MULTIDIMENSIONAL SPATIO-TEMPORAL MODELING OF COMMUNICABLE DISEASES USING TWITTER POSTS AND HEALTH REPORTS: VISUALIZATION, CORRELATIONS, AND PREDICTION

A THESIS SUBMITTED TO
THE GRADUATE FACULTY OF
THE SCHOOL OF SCIENCE AND ENGINEERING
IN CANDIDACY FOR THE DEGREE OF
MASTER OF SCIENCE IN
COMPUTER SCIENCE

DEPARTMENT OF INFORMATION SYSTEMS AND COMPUTER SCIENCE

BY

KENNEDY E. ESPINA

QUEZON CITY, PHILIPPINES

2015

ABSTRACT

Infodemiology is a relatively new field of study where the margin for innovation is still large. Using infodemiology, the aim of the paper is to determine whether **tweets** can be an indicator if there is a disease outbreak within a certain area in the Philippines. Different methodologies will be used to get to the result. This includes data extraction, analytics, modeling, and visualization. The modeling of the diseases will be primarily made using IBM's Spatio-Temporal Epidemiological Modeler (STEM). Once a model and a scenario is created in STEM, a simulation of possible disease outbreak locations will be shown on a map based on user-specified timeframe. The results of this research will be especially helpful to the Philippines' Department of Health, as this study complements already existing surveillance system.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

Introduction

In the era of online social networking, it can be said that people are more expressive than ever since there is now an avenue to express publicly emotions (sentiments and reactions) and actions (behavior, location) in realtime [9]. Because of this, Twitter has become a prime candidate to learn more about the daily conditions of people. The immediacy of the contents from Twitter makes it an ideal source of data for analysis in areas such as disaster controls, marketing strategization, population profiling, health emergency tracking, and more [27, 23]. In the context of the Philippines, a country considered to be the social networking capital of the world, the potential of Twitter posts can be actualized as a layer to improve on existing epidemiological systems employed in the country.

"Infodemiology" is a term coined by Eysenbach when dealing with surveillance of public health-related text in the Internet [14]. It aims to measure and track the behaviour of information online with regards to public health. Social networking websites, such as Twitter, had proved themselves to be key to creating infodemiological studies after previous successful studies [21, 7, 5, 13]. The primary focus of this research is the infodemiology of three (3) of the top commu-

nicable diseases - Measles, Influenza, Typhoid Fever - in the Philippines, using Twitter [22].

This research is aligned with the *Philippine eHealth Strategic Framework and Plan 2020* that may help in strengthening existing eHealth solutions through adding to the *Information Sources* pillar of the framework. Through the use of ICT, the state of eHealth in the Philippines should be improved to provide even faster universal healthcare to people. The research may pave the way for complementing the usual way of tracking diseases that usually requires work that take up a lot of time [17].

The power of social media lies on the immediacy it has, and the amount of context a researcher could get from the strings of texts that users post. Twitter, for example, has a maximum limit of 140 characters per tweets. This limit helps in making it easier to filter subjects, sentiments, and data through the use of computational algorithms [24]. The extracted data may help researchers in finding out patterns in them, which may lead into creating different kinds of models based on them.

The Department of Health (DOH) is the government entity in charge of collecting and storing statistics regarding diseases in the country. The Epidemiology Bureau (EB) is an office under the DOH whose mandate is to develop and evaluate disease surveillance systems. Currently, the data they gather

are in public domain, and can be accessed easily through the DOH's website. Other sources of anonymized aggregated data can come from Electronic Medical Record (EMR) Systems being used by different private and public hospitals, if given access. The data collected, along with the spatio-temporal details (geo-location and timestamp) embedded in it, can serve as a baseline for possible correlation with the collected Twitter posts to make an infodemiological model. Using these data, the goal of this research is to produce a model that shows and predicts the conditions of various locations regarding the chosen diseases.

The research is divided into two phases, which are 1) creation of a twitter collection system for health related topics and visualization on a map, and 2) the modeling and prediction of possible areas of outbreaks based on Twitter posts. The major part of the initial phase is in the collection, pre-processing, quantification, and visualization of related tweets. This phase's primary concern is creating a listener for health related words relevant to infodemiology, such as symptoms and associated outbreaks. Users' sentiments will disregarded, as the model to be made focuses on classifying whether a tweet is infodemiological in nature or not. Bilingual keywords or phrases will be collected, which is vital since the Philippines uses two primary languages, the Filipino language and the English language. The second phase is where the modeling of the collected tweets will happen. An infodemiological model will be developed to identify possible lo-

cations of outbreaks based on current tweets through IBM's Spatio-Temporal Epidemiological Modeler.

Social networking websites had already been used by researchers to track different diseases like Human immunodeficiency virus (HIV) and Dengue [26, 17]. This research would follow some methodologies used by previous studies, but this would cater more to the Philippines since the data comes from the Philippines' domain. Added to this, the factor of bilinguality may add more needed appropriations to the study since Filipinos are highly likely to use social networking websites in both English and in Filipino.

1.1 Research Questions

This research aims to answer the question "Can social media posts from Twitter be one of the determinants of disease outbreaks within the Philippines?"

Other subquestions are as follows:

- What are related keywords and phrases frequently mentioned by Twitter users regarding the three diseases, especially when they are prevalent across an area?
- What attributes of social media posts are accessible and significant to be able to produce a population surveillance report?

- How do we interface tweets with syndromic surveillance data from existing systems of the Department of Health as well as from electronic medical records (EMR)?

- How do we develop a validated spatio-temporal epidemiological model that shows and forecasts possible outbreak areas based on Twitter posts?

1.2 Significance of the Study

The research is highly anchored on concepts of data mining, data analytics, and data visualization, particularly focusing on social media. Social media, being a tool used by millions of people around the world, can be a good indicator of people's behavior when proper analytics is implemented. For this research, a technique for data mining and analysis of health related information will be developed to improve on the sector of public health in the Philippines. This study builds on previous work that uses social media surveillance by adding a spatio-temporal layer of abstraction.

Added to data mining and data analytics, modeling and visualization will also be used in the development of the predictive functionality. The gathered and analyzed data from Twitter will be used as the parameters for modeling using IBM's Spatio-Temporal Epidemiological Modeler (STEM). The model that will be

developed will track how diseases actually spread using an accurate simulation not only based on spatial, but also on temporal data. Added to this, the research may also pave the way for other related diseases to be predicted.

The significance of the study also comes in the social impact that it can make. The tool developed can primary help the government, particularly the Department of Health, in making decisions regarding these diseases. One concrete way to use the results of the tool is through targeted prevention. The DOH usually deploys medical missions around the Philippines, and through the use of the tool, they can easily target those areas showing high probability of the diseases' outbreak. Early detection plays a crucial role if it is used by the government. Added to this, if proven accurate, the work of producing an up-to-date disease surveillance report can be done in a more regular basis. Through doing this, information-dissemination to the public is easier. People can then be informed immediately if there are outbreaks within their community, and take necessary precautions for themselves.

All in all, the study aims to produce and validate a visualized predictive infodemiological model that may help on the reduction of the diseases' incidences within the Philippines.

CHAPTER II

Review of Related Literature

This research concerns itself in public health analytics based on people's social media usage. The review of related literature is divided into five sections and proceeds as follows: **Communicable Diseases: Measles, Influenza, Typhoid Fever**, **Public Health Analytics Practices**, **Infodemiology**, **Data Mining and Visual Analytics Tools**, and **Spatio-Temporal Epidemiological Modeler.**

2.1 Infodemiology

Infodemiology, a portmanteau of *"Information"* and *"Epidemiology"*, aims to measure the pulse of people's public opinions, behavior and knowledge through tracking their online activities [15]. A formal definition of the word is *"the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy [14]."* On the other hand, the term "infoveillance" has also made its way to the terminologies researchers use. Infoveillance can be considered as a subset of infodemiology, where its focus is on the surveillance of the accuracy and trueness of reports that are being captured online. With misinfor-

mation being widely sent on the Internet [12], the use of infoveillance is needed to countercheck a data's validity as a possible source of health data to be used in researches.

For years now, researchers have already acknowledged that social media can be a source of significant information regarding its users such as basic profile, interests, and current condition. The rise of mobilility - enforced by ubiquitous access to the internet and mobile applications - provides users the avenue to frequently update social media websites, and at the same time supply even more analyzable data to researchers [27]. It has been used by for-profit companies to market products, by the government to gauge vote turnouts, and recently, by the health sector to track and predict diseases. This subtopic focuses on studies on infodemiology and infoveillance using social media, and how it may be implemented in the context of this research.

2.1.1 Infodemiology vs. Epidemiology

The word "Infodemiology" is based on the term "epidemiology", which is the study of the determinants and distribution of a disease within an area or population [14]. The main goal of epidemiology is to seek the pattern of distribution of a disease that helps in gaining the context from which a public health policy may be made [14]. Infodemiology, on the other hand, adds a new layer for deal-

ing with epidemiology in such a way that its focus is more on how information about a disease is gathered, more than how a disease is spreading. This means that infodemiology prioritizes more the changes on the gathered data's patterns, which in turn signals a change in the behavior of the people or the users [14]. This is the key to the epidemiology aspect of infodemiology, since a "pulse" in health-related posts may indicate the prevalence of health concerns within an area.

2.1.2 Infodemiological Analysis

Social media websites have become one of the easiest avenues for people to communicate with one another. The context of the posts being generated by its users have countless possibilities, and this is what makes it a good area of analysis. Added onto this are the metadata found inside each posts that contain spatio-temporal details about it. This means that the posts being generated have timestamps "attached" to it, and can be pointed to a certain geographical location around the world [6]. On the other hand, because of the large volume of user-generated content social media websites acquire daily, the challenge of getting the context in them can then become a hard task for researchers.

Advantages and Limitations of Infodemiology

The main advantage of using infodemiological methodologies is the relatively fast and easy collection of data [15]. Added to this, the dataset collected through using infodemiology may contain data that may have not been gathered through using traditional methodologies due to the wider search space, which is really the Internet. For this research, another advantage that infodemiology brings is the ability of social media websites to geo-tag posts. Twitter uses this mechanism, which then would help the research in isolating collected tweets within the Philippines, and into more specific locations.

Although advantages can easily be identified, disadvantages can also be seen. One limitation identified by Eysenbach is the difficulty in interpreting collected data. Since the Internet may pose "misinformation" [12], quality of collected data may vary. For example, in searching for the keyword "measles," results may not always indicate that a person has the said disease. This calls for proper analysis of the gathered data through healh analytics.

2.1.3 Social Media as a Source of Data

Social media is a great avenue for researchers to survey diseases in an immediate manner. Unlike normal pen-and-paper surveys, there is no need for direct contact from patients to make a population-based surveillance of diseases. With

petabytes of data within the reach of healthcare institutions, it is just a manner of extracting significant and leaving out insignificant data to arrive at a near-accurate survey [10].

In one study, a system was developed called Medical Ecosystem, or M-Eco. The M-Eco system is responsible for the collection of data from social media websites until the visualization process of the results [10]. *"Signals,"* the anomalies that users may find significant from the gathered data, were used to detect patterns. Signals can come in the form of symptoms the user of the system is looking for. The results then get extracted to get the spatio-temporal metadata from the social media post to be used in its mapping [10]. For this research, the same methodology can be utilized through the tool to be used in extracting social media posts.

Studies on Viral Diseases

This subsection is for the discussion of previously successful tracking and surveillance of diseases done in other parts of the world using social media websites.

A. Human Immunodefficiency Virus (HIV) - The study on HIV utilized Twitter's Advanced Programming Interface (API) to monitor posts suggesting sexual or drug-related activities. The method used in the study starts from the collection of the tweets, and eventually trimming it down to tweets with

HIV data [26]. The data collected are then cross-referenced with actual HIV data gathered from *aidsvu.org* [26]. The filtering of tweets is seen in Figure 2.1.

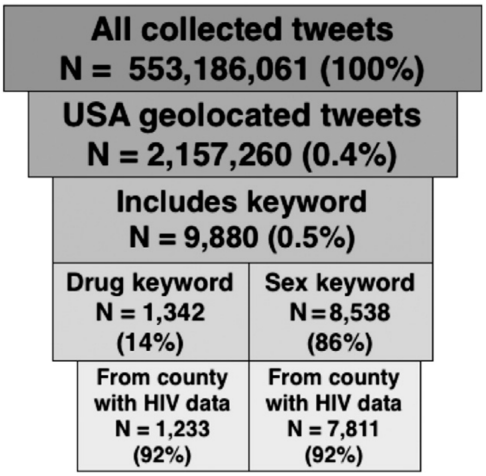

Figure 2.1: Twitter Data Filtering by Young et al [26].

The framework in Figurre 2.1 is one of the bases of this research [26]. From the collection of posts gathered, these data is then validated using aggregated public data that may come from the Department of Health. Through doing so, a correlation shall be identified to be used as a basis for the predictive model of the diseases.

B. Dengue - In the study made on Dengue surveillance in Brazil, four (4) parameters were used to filter social media posts, particularly tweets. These are *volume*, *location*, *time*, and *public perception* [17]. The same parameters are taken into consideration for the research on Measles, Influenza, and Typhoid Fever. Added to this, the research proposed five (5) taxonomies [Personal ex-

perience, Sarcastic tweets, Opinions, Resource, Marketing] to group the posts according to their sentiments [17]. For this research, sentiments are partially disregarded as this may yield to few tweets from Filipinos. If the need for sentiment analysis arises, it may be just to disregard sarcastic tweets, as mentioned in Gomide's research.

C. A[H1N1] - A study on how US-based social media users reacted to the A[H1N1] epidemic was also made to produce an infodemiological surveillance report [25].. Besides the initial filters using the words *"swine"*, *"influenza"*, and *"ah1n1"*, he also utilized a second filtering for other possible related keywords caused by the evolution of public concern. It shows a kind of filtering from specific to general, rather than general to specific. Some example search keywords include vaccines, flight details (to track movement), and even pork products, such as *"bacon"* [25].

2.2 Data Mining and Visual Analytics Tools

Data mining tools are used to get meaningful information from large amount of data. Big data usually consists of petabytes of data that's constantly growing over time. Tools are developed to sift through and organize these data, and these tools usually target data based on their sources. For this research, tools developed through Twitter's Advanced Programming Interface (API) is used.

2.2.1 Twitter Streaming API [4]

Social media APIs are known to provide developers access to the feeds of user-generated posts from different social media websites. For Twitter, they introduced *Twitter Streaming API*. This API allows developers to make tools going through the feeds, which leads to different tools especialized in different industries.

Twitter provides an all-in solution for developers who are only looking for Tweets, and not actually generating them. Twitter's streaming API provides developers three streaming endpoints, namely *"Public Stream"*, *"User Stream"*, and *"Site Stream"*. For the research, the main endpoint to be used is the Public Stream by individual users as this is mostly concerned for data mining for disease surveillance. For developers who want to generate tweets, such as those utilized in mobile applications, Twitter deployed its REST API.

2.2.2 Geocoding and the Google Maps Geocoding API

Geocoding is the act of transforming "aspatial locationally descriptive text" into a spatial representation, such as the latitude and longitude [16]. Geocoding can be considered as a subset of *Geographic Information Systems*, where geographic data are used for analysis [8]. Visualizations of GIS come in the form of maps. For public health and epidemiology, it is important to graphically rep-

resent disease registries on a map, making GIS an important tool to digitally analyze data. In the context of tweets, GIS and geocoding is specially helpful for locating tweets without the correct geo-location attached to them. Since users often include their addresses in their profiles, these can be used for geo coding. Although geocoding is a very helpful way in locating addresses, inconsistency may still happen [20]. This may be due to a number of factors, such as inputting misspelled street addresses and multiple places having the same name.

"The Google Maps Geocoding API is a service that provides geocoding and reverse geocoding of addresses." [1]. The Google Maps Geocoding API processes HTTP requests being thrown to it. The users have the option of choosing between *json* and *XML* format for its output.

2.2.3 R Programming Language and Environment

R is a statistical programming language that has also become a standard computational environment for many statisticians [19]. The R software is used in the research to compute for correlations between social media posts and the aggregated public health data. R will only be used as a supplementary tool as this may generate additional information in the modeling of the diseases.

2.3 Spatio-Temporal Epidemiological Modeler (STEM)

For this research, the Spatio-temporal Epidemiological Modeler (STEM) distributed by IBM through the Eclipse foundation is used to see the visualization of the gathered data [3]. STEM acts as a simulator of how diseases spread throughout an area, and through using this tool, the collection and analysis of data is made easier [3]. Through using STEM's default sets of epidemiology setups, it can help in giving the basis for the study of how outbreaks of the chosen diseases can be forecasted and mitigated.

STEM can handle custom experiments using different parameters. Developers can change a simulation's settings from the triggers of diseases, up to policies local to a country [3]. This is important for the research since the basis of the simulation is the gathered data from Twitter [3]. The software is extensible enough to cater this for the research.

2.3.1 Studies Using STEM

A group of researchers from the IBM Almaden Research center is heavily invested in producing researches using the Spatio-Temporal Epidemiological Modeler [2]. The group is headed by Mr. James Kaufman, one of the creators of STEM, who makes studies on the spread of infectious diseases. Two (2) of the researches they made are to be discussed. First is on Influenza focusing on the

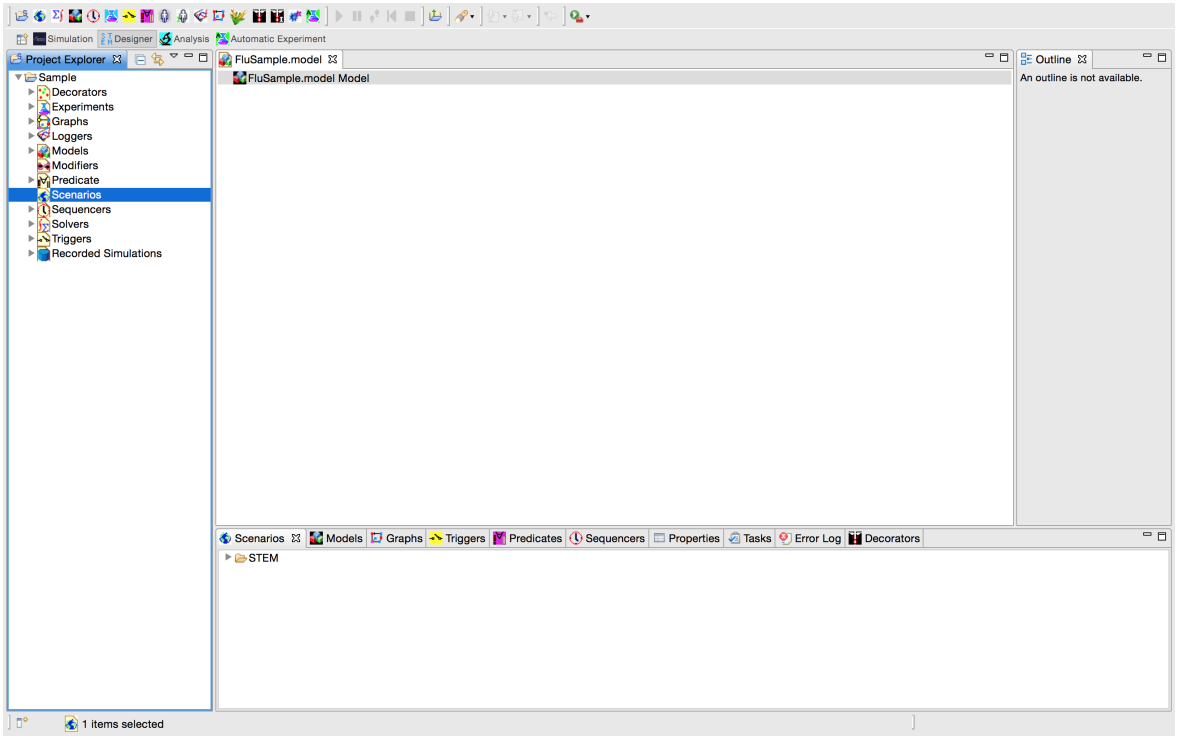

Figure 2.2: STEM's Integrated Development Environment by Eclipse.

SIR model used in predicting the disease in Israel, and on Dengue that shows the different compartmental models they used in modeling the disease.

Influenza

STEM was used to create a study entitled "A Spatiotemporal Model for Influenza" in the year 2011 [11]. The data used in the study focused on a 10-year data collected from 49 locations in Israel. The aim of the project is to create a predictive model using a deterministic SIR(S) compartmental disease model to account for seasonal variation of the disease [11]. The Susceptible-Infectious-

Recovered (SIR) compartmental model will also be used for modeling the diseases in this research, as supported by Edlund's research [11]. Through using this model, the population is assumed to be part of only one compartment at a certain point in time. Added to this, computations are used to adjust the predictive model from the data gathered.

As a conclusion, it was mentioned that the model they were able to make were accurate in the first two years of prediction, with a 3% error rate, but rapidly declined for the years after [11]. This shows that there are other factors that needs to be considered to be able to make a near-accurate predictive model of Influenza, and the social media abstraction may add value to this research.

Dengue

Another study led by the IBM Almaden Research Center focused on the dynamics of Dengue Fever [18]. The research focused on combatting Dengue fever reinfection by adding disease parameters such as the presence of multiple serotypes or strains of the virus. The study proceeded using a deterministic model, disregarding realistic stochastic forces outside the modelled environment. As a conclusion, the STEM software provided the researchers a platform to act as a baseline for their model.

2.3.2 STEM Loggers

STEM loggers stores simulation data into different formats. There are different loggers available in STEM, namely *CSV Logger*, *Map View Logger*, *Equirectangular Map Logger*, *Mercator Map Logger*, *Orthographic Map Logger*, and *Azimuthal Equidistant Map Logger*. Table 2.1 shows the difference among these.

Table 2.1: Comparison of STEM Loggers

Logger	**Compartment Logger**	Description
CSV Logger	Yes	Logs simulation data to a flat file using a configurable delimeter (such as a comma).,Useful for recording raw results of a STEM simulation for analysis.
Map View Logger	No	A simple logger for visually recording STEM simulations.,Captures the current view of the STEM Map and writes it to an image file.
Map Logger (Equirectangular, Mercator, Orthographic, Azimuthal Equidistant)	Yes	Highly configurable image drawers for capturing STEM simulations visually using various map projections.,Creates high resolution images using specific settings independent of the current STEM Map View.,Useful for creating production-quality STEM images for print and film animations.

s

For this research, the focus will be on the CSV logger. This contains the data that will be used for the re-creation of map using a separate tool.

CHAPTER III

Methodology

The primary focus of this research is the infodemiological surveillance of three (3) of the top communicable diseases - Measles, Influenza, Typhoid Fever - in the Philippines, as reported by the Philippines' Department of Health (DOH) [22]. The research is divided into two parts, which are 1) collection of health-related tweets and their visualization in a map, and 2) the modeling and prediction of possible areas of outbreaks based on Twitter posts.

This section breaks down the phases further into five (5) subsections, *Health Data Collection*, *Infodemiological Modeling and Classification*, *Twitter Data Visualization*, *Epidemiological Correlation and Modeling*, and *STEM Scenario Visualization*. The overview of the methodology can be seen in Figure 3.1.

3.1 Health Data Collection

The data that will be used for the research comes from two (2) sources, which are 1) publicly available individual posts collected from **Twitter**, which then will be validated through 2) publicly available statistical documents that largely leverages on the data by **Philippines' Department of Health (DOH)**. The users' posts from Twitter represents majority of the data to be processed, while

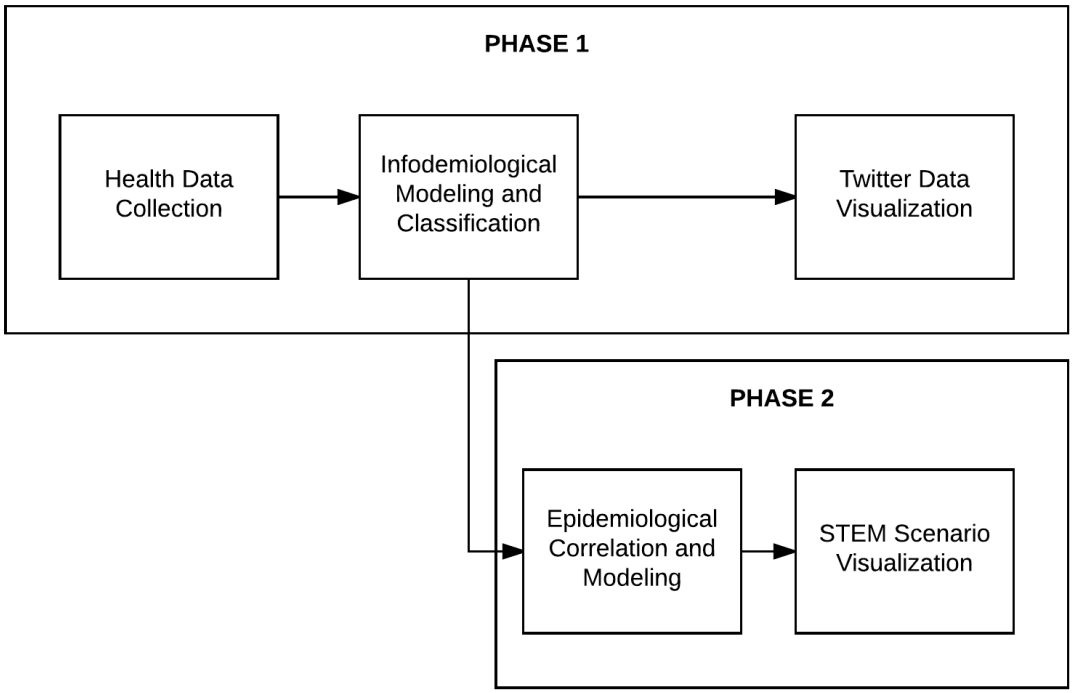

Figure 3.1: General view of the research's methodology.

the supporting records and documents act as a sustained supplementary data to be used as a baseline for comparison, correlation, and prediction.

3.1.1 Twitter Data Collection

Through the use of Twitter's public streaming Advanced Programming Interfaces (API), a tool will be used to gather relevant tweets through using keywords. The tool developed by the Ateneo Java Wireless Competency Center (AJWCC) is the chosen tweet collector that will be used. The tool is based on the *Ruby on Rails* programming language and framework, using the *tweetstream* gem. The

tool collects tweets in realtime based on the supplied keywords in its configuration file. It will be deployed on a local server with a constant connection to the Internet. This means that the collection will happen the whole day. The official start of collection for real-time tweet collection will be on *June 13, 2016* until *September 13, 2016*. Added to real-time collection, the research will also be accessing historical tweets through the AJWCC as well.

Keyword Compilation

Based on the related literature review, words associated to the disease will be gathered. Listed are possible keyword groups that may be found in social media posts.

- **Coloquial Terms** - Terms used by the people that may indicate the presence of flu.
  - Example: *"lagnat"*, *"flu"*, *"sick"*
- **Symptoms** - The symptoms of influenza. This includes both English and Filipino words.
  - Example: *"ubo"*, *"cough"*, *"sipon"*, *"sneeze"*
- **Behavior/Actions** - The behavior of people when they have the flu. For

example, people buy medicines during flu seasons, therefore drugstores, in general, may be used as a keyword.

  - Example: *"mercury drug"*, *"stay in bed"*, *"tulog"*, *"gamot"*

- **Weather Condition** - Lastly, the weather conditions of an area. Flu seasons arise when it's cold.

  - Example: *"rain"*, *"ulan"*, *"cold"*

These keywords, along with their conjugations if possible, will be used as the search parameters for Twitter.

3.1.2 Department of Health - Epidemiology Bureau Data

Regularly, the Philippines' Department of Health publishes through their websites their gathered disease surveillance statistics. This can be accessed by visiting *http://nec.doh.gov.ph/*. Since these data directly come from the DOH, these can be considered as a the gold standard for correlation and regression.

BIBLIOGRAPHY

[1] Getting started with google maps geocoding api. Web.

[2] Ibm research - almaden. Web.

[3] Public health research - the spatiotemporal epidemiological modeler (stem). Web.

[4] Twitter streaming api, 2015.

[5] Nicola Luigi Bragazzi. Infodemiology and infoveillance of multiple sclerosis in italy. *Mult Scler Int*, 2013:924029, 2013.

[6] Junghoon Chae, Dennis Thom, Yun Jang, SungYe Kim, Thomas Ertl, and David S Ebert. Public behavior response analysis in disaster events utilizing visual analytics of microblog data. *Computers & Graphics*, 38:51–60, 2014.

[7] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PLoS One*, 5(11):e14118, 2010.

[8] Ellen K Cromley and Sara L McLafferty. *GIS and public health*. Guilford Press, 2011.

[9] Ruth Deller. Twittering on: Audience research and participation using twitter. *Participations*, 8(1):216–245, 2011.

[10] K. Denecke, M. Krieck, L. Otrusina, P. Smrz, P. Dolog, W. Nejdl, and E. Velasco. How to exploit twitter for public health monitoring? *Methods of Information in Medicine*, 52(4):326–339, 2013.

[11] Stefan Edlund, Michal Bromberg, Gabriel Chodick, Judith Douglas, Daniel Ford, Zalman Kaufman, Justin Lessler, Rachel Marom, Yossi Mesika, Roni Ram, et al. A spatiotemporal model for influenza. 2009.

[12] Gunther Eysenbach. Infodemiology: The epidemiology of (mis)information. *Am J Med*, 113(9):763–5, Dec 2002.

[13] Gunther Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association, 2006.

[14] Gunther Eysenbach. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J Med Internet Res*, 11(1):e11, 2009.

[15] Gunther Eysenbach. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am J Prev Med*, 40(5 Suppl 2):S154–8, May 2011.

[16] Daniel W Goldberg. A geocoding best practices guide. *Springfield, IL: North American Association of Central Cancer Registries*, 2008.

[17] Janaína Gomide, Adriano Veloso, Wagner Meira Jr, Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, page 3. ACM, 2011.

[18] Kun Hu, Christian Thoens, Simone Bianco, Stefan Edlund, Matthew Davis, Judith Douglas, and James Kaufman. Modeling the dynamics of dengue fever. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 486–494. Springer, 2013.

[19] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[20] Nancy Krieger, Pamela Waterman, Kerry Lemieux, Sally Zierler, and Joseph W Hogan. On the wrong side of the tracts? evaluating the accuracy of geocoding in public health research. *American journal of public health*, 91(7):1114, 2001.

[21] Thiago D Nascimento, Marcos F DosSantos, Theodora Danciu, Misty DeBoer, Hendrik van Holsbeeck, Sarah R Lucas, Christine Aiello, Leen Khatib, MaryCatherine A Bender, Jon-Kar Zubieta, et al. Real-time sharing and expression of migraine headache suffering on twitter: A cross-sectional infodemiology study. *Journal of medical Internet research*, 16(4):e96, 2014.

[22] Department of Health. Reported cases of all reportable diseases and conditions by region. PDF, June 2015.

[23] William Ribarsky, Derek Xiaoyu Wang, and Wenwen Dou. Social media analytics for competitive advantage. *Computers & Graphics*, 38:328–331, 2014.

[24] Charles W Schmidt. Trending now: using social media to predict and track disease outbreaks. *Environ Health Perspect*, 120(1):30–33, 2012.

[25] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS One*, 6(5):e19467, 2011.

[26] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Prev Med*, 63:112–5, Jun 2014.

[27] Daniel Zeng, Hsinchun Chen, Robert Lusch, and Shu-Hsing Li. Social media analytics and intelligence. *Intelligent Systems, IEEE*, 25(6):13–16, 2010.

APPENDIX A

Geocoded XML of "*Metro Manila*" using Google Maps Geocode API


```
<GeocodeResponse>
    <status>OK</status>
    <result>
        <type>administrative_area_level_1</type>
        <type>political</type>
        <formatted_address>Metro Manila, Philippines</formatted_address>
        <address_component>
            <long_name>Metro Manila</long_name>
            <short_name>NCR</short_name>
            <type>administrative_area_level_1</type>
            <type>political</type>
        </address_component>
        <address_component>
            <long_name>Philippines</long_name>
            <short_name>PH</short_name>
            <type>country</type>
            <type>political</type>
        </address_component>
        <geometry>
            <location>
                <lat>14.6090537</lat>
                <lng>121.0222565</lng>
            </location>
            <location_type>APPROXIMATE</location_type>
            <viewport>
                <southwest>
                    <lat>14.3493861</lat>
                    <lng>120.9172569</lng>
                </southwest>
                <northeast>
                    <lat>14.7812170</lat>
                    <lng>121.1320120</lng>
                </northeast>
            </viewport>
            <bounds>
                <southwest>
                    <lat>14.3493861</lat>
                    <lng>120.9172569</lng>
                </southwest>
                <northeast>
                    <lat>14.7812170</lat>
                    <lng>121.1320120</lng>
                </northeast>
            </bounds>
        </geometry>
        <partial_match>true</partial_match>
        <place_id>ChIJbTgmYNLIlzMR0HiSrNoj7V8</place_id>
    </result>
</GeocodeResponse>
```