHO and WHO to create my reference database.

3.1.2 Twitter data collection

Twitter enables public access to tweets through two key functions of its Application Programming Interface (API): the streaming API and the search API. The streaming API can be used to subscribe to a continuing stream of new tweets containing specific keywords or originating from specific users or defined locations. Whereas the streaming API is relevant to monitor tweets

continuously based on defined criteria, the search API, by contrast, can be used to retrieve past tweets according to a range of criteria including keywords/hashtags, senders, location, etc. The search API will only return a limited number of tweets, and therefore cannot be used to retrieve a comprehensive archive of past tweets containing specific hashtags. Furthermore, there are in-built limits on the number of keywords or Twitter accounts that can be queried at any given time or within a certain timeframe. The major issue, however, might be the fact that the search API can only be used to retrieve tweets posted within the last 6 to 9 days depending on Twitter activity. For hot topics that generate a lot of tweets, the requests can only go back to one or two days. For example, frequent requests were accepted when I was collecting tweets with the key word "Ebola", and I could track back all Ebola related tweets to about 10 days earlier. However, when it came to Zika outbreak, one request was restricted to one and half day since it was returning too many tweets. It should be noted that some of these limits can be overcome, at a cost, by accessing the Twitter API through one of the third-party resellers of Twitter content. Playing with the programming code can also help with overcoming these difficulties. For instance, to ensure a complete collection of tweets over time, I manually set the time period for data collection and the specific day for Twitter data harvesting. After testing different options and strategies, I decided to use the search API every day for a period of five months to collect enough data for my analysis. Unfortunately, even doing so, the Twitter data collected during May remained incomplete and could not be used for this research. Tweets harvested for March, April, June and July were then stored in a database that was being queried for the analysis.

3.2 CLAVIN

After carefully reviewing the potential of the various geoparsers mentioned in the geoparsing section, I decided to use CLAVIN for this research. This choice is based on two reasons: (1) CLAVIN was developed based on Stanford NER (Named Entity Recognition), which is a stable and much acknowledged NER system (Bontcheva et al, 2013). (2) CLAVIN uses Geonames as its gazetteer, which contains a list of place names around the world in various scales. It was the most comprehensive gazetteer freely available at the time my research was conducted.

Two versions of CLAVIN geoparser are available online: the desktop version and the web version. The desktop CLAVIN package is available on GitHub, where it can be downloaded and installed on a personal computer, and then used without internet connection. To use the desktop version, the user needs to download and install the code package required for geolocation extraction, and the gazetteer that includes a large number of place names for the countries and major cities in the world. By contrast, to apply an online testing version of CLAVIN, users must go to CLAVIN website then copy and paste their own data into the location extracting data box.

For the web version, there are two data extracting options called, extract location and resolve location. “Extract location” returns the extracted place names and their frequency (i.e. the number of times each place name appears in the text). “Resolve location” returns the geographic coordinates of all the extracted place names, but without returning any frequency. For example, in the text “I am from China and I came to Canada to study. I go back to China every year”, “extracted location” will return “China (2) Canada” (only the place name that is extracted more than one time will be followed

by a number representing the frequency), while “resolve location”, will return “People’s Republic of China 35.65861, 104.06472 CN” and “Canada 60.10867, -113.64258 CA”. By combining the two functions, users could identify all the place names, their frequency as well as their geographic coordinates.

To evaluate the geoparsing performance of CLAVIN, I used the data I mined from Twitter. I randomly selected 1000 tweets (Note: these tweets were selected from the 10,623 tweets collected in March 2016 since I did this test in April 2016). I then read all the selected tweets and identified all the place names in the tweets. This manually extracted database serves as a reference to test the performance of CLAVIN. I then used both the online version and the desktop version of CLAVIN to geoparse the place names. I compared the results with the manually extracted data based on: (1) The total number of identified place names, (2) the comprehensiveness of geoparsing results. The assessment results are listed in table 1.

Place names	**Manual**	**CLAVIN online**	**CLAVIN desktop**
Américas Park	None	1	None
Atlanta	1	2	None
Baltimore	None	2	None
Brazil	36	36	38
Colombia	23	18	26
Colorado	1	None	None
Columbus	12	None	None
Cuba	3	4	4
England	1	None	None
Florida	1	1	None
Floridablanca	None	2	None
French Polynesia	None	1	None
Haiti	1	1	1
Îles du Vent	None	1	None

India	1	1	None
Jamaica	27	27	None
Japan	1	None	1
Kenya	1	1	1
La Romain	1	3	None
Laos	3	5	None
LosAngeles	1	None	None
Majorca	1	1	1
Mexico	1	None	1
Manila	None	1	None
Missouri	2	3	None
Nigeria	1	1	1
Orlando	1	1	None
Paris	2	2	None
Peru	2	2	1
Pucusana	None	2	None
Puerto Rico	3	5	6
Rio de Janeiro	13	11	None
San Fernando	1	None	None
San Francisco	12	12	None
Spain	1	1	1
Texas	3	3	None
Tobago	2	2	None
Trinidad and Tobago	2	1	1
Tokyo	1	None	None
United States	6	8	None
Venezuela	2	1	1
Yuma County	None	4	None
Zikah	None	1	None
Zimbabwe	1	1	1

Table 1: Comparative analysis between place names identified by manual, CLAVIN online and CLAVIN desktop geoparsing methods.

From the 1000 selected tweets, I manually extracted 30 different places, while CLAVIN online extracted 36 different places and CLAVIN desktop only extracted 14 different places. These results indicate that desktop CLAVIN returns much less satisfactory results. In fact, this desktop version misses all the sub-country-level toponyms, including the popular ones such as Paris and Tokyo. This

may be improved by trying to link it to a different gazetteer. It is important to note that even at the country-level, desktop CLAVIN fails to identify some countries such as India and Jamaica. The results obtained with online CLAVIN are more reliable even though it missed a few place names and created a couple of false positives such a Zikah.

Although the assessing performance of the online version of CLAVIN is quite good, it does not allow users to track back the location of place names within the text parsed. This raises another issue related to the level of ambiguity associated with city-level place names. It is a pretty common issue with geoparsing. This means that by using CLAVIN, it is not possible to know if a place name has been properly identified or if it is a false positive, since we cannot track back where in the text the place name was identified. This issue is confirmed by Liu (2014) who emphasizes both the strong performance of Stanford NER in terms of Named Entity Recognition work, as well as its ambiguity issues, mainly at the level of recognizing local names.

In other words, the online version of CLAVIN is pretty good at identifying place names, and it is easy to use. However, it tends to extract false positives and ambiguous results at the sub-country level. This is not an issue for this project since my reference data, and therefore my scale of analysis, are at the country level.

3.3 Semi-automatic Geoparser

The principle of semi-automatic geoparsing is quite simple: using search function to identify keywords (i.e. place names) in a text. For this project, I used Java programming language to identify place names that are matched with specific keywords (i.e. country names) while scanning through all

the tweets collected for this research. Only the country names relevant for this project are selected. For instance, in the official dataset, countries like Brazil, Mexico and Colombia are consistently reported during the data collecting period. These country names are used as the keywords to extract the tweets that contain at least one of these names. Therefore, this semi-automatic geoparsing method is relevant in identifying the specific tweets that contain place names mentioned both in official and Twitter datasets that are of interest to users. For my project, this will enable me to look for specific tweets that contain place names mentioned in official reports.

To assess the performance of semi-automatic geoparser, the same test sample as previously done to assess the performance of CLAVIN was applied again, but I only used the country-level toponyms geoparsed in CLAVIN assessment as the input keywords. See table 2.

Place names	Manual geoparsing	Semi-automatic
Brazil	36	38
Colombia	23	26
Cuba	3	4
French	None	1
Haiti	1	1
India	1	1
Jamaica	27	27
Japan	1	1
Kenya	1	1
Laos	3	5
Mexico	1	1
Nigeria	1	1
Peru	2	2
Puerto Rico	3	6
Spain	1	1
Trinidad and	2	1
United States	6	8
Venezuela	2	2
Zimbabwe	1	1

Table 2: Comparative analysis between place names identified by manual and semi-geoparsing methods.

The assessment result indicates that semi-automatic geoparsing offers promising performances. The main advantage of this approach is that it allows for the identification of the place names within each tweet. It is also interesting to notice that this method identified a couple of place names that I missed during my manual analysis (there is no perfect method…). I did not test the performance of this semi-automatic approach at the sub-country-level since in this specific research, I will limit my data collection from tweets only at the country-level as emphasized previously.

3.4 Data processing

3.4.1 Extracted toponyms dataset preparation

Since the reference data (place names mentioned in official reports) are only provided at the country level, I focused on country names to extract data from Twitter. It is also important to

emphasize that I have decided not to collect data at the sub-country level - even if they could be aggregate at the country level - because of the inconsistent performance of geoparsing at this scale.

28 countries/administration entities are included in the reference database: Aruba, Barbados, Belize, Bolivia, Colombia, Cayman Islands, Dominica, Ecuador, El Salvador, French Guiana, Guadeloupe, Grenada, Guatemala, Haiti, Honduras, Jamaica, Martinique, Mexico, Panama, Paraguay, Peru, Puerto Rico, Saint Barthelemy, Saint Maarten, Saint Martin, Saint Kitts & Nevis, Suriname and Venezuela. New Zika cases were continuously reported only for 22 of these countries/administration entities, and for the other 6 just occasionally updates. 13 of these 22 countries were mentioned regularly on Twitter during the period under study: Aruba, Brazil, Colombia, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Panama, Puerto Rico, Venezuela. To ensure a consistent comparison between the reference data and the Twitter data, I focused on these 13 countries.

3.4.2 Noise and noise cleaning

There are two types of noises in semi-automatic geoparsing. The first type refers to the mistaking of adjectives as place names. The principle of semi-automatic geoparsing is to scan through the whole string to identify specific keywords. However, certain place names are similar to adjectives. For instance Jamaican is not a place but would be recognized as the place Jamaica since “Jamaica” is included in “Jamaican”. The second type of noise is caused by shared names between country-level places and sub-country-level places. For example, Panama is the country name and Panama City is a sub-country-level place name. By inputting Panama, both of them will be extracted. To eliminate

these noises, I created a sub-database with all the possible mistakes (e.g. Jamaican and Panama City) and I used it to remove all such mistakes from my main database. Then I used this cleaned up data for my analysis.

3.4.3 Data normalization and organization

All the tweets were periodically collected for five months, March 2016 to July 2016, but only the data for March, April, June and July were usable since the data collected for May was uncompleted due to high Twitter traffic during that month. Even for the other months, it was sometimes difficult to harvest all the tweets for a complete collecting period. For example, normally, one collecting action should gather all the tweets that were posted 7, 8 days earlier. However, when there is too much traffic or there has been too many requests for data collection, the harvesting could get cut off in just a few hours from the mining start time. This resulted in tweets being mined for only a few hours instead of a whole day, for certain days during the data collection period. For instance, the collecting period for March 10th is 19 hours, and 17 hours for April 21st. Therefore, the data collecting hours are not perfectly consistent. To address this problem, a normalization procedure is required to make Twitter data comparable across different weeks/periods, as explained below (section 3.4.4).

Another necessary procedure is data organization. The main idea of this research is to compare the epidemic outbreak data collected through Twitter with the data reported by official institutions. Since the spread of infectious disease is very time sensitive, maintaining a good and consistent time scale to compare two datasets (official dataset and Twitter dataset) is crucial. As noticed in many official Zika outbreak descriptions, Epidemiology Week (EW) is used to divide each epidemic

reporting period and is used as time stamp associated to every official new Zika case reported. The definition of EW is "A standardized method of counting weeks to allow for the comparison of data year after year" (CDC). Epidemiology Week begins from the first week of January (http://www.cmmcp.org/epiweek.htm). Thus, in order to be comparable with official records, the Twitter data is also aggregated into EW according to the dates that the tweets were mined. The social media data harvesting period ranges from March to July, except for 5 weeks in May. This period completely covers EW 10 to EW 17, and then EW 23 to EW 30. In total, I was able to collect Twitter data for 15 complete Epidemiology Weeks, which is what I used for all my analysis.

3.4.4 Twitter data Normalization

The normalization process follows two steps: First, normalizing toponyms collected daily in a 24-hour unit. If h(d) represents the total hours of Twitter mining period in a certain day (d), and frd (t,d) represents the total times a toponym is mentioned in the data collection of this day, then the normalization function could be FNd (t,d) = [frd (t,d) / h(d)] * 24. FNd (t,d) is the normalized frequency of this toponym in this specific day. It should be noted that this normalization process assumes that place names are evenly posted every hour during a day. Second, every EW's data is normalized in a 7-day unit. In certain days, the tweets are completely missing. For example, EW 10 includes the days from March 6th to March 12th, however, the data for March 11th is missing. In this circumstance, I used the average of the days that have Twitter data to multiply by 7 in order to make each EW's data represent a complete EW period. For example, EW 10 ranges from the 6th to the 12th of March, but Twitter data is missing for March 11th. I, therefore took the average frequency of each

toponym (e.g. Brazil's data on 6th, 7th, 8th, 9th, 10th and 12th of March) and added it to the 6 available to obtain a full week of data.

3.5 Final dataset

Having processed datasets from Twitter and official data for the 13 countries under study, I organized them into tables. See table 3-1 to 3-13.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	29	12	4	1	5	0	3	1	1	0	7	0	0	0	1
Official	29	12	4	1	5	0	3	1	1	0	7	0	0	0	1

Table 3- 1: The final dataset for Aruba.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	4782	5299	5239	4641	1533	2971	3228	4530	5467	2188	2856	2253	1415	3194	8832
Official	22200	21250	17100	16000	13700	12250	10500	9200	3550	3250	3050	2550	1890	1800	1400

Table 3- 2: The final dataset for Brazil.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	929	337	183	515	225	2062	368	911	106	858	132	1969	89	202	3313
Official	3600	2950	2450	3750	3250	3200	3000	2850	1800	1500	1250	1240	725	510	500

Table 3- 3: The final dataset for Colombia.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	2	9	11	2	2	22	416	14	0	2	34	1	7	38	11
Official	9	10	2	10	11	1	2	36	245	375	415	320	325	250	200

Table 3- 4: The final dataset for Ecuador.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	77	42	23	218	19	11	30	14	9	548	22	122	46	7	48
Official	110	80	50	60	70	45	75	35	40	55	80	60	45	45	30

Table 3- 5: The final dataset for El Salvador.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	22	14	50	27	14	14	32	11	8	2	6	17	7	31	15
Official	53	55	3	36	52	50	60	54	107	80	87	81	54	48	50

Table 3- 6: The final dataset for Guatemala.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	104	134	152	57	78	55	121	646	18	27	496	108	20	25	49
Official	146	100	83	70	69	45	45	38	63	113	94	77	38	17	17

Table 3- 7: The final dataset for Haiti.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	114	148	10	19	12	15	7	12	105	10	27	28	1	16	1689
Official	625	500	125	300	495	500	375	510	990	1000	830	800	940	650	550

Table 3- 8: The final dataset for Honduras.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	138	309	220	199	151	113	30	148	319	308	319	178	96	655	272
Official	85	84	83	81	95	101	75	99	505	480	432	312	263	247	197

Table 3- 9: The final dataset for Jamaica.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	189	168	306	39	197	106	267	72	77	49	216	71	31	240	114
Official	17	5	3	20	201	17	23	27	96	137	125	180	237	273	195

Table 3- 10: The final dataset for Mexico.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	43	669	889	75	71	25	14	22	6	37	3	16	0	18	9
Official	78	80	91	103	52	69	73	63	72	55	54	41	46	31	36

Table 3- 11: The final dataset for Panama.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	2282	1952	3153	1367	447	365	3214	10299	1215	3213	1315	1935	407	755	4445
Official	20	40	10	10	10	90	150	100	800	1300	1350	1450	1300	1550	1320

Table 3- 12: The final dataset for Puerto Rico.

Week	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**23**	**24**	**25**	**27**	**28**	**29**	**30**
Twitter	299	273	48	27	45	207	71	60	88	28	27	4	7	3	10
Official	3000	2900	1500	1300	1400	1250	1500	1250	1100	1800	1500	1050	1050	800	700

Table 3- 13: The final dataset for Venezuela.

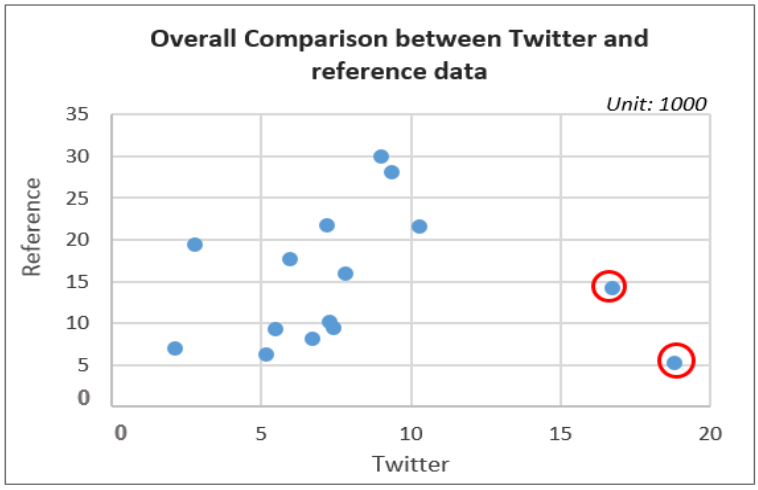
IV. Analysis and Results

The goal of this research is to seek if the place names mentioned in Zika-related tweets can help us better understand and track the spread of Zika in the reality. To achieve this goal, I conducted a two-step analysis. The first step includes 2 sections. The first section provides an overview of the two types of data (Twitter and official), and tests their correlation from 2 perspectives – overall and by individual country. The second section critically interprets and discusses the possible factors that may have influenced the results of section 1. The second step is an in-depth exploration of what might be the causes of the results from step 1 and an attempt to identify these causes such as demographic, tourism and languages. Together these 2 steps provide elements to better understand how to optimize the quality of data harvested from Twitter and how non related events interact and eventually generate Twitter data that may or may not be relevant to study and track disease outbreaks.

4.1 Primary data analysis

4.1.1 Overall analysis

In this section, I present the results of the comparison between the number of tweets per EW in which both the term “Zika” and a country name are mentioned, and the number of new official cases identified for that specific EW. I start by comparing the overall results of the 13 countries identified previously, and then compare the results for each country. Graph 1 shows the overall correlation between the twitter database and the official database for the 13 selected countries for each EW as well as the two EWs considered as outliers (circled in red). Table 4 presents the Pearson test results before and after removing these two apparent outliers.



Graph 1: Number of tweets (in 1000) mentioning Zika and one of the 13 selected country names per weeks Vs total number of official cases (in 1000) for this same week.

Data	Correlation coefficient	t-value	Degree of freedom	Significance
Original	0.01	0.029	13	< 50%
Outlier removed	0.56	2.233	13	95%

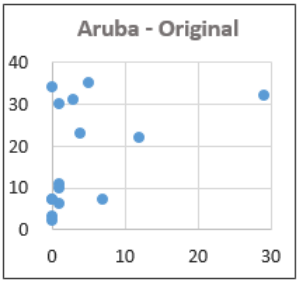
Table 4: The statistical correlation between reference data and data mined from Twitter, before and after the removal of outliers.

If we compare our two databases with a simple statistical measure such as the Pearson coefficient, our first results show no correlation between both databases (see table 4). However, this result is obviously affected by the two apparent outliers identified in red circles which correspond to EW 17 and EW 30. Although data was properly collected for these 2 weeks, they are quite unusual. In fact, if we look more closely at the data collected during EW 17, it appears that Puerto Rico is mentioned in almost 9000 tweets, whereas the average frequency-of-mentions in the other EWs for Puerto Rico is only about 1000 tweets. This unusually high volume of tweets mentioning Puerto Rico in EW 17 explains why this EW is an apparent outlier. The cause of this apparent outlier is due to an extremely intensive discussion on the first identified case of microcephaly caused by Zika virus in

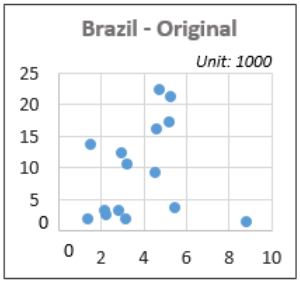
Puerto Rico. A very similar apparent outlier is identified in EW 30, but this time with an unusually high volume of tweets mentioning Brazil. It is important to emphasize that these apparent outliers are not associated to errors in the data collection but are rather an outcome of the nature of the Twitter data, which is highly influenced by certain events. By removing these 2 apparent outliers, we obtain better results with 0.56 Pearson correlation coefficient (see table 4). These results seem to point out that there may be a correlation between the official data and the data mined from Twitter. To further explore this possibility, I have applied the same method for each of the 13 countries under study.

4.1.2 Individual analysis

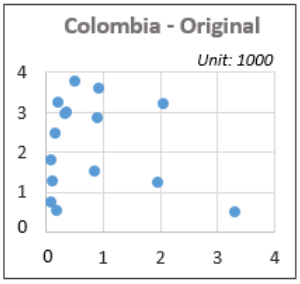
In the first series of analysis (see graph 2-1 to 2-13) it appears that almost every country has one or several apparent outliers. These results are confirmed in table 5, in which only one country, Venezuela, demonstrates a relatively strong positive correlation.



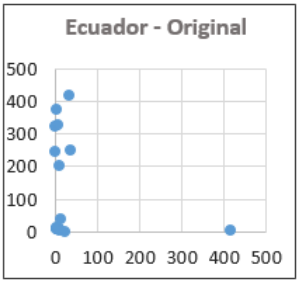
Graph 2-1



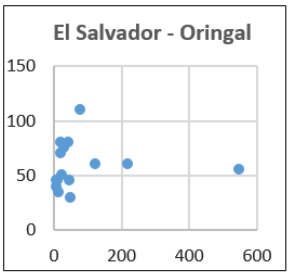
Graph 2-2



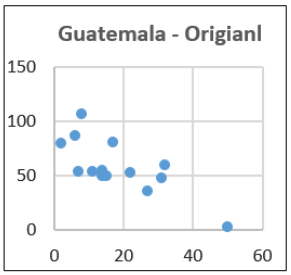
Graph 2-3



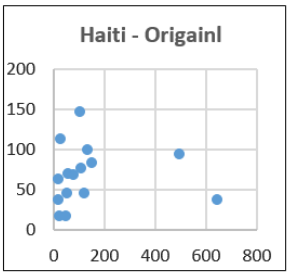
Graph 2-4



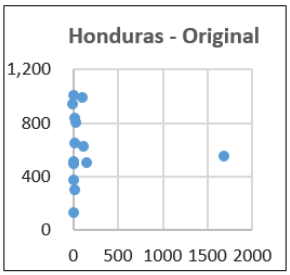
Graph 2-5



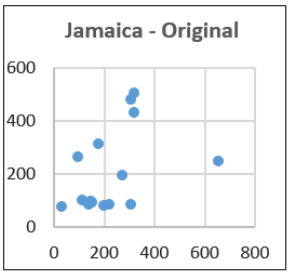
Graph 2-6



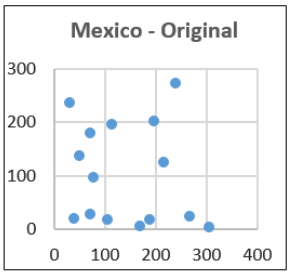
Graph 2-7



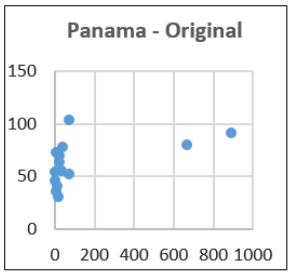
Graph 2-8



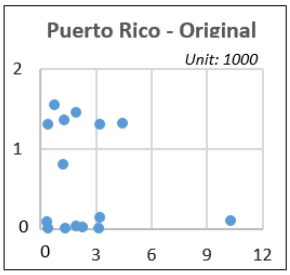
Graph 2-9



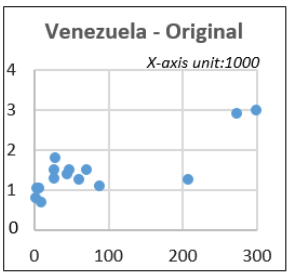
Graph 2-10



Graph 2-11



Graph 2-12



Graph 2-13

Graph 2-1 to 2-13: Distribution patterns between number of tweets and official new Zika cases (reference data) for each country. In all cases, the x-axis represent values associated to official data and the y-axis represent values associated to Twitter data. “Unit” applies to both x and y axis, while “x-axis unit” or “y-axis unit” only applies to the x-axis or the y-axis.

Country	Correlation coefficient	t-value	Degree of freedom	Significance
Aruba	0.41	1.624	13	80%
Brazil	0.17	0.603	13	<50%
Colombia	-0.19	-0.690	13	50%
Ecuador	-0.23	-0.839	13	60%
El Salvador	0.04	0.131	13	70%
Guatemala	-0.74	-3.955	13	98%
Haiti	0.03	0.108	13	60%
Honduras	-0.05	-0.192	13	70%
Jamaica	0.41	1.640	13	80%
Mexico	-0.15	-0.554	13	<50%
Panama	0.49	2.040	13	90%
Puerto Rico	-0.17	-0.637	13	<50%
Venezuela	0.82	5.125	13	≈ 100%

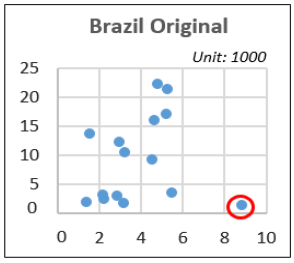
Table 5: Correlation test for the 13 selected countries between Twitter data and reference data from official new Zika cases.

Overall these results show a weak correlation between both datasets, with a few exceptions such as Guatemala that has a negative outcome (-0.74), Panama (0.49) and Venezuela that has the only significantly strong correlation coefficient of 0.82. Given the extensive presence of apparent outliers in the data, I conducted a second series of analysis after removing apparent outliers to see how this could improve the correlations.

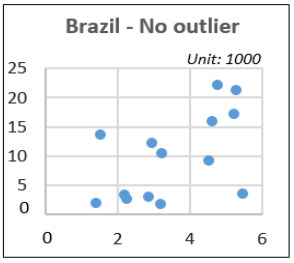
4.1.3 Removal of apparent outliers

This series of analysis aims to investigate how much correlation between twitter data and official Zika data could be improved by removing apparent outliers. Since I only have 15 values (i.e. weeks) for each data set, I decided to limit the number of apparent outliers removed to a maximum of 2 in order to keep enough data for some statistical analyses.

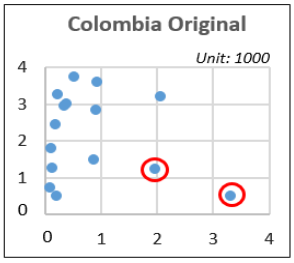
Graph 3-1 (a/b) to 3-9 (a/b) show results before and after the removal of statistical apparent outliers (note: no apparent outliers were removed from Aruba, Guatemala, Mexico and Venezuela since there were no obvious apparent outliers identified for these countries).



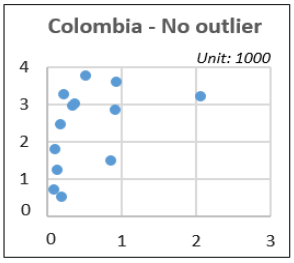
Graph 3-1 (a)



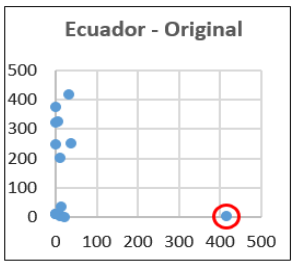
Graph 3-1 (b)



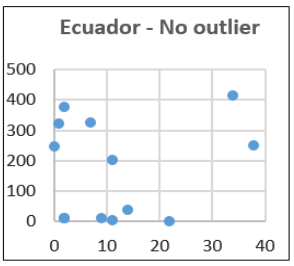
Graph 3-2 (a)



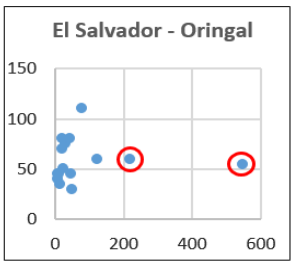
Graph 3-2 (b)



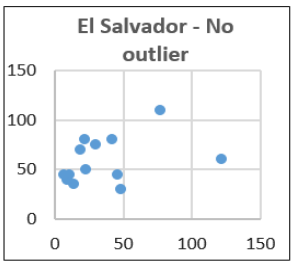
Graph 3-3 (a)



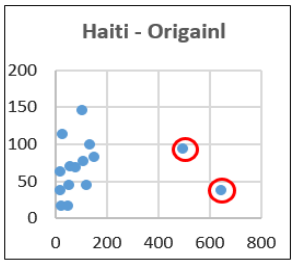
Graph 3-3 (b)



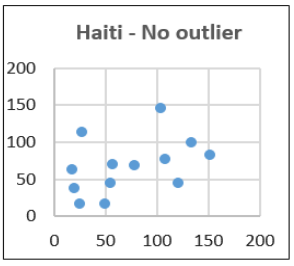
Graph 3-4 (a)



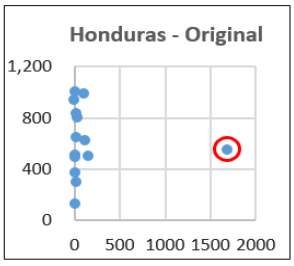
Graph 3-4 (b)



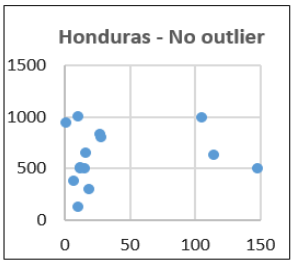
Graph 3-5 (a)



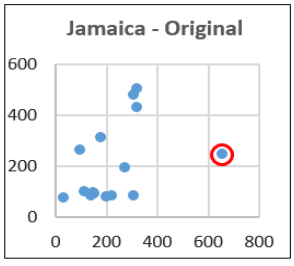
Graph 3-5 (b)



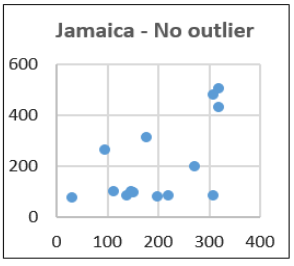
Graph 3-6 (a)



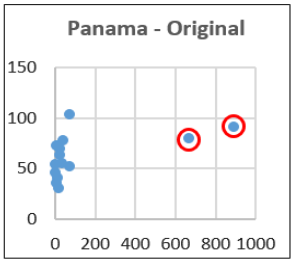
Graph 3-6 (b)



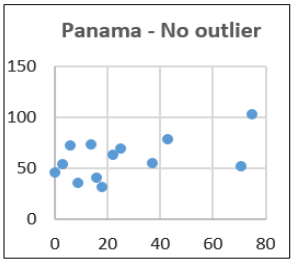
Graph 3-7 (a)



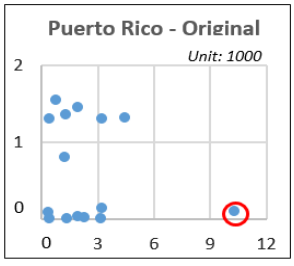
Graph 3-7 (b)



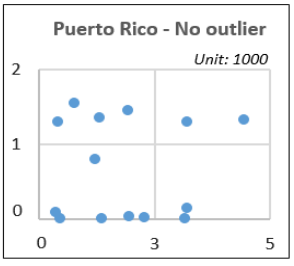
Graph 3-8 (a)



Graph 3-8 (b)



Graph 3-9 (a)



Graph 3-9 (b)

Tables 3-1 to 3-9: The comparisons before and after removing apparent outliers of the 9 countries with extreme apparent outliers. “Unit” applies to both x-axis and y-axis, while “x-axis unit” or “y-axis unit” only applies to x-axis or y-axis respectively.

The correlation test on the apparent outlier-removed data and how the correlations are improved after removal of the apparent outliers are listed in table 6.

Country	Original	Without apparent outlier	t-value	Degree of freedom	Significance
Aruba	0.41	N/A	N/A	N/A	N/A
Brazil	0.17	0.55	2.278	12	98%
Colombia	-0.19	0.43	1.597	11	80%
Ecuador	-0.23	0.21	0.746	12	50%
El Salvador	0.04	0.35	1.247	11	70%
Guatemala	-0.74	N/A	N/A	N/A	N/A
Haiti	0.03	0.42	1.515	11	80%
Honduras	-0.05	0.13	0.449	12	<50%
Jamaica	0.41	0.59	2.558	12	95%
Mexico	-0.15	N/A	N/A	N/A	N/A
Panama	0.49	0.51	1.95	11	90%
Puerto Rico	-0.17	0.04	0.153	12	<50%
Venezuela	0.82	N/A	N/A	N/A	N/A

Table 6: A comparison between the correlations of the original data, and the correlations of the data after apparent outliers are removed. The t-value, degree of freedom and significance is for the correlation test of the data after the removal of apparent outliers.

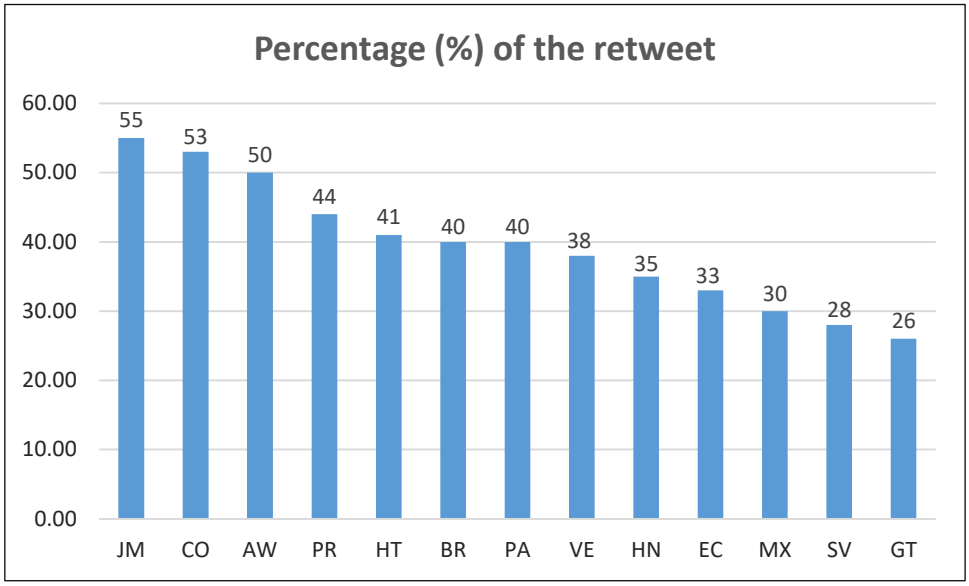
As it appears in the table above (table 6), most countries' correlations are strengthened after removing the apparent outliers. Although these results are expected, what is more interesting is that several of these correlations become stronger after removal of the apparent outliers. For instance, Brazil's correlation coefficient increases from 0.17 to 0.55 and Jamaica from 0.41 to 0.59. These results emphasize the impact of apparent outliers in Twitter data. In our case, apparent outliers occur in EWs where there are way more tweets than what would have been expected based on the official number of new cases during that week. The question then becomes: what are the elements that contribute to these generated apparent outliers?

To address this question, I have looked at the context of the tweets for each of the apparent outliers identified here. Of the 13 apparent outliers identified, I noticed that 8 were caused by retweets. For example, for Brazil EW 30 was identified to be an apparent outlier; during that week, it appears that a few thousand tweets about Zika were also mentioning the Olympic games and were extensively retweeted. Another example involves Panama: during EW 11 and EW 12, the number of tweets mentioning Zika and Panama exploded because of the fierce retweet of "*Panama reported the first case of microcephaly tied to Zika*".

These phenomena seem to indicate that retweets could be a potential cause for the statistical apparent outliers. As pointed out by Naveed and colleagues (2011), retweet reflects people's interest. It is based on the idea that people tend to retweet when they think their followers would be interested in the tweet's content. Thus, dramatic news such as "*Olympic Games is coming up but Zika is still out of control in Brazil!*" will be retweeted thousands of times which in return will contribute to the very high frequency-of-mention for Brazil and Zika as seen with EW 30. To assess the impact of retweets on the data generated through Twitter, I have ran another series of correlation analysis without retweets.

4.2 Removal of retweets

This section seeks to investigate the impact of retweets on the results. I first identified all retweets in the total number of tweets for each country except for Aruba, Guatemala, Mexico and Venezuela. I then conducted the same series of analysis as previously (section 4.1) to assess how the removal of retweets affects the correlation between data mined from Twitter and reference Zika cases for each country.

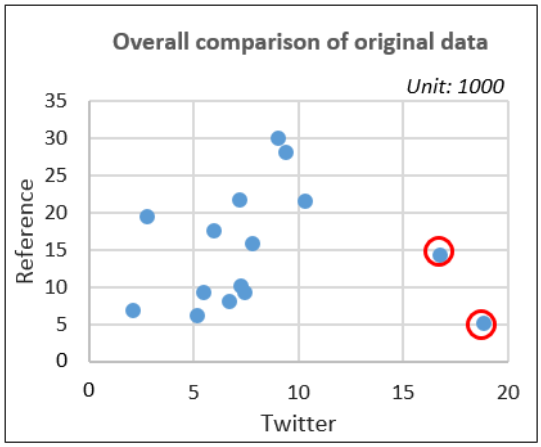


Graph 4: Total percentage of retweets per country. AW – Aruba, BR – Brazil, CO – Colombia, EC – Ecuador, SV – El Salvador, GT – Guatemala, HT – Haiti, HN – Honduras, JM – Jamaica, MX – Mexico, PA – Panama, PR – Puerto Rico, VE – Venezuela.

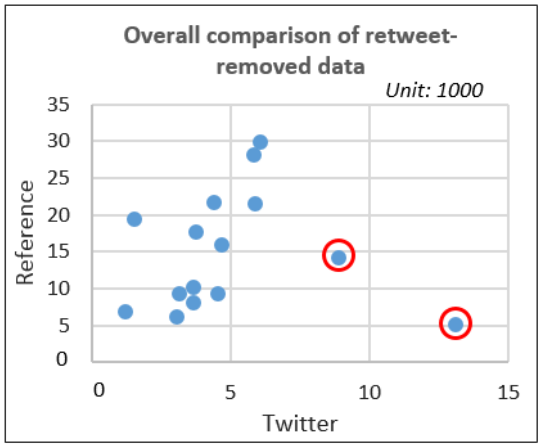
This graph shows that the percentage of retweets varies from 26% (Guatemala) to 55% (Jamaica), emphasizing the importance of retweets in my data. To assess the impact of these retweets on my results, I have removed them from the database from which apparent outliers had already been removed, and then conducted a similar series of statistical analysis.

4.2.1 Overall analysis

Graphs 5 (a) and graph 5 (b) show the total number of tweets for the 13 countries before and after filtering out the retweets, while table 7 shows the Pearson Correlation Coefficient test results before and after removing apparent outliers and retweets.



Graph 5 (a): The two points in red circles are outliers.



Graph 5 (b): The two points in red circles correspond to the outlier in the original data.

Data	Correlation coefficient	t-value	Degree of freedom	Significance
Original	0.01	0.029	13	< 50%
Retweets removed	0.01	0.041	13	< 50%
Original apparent outliers removed	0.56	2.233	11	95%
Retweets and apparent outliers are both removed	0.65	2.839	11	98%

Table 7: A comparison between the correlations before and after removing retweets and apparent outliers.

The results show that there is not much difference between the results with and without the retweets. In fact, the main criteria that influences the results is the removal of apparent outliers. It is also interesting to notice that once the apparent outliers are removed, removal of retweets significantly impacts the results (Pearson correlation coefficient increased from 0.56 to 0.65). To further understand this impact, I removed both the apparent outliers and the retweets for each country and conducted a series of correlation coefficient analysis.

4.2.2 Individual analysis

In this series of analysis, I have removed retweets from both my full data set and the data set without the apparent outliers identified previously. The goal of this analysis was to assess the impact of retweets on the results. See tables 8 (a/b) and table 9.

Country	Correlation coefficient without retweets	t-value	Degree of freedom	Significance
Aruba	0.39	1.534	13	>80%
Brazil	0.17	0.606	13	<50%
Colombia	-0.17	-0.623	13	<50%
Ecuador	-0.21	-0.771	13	>50%
El Salvador	-0.01	-0.036	13	<50%
Guatemala	-0.69	-3.404	13	>98%
Haiti	-0.08	-0.296	13	<50%
Honduras	-0.07	-0.067	13	<50%
Jamaica	0.13	0.458	13	< 50%
Mexico	-0.17	-0.617	13	<50%
Panama	0.48	1.989	13	>90%
Puerto Rico	-0.16	-0.594	13	<50%
Venezuela	0.69	3.419	13	>90%

Table 8 (a): The correlation test on retweet-removed data of 13 countries.

Country	Correlation coeff. with no retweets or apparent outlier	t-value	Degree of freedom	Significance
Brazil	0.56	2.34	12	95%
Colombia	0.58	2.272	11	95%
Ecuador	0.26	0.946	12	60%
El Salvador	0.12	0.413	12	<50%
Haiti	0.2	0.674	11	<50%
Honduras	0	0.005	12	<50%
Jamaica	0.16	0.55	12	<50%
Panama	0.44	1.604	11	80%
Puerto Rico	0.04	0.142	12	<50%

Table 8 (b): The correlation test on apparent outlier-removed data based on retweet-removed data. Aruba, Guatemala, Mexico and Venezuela are removed since they do not have identified apparent outliers.

Country	Original	No retweets	No apparent outliers	No retweets and apparent outliers
Aruba	0.41	0.39	N/A	N/A
Brazil	0.17	0.17	0.55	0.56
Colombia	-0.19	-0.17	0.43	0.58
Ecuador	-0.23	-0.21	0.21	0.26
El Salvador	0.04	-0.01	0.35	0.12
Guatemala	-0.74	-0.69	N/A	N/A
Haiti	0.03	-0.08	0.42	0.2
Honduras	-0.05	-0.07	0.13	0
Jamaica	0.41	0.13	0.59	0.16
Mexico	-0.15	-0.17	N/A	N/A
Panama	0.49	0.48	0.51	0.44
Puerto Rico	-0.17	-0.16	0.04	0.04
Venezuela	0.82	0.69	N/A	N/A

Table 9: Systematic comparison of the different correlation coefficients.

There are two points that can be made from this series of results. First, the impact of apparent outliers on the results is much more important than the impact of retweets, which implies that all the apparent outliers are not just due to retweets as hypothesized earlier in the thesis. Second, the

removal of retweets does not systematically improve the results of the analysis; while it seems to improve it for some countries (e.g. Colombia) it does the opposite for others (e.g. Panama and Jamaica).

Based on these different results, it is clear that apparent outliers are a key issue when trying to use tweets to assess the impact of a phenomenon, while the impact of retweets seems to be important in some cases but not overall. In other words, systematically removing retweets is not a good option to improve the correlation between a phenomenon (i.e. Zika new cases) and the volume of tweets related to the phenomenon and associated to a country name. In order to further explore how Zika related tweets could be utilized to better understand the spread of the disease, I have decided to run a qualitative analysis of my results for the only country that consistently show a strong correlation throughout the previous tests: Venezuela.

4.3 Venezuela case study

In light of the previous analyses, only the correlation between Venezuela's data mined from Twitter and the official cases appear to be consistently strong, compared to the correlations of data for the other countries. In this section, I further explore the reasons for this correlation in order to better understand how this specific case could reveal some elements for understanding the potential of using Twitter to study the spread of the Zika virus. This correlation is explored through two sets of analysis: The first set of analyses seeks to identify the link between the content of the tweets and the new official Zika cases, and the second set of analyses aims to identify potential individual Twitter accounts that might be a reliable source of data for identifying new cases.

4.3.1 Twitter context study

Given the large amount of tweets collected during the 15 weeks for Venezuela (1197 tweets), I have decided to focus my qualitative analysis on 5 EWs; three with a high volume of tweets (EW 10, EW 11 and EW 15) and two with a low volume of tweets (EW 13 and EW 16). These five weeks represent a total of 877 tweets.

I then read all the selected tweets to better understand their content and how they may be related to new Zika cases. Throughout this process, I quickly noticed that most of the tweets were referring to a small number of specific topics. For instance, during EW 10, 173 of the 299 tweets posted that week with the key words "Zika" and "Venezuela" were about the struggles of Venezuela to control Zika outbreak amid economic crisis. The most common topics and tweet contents are listed in table 10.

EW	Total number of tweets	Number of official Zika cases	Most common content (tweeted and retweeted)	Percentage
10	299	3000	Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis	58% (173)
			@PDChina: #Guangdong Province confirms 2 more new #Zika cases of a father and daughter traveled to #Venezuela	13% (39)
			@DrJaneChi: In Venezuela, Zika is on the rise; abortion is illegal; & condoms, if you can find them, cost up to $170 per 3-pack.	5% (15)
11	273	2900	RT @NewsBreaksLive: Zika Virus In #China: 34-Year-Old Man #Who Traveled To #Venezuela Quarantined https://t.co/WkiTjHDcMA	30%
			RT @nytimes: In Venezuela, basic	24%

			facts about Zika remain hidden. How a photographer is covering the story https://t.co/tJzejNhOLU	
			RT @alfonslopeztena: Venezuela 's economy to shrink 8%, inflation at 700 %, Caracas has one of highest murder rates, 400.000 have Zika.	18%
13	27	1250	For pregnant women in Venezuela, the possibility of getting the Zika virus is scary.	24%
			#Urgente centro #Barquisimeto full #zancudos #venezuela #zika https://t.co/rpI9vgVue1	21%
			New Guangdong ex-Venezuela #Zika Cluster #microcephaly https://t.co/YOhwH49w36	8%
15	207	1800	@shomaristone: #BREAKING: Dallas man who contracted #Zika in Venezuela transferred it to a male sexual partner, CDC says.	66%
			Zika Virus Can Be Transmitted Through Anal Sex, Too: A Texas man who had traveled to Venezuela passed the Zika... https://t.co/Ua7HmZwdpB	12%
			@VzlaBeg4Justice: #Venezuela. The Zika outbreak has exposed in the public health sector: misinformation, scarcity and govt mismanagement.	3%
16	71	1500	RT @XHNews: China will provide #Zika medical supplies to #Venezuela as humanitarian aid https://t.co/rvRxUCVuZ0 https://t.co/IwN8AGqh2t	67%
			@VzlaBeg4Justice: #Venezuela. The Zika outbreak has exposed in the public health sector: misinformation, scarcity and govt mismanagement	6%
			A Dallas man who contracted Zika in Venezuela transferred it to a male sexual partner after returning home in January, CDC says.	6%

Table 10: The number of tweets and new Zika cases in the sample EWs, and the top 3 most common tweet contents in each EW.

Looking at these results, we can identify a few interesting results: First, a very small number of tweets contribute to a large amount of the tweets collected for each given week. The 3 most common tweets always represent at least 50% of all the tweets for a given week (e.g. EW 13), and up to 2/3 of all the tweets for a given week in some cases (e.g. EW 16). None of these tweets talks only about Zika. In fact, for all of these tweets, Zika is associated with another phenomenon such as economic crisis (EW 10), as well as humanitarian and political issues (EW 16). Second, there are some categories that emerged from these results. Several tweets link the spread of Zika with economic crisis (EW $10^{\#1}$, EW $13^{\#3}$), while other tweets link Zika with the lack of public health service/management (EW $13^{\#1}$, EW $15^{\#3}$ and EW $16^{\#2}$) and with the spread of Zika from Venezuela to other places such as China or the United Sates. Furthermore, it is interesting to notice that some topics re-occur throughout the weeks. While some seem more informative such as the ones talking about the transmission of the virus, others are way more political such as the ones posted by @VzlaBeg4Justice, that emphasize the capacity of the government to manage the crisis (EW $15^{\#3}$ and EW $16^{\#2}$). These two main conclusions seem to imply that there is no direct correlation between the official cases and the content of the tweets released. Indeed this correlation seems rather random based on the content of the tweets selected. This conclusion is confirmed when we look at the content of tweets for the two weeks, EW 10 and EW 11, with the highest number of both tweets and official new cases. None of the tweets are related to new cases, but rather to social economic and public health issues. Based on this comparison, it appears that Zika serves to talk about broader issues at country level. Thus, even though EW 10 and EW 11 have a high number of Twitter that seem to correlate with the high number of official cases, the content of these tweets is unrelated to the record of new cases.

This qualitative analysis clearly demonstrates that the strong correlation between new official cases and Twitter activity is a coincidence. These results confirm the previous conclusion that Twitter data as collected throughout the process described in this project cannot serve to assess the spread of Zika at the country level. That said, in the last series of analysis, I want to see if it is possible to identify specific Twitter accounts that systematically provide relevant information about new Zika cases. As emphasized by Choudhury (2012), certain Twitter accounts provide valuable information on specific events/topics. So if Twitter cannot be mined systematically to track the spread of Zika, may be certain Twitter accounts could contribute to better follow this spread. In the last section I try to identify some of these accounts.

4.3.2 Informative account detection

In this section I propose to look systematically at all the Twitter accounts used to post tweets about Zika and Venezuela during all the 15 EWs of my project in order to identify if any of these accounts could serve as a reliable source of information to track new Zika cases. It is important to mention that all the accounts that are retweeted or mentioned in a tweet can be identified and that these accounts represent the large majority of the accounts used to post the tweets harvested in this project (See table 11).

Accounts \ EW	10	11	12	13	14	15	16	17	23	24	25	27	28	29	30
AEIfdp	■														
alfonslopeztena		■													
AudiByrneHaema										■					
CAllstadt	■														
cctvnews	■								■						
ceumedrelations										■					
CNNMoneyInvest								■							
DavidRoet		■													
Donte4419							■	■							
DrJaneChi	■	■													
elnacinews365		■													
fernandoTLMDO		■	■												
FraendyNewsman						■		■							
giustoyap	■														
glassmanamanda		■													
HealthcareBlvd	■														
HealthyAmerica1	■														
helpvenezuela99					■										
INVISTI														■	
isalara					■										
jornalistavitor	■														
julio_aliaga						■									
keophus	■														
la_pa247news			■												
LadaTweets		■													
MaryMurrayNBC	■														
Maxinflation		■													
MisaJC												■			
nataliabonilla			■												
nbcwashington						■		■							
NewsBreaksLive	■	■													
NPRHealth	■														
nprnews	■					■									
nytimes		■													
nytimesphoto		■	■												
PDChina	■								■						
people	■								■						
Rick_Permanand8	■														
Senator_Assange													■		
shomaristone						■									
SocialInDC						■									
SW7018	■														
TatoskyD						■									
TherapyForum	■														
trafficLARA				■											
VzlaBeg4Justice						■	■								
willcarless					■	■									
XHNews			■	■			■								
ZakenLife		■							■						
Zika_News	■	■													

Table 11: Presence (grey) or absence (white) of Zika + Venezuela related tweets for each EW.

The first surprising result is that none of the accounts identified posted tweets for more than three weeks about Zika and Venezuela, while new cases were reported during each of the 15 EWs (see table 11). Even the account named “Zika News” posted tweets about Zika and Venezuela for only 2 weeks (EW 10/11). In other words, none of these accounts could serve as the main source of information to keep track of the Zika outbreak in Venezuela. That said, I wanted to see if it could be possible to follow a selection of accounts that when combined together could provide an interesting source of information. To assess the potential of these different accounts combined, I have looked at the content of all the tweets mentioning “Zika” and “Venezuela” that were posted from these accounts (see table 12).

Account	**Tweet content**	**EW**
AEIfdp	#Venezuela already collapsing health care system seems poised to set off a regional humanitarian crisis #ZikaVirus	10
alfonslopeztena	Venezuela 's economy to shrink 8%, inflation at 700 %, Caracas has one of highest murder rates, 400.000 have Zika	11
AudiByrneHaema	evidence is Zika causes CNS defects - Rio, Venezuela, French Polynesia, Honduras; serious even if not the specific microcephaly case.	24
CAllstadt	via @npr: Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis https://t.co/AS9FqjuVE7	10
cctvnews	Guangdong Province reports 2 more new cases of #Zika virus, a father and daughter pair who traveled to Venezuela	10
	China's Guangdong Province reports new imported #Zika case: a 32-year-old patient returning from Venezuela	23
ceumedrelations	@SPPCEU's @BuxtonJulia in @ConversationUK on #Venezuela's #water #crisis: https://t.co/bsa89yTWe8 #ElNino #Zika #Human	24
CNNMoneyInvest	Life in #Venezuela: Blackouts, Zika, recession & now this: a 2-day work week https://t.co/DMd3SoISgL @Pat_Gillespie	17
DavidRoet	#Venezuela 's economy to shrink 8% , Inflation at 700 %, Caracas has 1 of highest murder rates & 400 k have Zika	11
Donte4419	A Dallas man who contracted Zika in Venezuela transferred it to a male sexual partner after returning home in January.	16, 17
DrJaneChi	In Venezuela, Zika is on the rise; abortion is illegal; & condoms, if you can find them, cost up to $170 per 3-pack.	10,11
elnacinews365	Venezuela Confirms Three Zika Deaths, 21 Suffering Related Nerve	11

	Disorder... https://t.co/PWhJsjlmUh https://t.co/np3kvJ	
fernandoTLMDO	Covering #Zika in Hushed-Up #Venezuela https://t.co/ZZq1fpX8Na @Sororita @marupita	11,12
FraendyNewsman	BREAKING: @CDCgov says Dallas man who contracted #Zika while in Venezuela transferred it to another man while they had sex	15,17
giustoyap	Three people have died in Venezuela from complications related to the Zika virus, President Nicolas Maduro said.	10
glassmanamanda	Covering #Zika in Hushed-Up #Venezuela https://t.co/x4oRPTpCUd via @nytimesphoto	11
HealthcareBlvd	DEPRESSION Venezuela medical shortages put Zika-linked Guillain-Barre cases at risk https://t.co/dyr797vmbl	10
HealthyAmerica1	Via @NPR: Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis https://t.co/oPQ9eQdQ0n	10
helpvenezuela99	While the government keeps a criminal silence, Zika spreads unchecked in Venezuela.	14
INVISTI	The New York Times cita sobre Venezuela. "Factbox: Why the Zika Virus Is Causing Alarm" por el escritor REUTERS	29
isalara	"Government's Secrecy Contributes To Zika Outbreak In Venezuela, Critics Say" by @JohnOtis in @MorningEdition	14
jornalistavitor	Venezuela faces worst-case scenario?as Zika outbreak expands https://t.co/6Gva1yftSA	10
julio_aliaga	LatAm News 14A https://t.co/VeoLAgwm0Y #Brazil #Venezuela #RevocatorioYA #Argentina #Mexico #Rio2016 #Cuba #USpoli #Zika	15
keophus	Zika swamps Venezuela ailing healthcare system https://t.co/vmIlBweJ4g None	10
la_pa247news	Report: 11 Dead of Crippling Nerve Disorder Tied to Zika in Venezuela... https://t.co/G2iMHPMWKp https://t.co/nm2ebJkCf7	12
LadaTweets	#Equador & #Venezuela: Experts find #US #CIA trace in L. American protests: https://t.co/WBg2woYNYe Told you so: https://t.?None	11
MaryMurrayNBC	Cuba reports 3rd case of Zika, a lab tech who had been working in Venezuela. https://t.co/PuZsT7crrS	10
Maxinflation	Estimated 400 000+ Zika cases in Venezuela?	11
MisaJC	And Puerto Rico, Zika rates astronomical over there RT @O_Dolly: Venezuela's economy, add that too. https://t.co/h8oF6JiK2k	27
nataliabonilla	Covering #Zika in Hushed-Up #Venezuela cc. @jorgejmuniz @samynemir https://t.co/pJDjmdC0NY via @nytimesphoto	12
nbcwashington	Dallas man who contracted Zika in Venezuela transferred it to a male sexual partner, CDC says. https://t.co/i8YzVnvsG0 None	15,17
NewsBreaksLive	Zika Virus In #China: 34-Year-Old Man #Who Traveled To #Venezuela Quarantined https://t.co/WkiTjHDcMA	10,11
NPRHealth	Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis https://t.co/EUoSCbEzvg	10
nprnews	#Venezuela Struggles To Contain #Zika Outbreak Amid Economic Crisis	10

	https://t.co/uepRyfxMid via @nprnews	
	Zika Virus Can Be Transmitted Through Anal Sex, Too: A Texas man who had traveled to Venezuela pas... https://t.co/NByOsHaz2P	15
nytimes	In Venezuela, basic facts about Zika remain hidden. How a photographer is covering the story https://t.co/tJzejNhOLU	11
nytimesphoto	Covering Zika in Venezuela, where the government has yet to proclaim a public health crisis https://t.co/yBOunWX9Tx	11,12
PDChina	#Guangdong Province confirms 2 more new #Zika cases of a father and daughter traveled to #Venezuela https://t.co/ODwflKzVSa	10
	#China's #Guangdong reports new imported #Zika case from a patient returning from #Venezuela https://t.co/5qGmguMNBR	23
people	Detroit Tigers pitcher who caught Zika Virus in Venezuela warns about travel to Rio Olympics https://t.co/maUuG9rZPF	23
Rick_Permanand8	Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis https://t.co/UlhRlMbDkk #RickPermanand	10
Senator_Assange	One odd effect of the Zika virus has been total amnesia in middle-class socialists who endlessly praised Venezuela.	28
shomaristone	Dallas man who contracted #Zika in Venezuela transferred it to a male sexual partner, CDC says.	15
SocialInDC	CDC Sees Same-Sex Zika Transmission: A Dallas man who contracted Zika in Venezuela transferred it to a male.	15
SW7018	Pregnant women in scarcity-hit Venezuela battle to dodge Zika: CARACAS (Reuters) - Carolina, who lives on the ...	11
TatoskyD	#Venezuela: Even for a normal case of Zika, which can involve a mild fever, a rash and joint pain.	15
TherapyForum	sharing #suaju Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis - NPR #outbreak https://t.co/zHPrCHZJsJ	10
trafficLARA	#Urgente centro #Barquisimeto full #zancudos #zika #Venezuela https://t.co/sDXG7MbQ2E	13
VzlaBeg4Justice	#Venezuela.The Zika outbreak has exposed in the public health sector:misinformation, scarcity and govt mismanagement	15,16
willcarless	ICYMI: Experts are watching Colombia & Venezuela with bated breath for signs of microcephaly linked to #Zika	14
	So, so far 2 suspected cases of microcephaly in Colombia, and 1 am in Venezuela. Tip of the spear?	15
	I still come back to: If #Zika causes #Microcephaly why aren't we seeing elevated #s of cases in Colombia/Venezuela? It doesn't?	
XHNews	China's Guangdong reports another #Zika case. Patient returned from Venezuela in Feb https://t.co/HlrVTOmb3q	12
	China reports 1st imported #Zika case. Man from Venezuela via HK, Shenzhen https://t.co/jvDDwnWe4A	13
	China will provide #Zika medical supplies to #Venezuela as humanitarian aid https://t.co/rvRxUCVuZ0 https://t.co/IwN8AGqh2t	16

	China provides #Venezuela with 96 tons of much-needed medicine against #Zika	23
ZakenLife	Venezuela takes on Zika amid shortages, information blackout - https://t.co/f61nPxftlk via @SocialMedia_Tea	11
Zika_News	Cuba announces first case of Zika, imported from Venezuela #health #google https://t.co/yJ40ZLUPbB https://t.co/1etF7uOJ88	10
	Venezuela's Meltdown Continues #Venezuela https://t.co/thQA4jsDaK https://t.co/m0oepfxCau	11
	Cumulative Locally Acquired Zika Cases by Country... #crisismanagement #Venezuela #Jamaica https://t.co/TVbhyiKfS7	

Table 12: The identifiable accounts and their tweet contents. The Twitter account are listed in alphabetical order.

These results confirm the previous ones: The contents of the tweets mentioning Zika and Venezuela are mainly about topics that are more or less directly related to Zika. For instance the account named HealthcareBlvd tweeted that the Zika-linked Guillain-Barre cases are at risk because of Venezuela's medical shortage. What is even more interesting is that when new cases are identified, it is often new cases that are related to Venezuela but that are emerging elsewhere such as in China's Guangdong province (e.g. tweeted by XHNews: *China's Guangdong reports another #Zika case. Patient returned from Venezuela in Feb*) or in Brazil, French Polynesia and Honduras (e.g. tweeted by AudiByrneHaema: *evidence is Zika causes CNS defects - Rio, Venezuela, French Polynesia, Honduras; serious even if not the specific microcephaly case*.). In fact, in several of these cases there are more than one place mentioned in the body of the tweet, emphasizing the difficulty of connecting Zika to one place/country using geoparsing.

Obviously, there are no accounts that consistently or regularly tweet about new Zika cases. Although, tracking several accounts at once to obtain the latest updates about new Zika cases seems feasible, the qualitative analysis shows that very few tweets talk about the new Zika cases. In fact, in

Twitter, Zika seems to be used to push forward some other more or less related agendas such as political issues: "*Venezuela already collapsing health care system seems poised to set off a regional humanitarian crisis #ZikaVirus*" (tweeted by AEIfdp) and "*Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis*" (tweeted by CAllstadt).

Throughout this case study and the qualitative analysis, I was able to demonstrate that the strong statistical correlation between the amount of Tweets containing Zika and Venezuela, and the number of new Zika cases for each week was in fact random. By looking into the content of the tweets, it becomes clear that none of these accounts were used to provide information about new Zika cases. In other words, based on this specific case study, it doesn't seem that Twitter is a relevant platform on its own to track the Zika outbreak neither by analyzing a large volume of tweets or by focusing on specific accounts.

V. Discussion and conclusion

This research employs both quantitative and qualitative measures to investigate the potential of using Twitter as a data source to help us better understand the spread of an epidemic disease. This series of analysis led me to draw three main conclusions that I will be discussing in this section: First, when doing analysis at the country-level, there is no obvious correlation between Zika-related tweets and official records of new Zika cases. Second, the word "Zika" appears often in tweets in which it is not the main topic of discussion. In fact, Zika often appears as a secondary topic used to push forward other agendas that could be either political, economic, social or humanitarian. Finally, I was not able to identify any Twitter accounts that would provide systematically relevant information about new Zika cases for any of the 13 countries under study.

The absence of a correlation between the number of tweets mentioning both Zika as well as a country name, and the official number of new Zika cases for the country in question was identified in this study through a series of statistical analysis. I first compiled a database of 1 million tweets mentioning Zika over a period of 15 weeks, then identified 13 countries that were often mentioned with Zika in the body of tweets using different geoparsing methods and tools. The 13 identified countries are Aruba, Brazil, Colombia, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Panama, Puerto Rico and Venezuela. The first series of analysis utilizes Pearson correlation coefficient test to provide an overview of the statistical relationships between my Twitter data and my reference data set. The test shows that most countries have weak correlation between the amount of Twitter data mentioning the countries and the number of official new Zika cases per week for those countries. These correlations often become stronger after removing one or two apparent outliers (i.e.

weeks containing extreme values in the Twitter data). This improvement suggests the possibility of a correlation between the volume of tweets mentioning Zika and new Zika cases, however this correlation is blurred by noise generated through Twitter. To reduce this noise, I removed retweets and ran another series of statistical analysis. While the simple removal of retweets didn't improve the results, the combined removal of both apparent outliers and retweets improved the correlation significantly, but not systematically for all the countries. In other words, it is possible to improve results by removing some of the noise in Twitter data, but the results are not systematically improved which makes it difficult to apply this approach in a systematic way to study the spread of Zika. My previous analyses also show that most countries have more than 40% retweets in their dataset, which means a considerable amount of tweets serve to amplify a phenomenon or news. These retweets could serve as an indicator to identify and study hot topics (Macskassy & Michelson, 2011).

What is even more interesting is that after the removal of retweets, it becomes obvious that the apparent outliers are caused by breaking news irrelevant to Zika, which supports my second main conclusion. When a breaking news mentions Zika and a given country name, the number of tweets mentioning this country rapidly increases in a certain week (EW). Due to the unpredictable occurrence of breaking news, the possibility of having apparent outliers in Twitter data is not just due to the amount of retweets but rather due to the overall interest in a breaking news generated in the Twittosphere. One of the solutions to filter out these breaking news from Twitter data could be applying text classification on tweets with machine learning. Text classification conducted by machine learning is able to group text into categories based on their contents (Zhang et al, 2015). This technology enables classifying tweets according to their contents and topics, then by identifying the

topic class with the highest number of tweets, breaking news could possibly be removed. A possible approach to alleviate this issue would be the use of more specific key words related to Zika symptoms such as, joint pain, conjunctivitis and muscle pain, to detect tweets indicating potential Zika-virus infected cases. This hypothesis had been tested during a H1N1 flu pandemic in the United Kingdom, and it yielded remarkable results (Lampos & Cristianini, 2010).

Finally, my last conclusion emerged from a series of qualitative analysis focusing on a specific case study that showed a very strong correlation between tweets mentioning both Venezuela as well as Zika, and new official Zika cases for Venezuela. To investigate the reasons of this correlation, I conducted a series of qualitative analysis on the 1197 tweets mentioning both Venezuela and Zika. Through this analysis, it appeared that in a large majority of these tweets, Zika is a secondary topic, as can be seen in the following example; "*Covering Zika in Venezuela, where the government has yet to proclaim a public health crisis*". What is even more interesting is that none of the tweet contents is directly related to new Zika cases, confirming the previous conclusion that Zika is often used in tweets to push forward other agendas. Moreover, as emphasized in the literature review, the multiplicity of tweets sometimes hides misleading information in terms of the topic discussed. Through this qualitative analysis, I was able to identify that the main topics associated to Zika are often political "*Venezuela Struggles To Contain Zika Outbreak Amid Economic Crisis*" and "*China will provide #Zika medical supplies to #Venezuela as humanitarian aid*". These examples illustrate how far the information mentioned in tweets is from the real spread of the disease. In other words, Twitter might be a better source of information to study how social media is utilized to push political agendas using epidemic disease, rather than a source of data to track the spread of an epidemic

disease. Finally I read all the tweets from my Venezuela sub-database to try to identify any accounts that might provide systematic tweets related to new Zika cases. Unfortunately none of the accounts was consistently tweeting about Zika and none of the tweets contents was directly related to new Zika cases. Even institutional accounts such as “nytimes” (New York Times) and “PDChina” (People’s daily, China) were not talking about new Zika cases. This last result confirms that even institutional accounts tend to report broader news than facts about the spread of the Zika outbreak. Although the original idea of this last part of my project was to see if it was possible to identify reliable Twitter accounts that would provide systematically relevant data about new Zika cases in order to use them to follow the spread of the disease, the results show that such accounts do not exist for any of the 13 countries under study for the time period of my analysis.

Limitations of the research and future studies

There are several limitations to this research. The first one is that I only mined tweets in English. Given the importance of Spanish and Portuguese in Latin America, using these languages to mine Twitter would have definitely improved the comprehensiveness of the database. This would have also required the use of a different geoparser since CLAVIN currently cannot recognize toponyms in Spanish or Portuguese. It would have been possible to extract toponyms in these two languages using semi-automatic geoparser, but it would have required a good understanding of these two languages which I do not have. Obviously geoparsers and gazetteers exist in these languages but it was beyond the scope of this project to try to utilize them to track the spread of Zika. Another important point is that the geoparsing process in this project could have been improved by using Natural Language Processing Tool, such as Gate. Among the multiple things that are possible with

Gate, there is the possibility of linking it to any gazetteers, such as Geonames, to geoparse place names in different languages and at different scales. Exploring further the potential of the geoparsing process to refine data analysis could definitely be of interest. Although I have done a broad range of statistical analysis in this project, it could have been interesting to focus more on "new Zika cases" from geolocated tweets and less on place names to see how relevant this data could be for our research. However there was very few tweets directly mentioning "new Zika cases" in terms of my data collected from Twitter which makes it impossible to conduct this kind of study. Another interesting area of research to pursue could be to further study how different places are connected in tweets. For example, in the Venezuela case study (section 4.3) different place names appear in individual tweets such as "China will provide #Zika medical supplies to #Venezuela as humanitarian aid." and "evidence is Zika causes CNS defects - Rio, Venezuela, French Polynesia, Honduras; serious even if not the specific microcephaly case." This type of content could be analyzed to better understand what kind of spatial connections that are generated via Twitter (i.e. different place names in one tweet) and how these connections sketch new types of digital spatial connections. Although it is clear that there is too much noise in Twitter data to be able to use it to accurately track the spread of a disease such as Zika, the richness and the quantity of this data could be exploited to better understand the relationships between the physical spread of a disease and its digital spread via social media data. Trying to understand "real" world problems via the prism of social media requires much more work as illustrated in this project and might probably require to begin with learning more about the real nature of the digital world on its own. I hope this thesis has made contribution to this field.

VI. References

About GPHIN,. (2017). Retrieved from https://gphin.canada.ca/cepr/aboutgphin-rmispenbref.jsp?language=en_CA

Amaral, L. A. N., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. Proceedings of the national academy of sciences, 97(21), 11149-11152.

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences, 106(51), 21484-21489.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013, September). TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In RANLP (pp. 83-90).

Cao-Lormeau VM, Roche C, Teissier A et al. Zika virus, French Polynesia, South Pacific, 2013. Emerg Infect Dis 2014; 20: 1084–1086.

Cheng, Z., Caverlee, J., & Lee, K. (2010, October). You are where you tweet: a content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 759-768). ACM.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one, 5(11), e14118.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS one, 5(11), e14118.

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. The American Journal of Tropical Medicine and Hygiene, 86(1), 39-45.

Coker, R. J., Hunter, B. M., Rudge, J. W., Liverani, M., & Hanvoravongchai, P. (2011). Emerging infectious diseases in southeast Asia: regional challenges to control. The Lancet, 377(9765), 599-609.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). The livehoods project: Utilizing social media to understand the dynamics of a city.

De Choudhury, M., Diakopoulos, N., & Naaman, M. (2012, February). Unfolding the event landscape on twitter: classification and exploration of user categories. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 241-244). ACM.

Duffy MR, Chen TH, Hancock WT et al. Zika outbreak on Yap Island, Federated States of Micronesia. N Engl J Med 2009; 360:2536–2543.

Ecker, D. J., Sampath, R., Blyn, L. B., Eshoo, M. W., Ivy, C., Ecker, J. A., ... & Hofstadler, S. A. (2005). Rapid identification and strain-typing of respiratory pathogens for epidemic surveillance. Proceedings of the National Academy of Sciences, 102(22), 8012-8017.

Edosomwan, S., Prakasan, S. K., Kouame, D., Watson, J., & Seymour, T. (2011). The history of social media and its impact on business. Journal of Applied Management and entrepreneurship, 16(3), 79.

Epidemic intelligence – systematic event detection,. (2017). Retrieved from http://www.who.int/csr/alertresponse/epidemicintelligence/en/

Faye, O., Freire, C. C., Iamarino, A., Faye, O., de Oliveira, J. V. C., Diallo, M., & Zanotto, P. M. (2014). Molecular Evolution of Zika Virus during Its Emergence in the 20 th Century. PLoS Negl Trop Dis, 8(1), e2636.

Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval: data structures and algorithms.

Frank, J. R. (2007). U.S. Patent Application No. 11/932,438.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., ... & Jombart, T. (2009). Pandemic potential of a strain of influenza A (H1N1): early findings. science, 324(5934), 1557-1561.

GermTrax,. (2017). Retrieved from http://www.germtrax.com/.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017). What's missing in geographical parsing?.

Hariharan, R., Hore, B., Li, C., & Mehrotra, S. (2007, July). Processing spatial-keyword (SK) queries in geographic information retrieval (GIR) systems. In Scientific and Statistical Database Management, 2007. SSBDM'07. 19th International Conference on (pp. 16-16). IEEE.

HealthMap,. (2016). Retrieved from http://www.healthmap.org/zika/#timeline

Infection control strategies for specific procedures in health-care facilities,. (2008, Feburary). Retrieved from http://www.who.int/csr/resources/publications/WHO_CDS_HSE_2008_2/en/.

Ivnitski, D., Abdel-Hamid, I., Atanasov, P., & Wilkins, E. (1999). Biosensors for detection of pathogenic bacteria. Biosensors and Bioelectronics, 14(7), 599-624.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

Lampos, V., & Cristianini, N. (2010, June). Tracking the flu pandemic by monitoring the social web. In Cognitive Information Processing (CIP), 2010 2nd International Workshop on (pp. 411-416). IEEE.

Liu, F., Vasardani, M., & Baldwin, T. (2014, November). Automatic identification of locative expressions from social media text: A comparative analysis. In Proceedings of the 4th International Workshop on Location and the Web (pp. 9-16). ACM.

Macskassy, S. A., & Michelson, M. (2011). Why do people retweet? Anti-homophily wins the day! Paper presented at the Icwsm, 209-216.

Malone, R. W., Homan, J., Callahan, M. V., Glasspool-Malone, J., Damodaran, L., Schneider, A. D. B., ... & Smith-Gagen, J. (2016). Zika virus: medical countermeasure development challenges. PLoS Negl Trop Dis, 10(3), e0004530.

Mandl, T., Gey, F., Di Nunzio, G., Ferro, N., Sanderson, M., Santos, D., & Womser-Hacker, C. (2008). An evaluation resource for geographic information retrieval. In Proceedings of the 6th Language Resources and Evaluation Conference. Sheffield.

Martins, B., & Calado, P. (2010, February). Learning to rank for geographic information retrieval. In Proceedings of the 6th Workshop on Geographic Information Retrieval (p. 21). ACM.

Microcephaly & Other Birth Defects,. (2016). Retrieved from https://www.cdc.gov/zika/healtheffects/birth_defects.html.

Miller, J. C. (2007). Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. Physical Review E, 76(1), 010101.

Naveed, N., Gottron, T., Kunegis, J., & Alhadi, A. C. (2011, June). Bad news travel fast: A content-based analysis of interestingness on twitter. In Proceedings of the 3rd International Web Science Conference (p. 8). ACM.

Neuzil, K.M., Hohlbein, C., Zhu, Y.: Illness among schoolchildren during influenza season: effect on school absenteeism, parental absenteeism from work, and secondary illness in families. Arch. Pediatr. Adolesc. Med. 156(10), 986–991 (2002)

Newman, M. E. (2002). Spread of epidemic disease on networks. Physical review E, 66(1), 016128.

Nicholson, A., Snair, M. R., Herrmann, J., Institute of Medicine (U.S.)., National Academies of Sciences, Engineering, and Medicine (U.S.),, & Global Health Risk Framework: Resilient and Sustainable Health Systems to Respond to Global Infectious Disease Outbreaks (Workshop).

(2016). Global health risk framework: Resilient and sustainable health systems to respond to global infectious disease outbreaks : workshop summary.

Noordhuis, P., Heijkoop, M., & Lazovik, A. (2010, July). Mining twitter in the cloud: A case study. In Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on (pp. 107-114). IEEE.

Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions),. (2017) Retrievel from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery, 9(3), 203-214.

Philipson, T. (2000). Economic epidemiology and infectious diseases. Handbook of health economics, 1, 1761-1799.

Plewe, B. (1997). GIS online: Information retrieval, mapping, and the Internet. OnWord Press.

Poser, K., & Dransch, D. (2010). Volunteered geographic information for disaster management with application to rapid flood damage estimation. Geomatica, 64(1), 89-98.

Russell, M. A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " O'Reilly Media, Inc.".

Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol, 7(10), e1002199.

Scanfeld, D., Scanfeld, V., & Larson, E. L. (2010). Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control, 38(3), 182-188.

Schmidt, C. W. (2012). Trending now: using social media to predict and track disease outbreaks. Environ Health Perspect, 120(1), 30-33.

Sutton, J., Palen, L., & Shklovski, I. (2008, May). Backchannels on the front lines: Emergent uses of social media in the 2007 southern California wildfires. In Proceedings of the 5th International ISCRAM Conference (pp. 624-632). Washington, DC.

The history of Zika virus,. (2016) Retrievel from http://www.who.int/emergencies/zika-virus/timeline/en/

Transimmion,. (2017). Trtrieved from https://www.cdc.gov/zika/transmission/index.html

Teutsch, S. M., & Churchill, R. E. (2000). Principles and practice of public health surveillance. Oxford University Press.

Travers, J., & Milgram, S. (1967). The small world problem. Phychology Today, 1, 61-67.

Twu, S. J., Chen, T. J., Chen, C. J., Olsen, S. J., Lee, L. T., Fisk, T., ... & Dowell, S. F. (2003). Control measures for severe acute respiratory syndrome (SARS) in Taiwan. Emerging infectious diseases, 9(6), 718.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. science, 312(5772), 447-451.

Yang, C., Yang, J., Luo, X., & Gong, P. (2009). Use of mobile phones in an emergency reporting system for infectious disease surveillance after the Sichuan earthquake in China. Bulletin of the World Health Organization, 87(8), 619-623.

Yates, R. B., & Neto, B. R. (2011). Modern Information Retrieval: the concepts and technology behind search. Addison-Wesley Professional.

Zhong, N., & Zeng, G. (2006). What we have learnt from SARS epidemics in china. BMJ: British Medical Journal, 333(7564), 389.