THE AUGMENTATION, POTENTIAL, AND PRACTICALITY OF TWITTER DATA FOR PREDICTING INFLUENZA EMERGENCY ROOM ADMISSIONS

By

Joshua J. Vertalka

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Geography - Doctor of Philosophy

2018

ABSTRACT

THE AUGMENTATION, POTENTIAL, AND PRACTICALITY OF TWITTER DATA FOR PREDICTING INFLUENZA EMERGENCY ROOM ADMISSIONS

By

Joshua J. Vertalka

Every year, millions of people become infected with one of the many seasonal influenza viruses. These infections may have dire consequences as local hospital Emergency Rooms (ERs) experience sudden surges of influenza patients, causing ambulance diversions and shortages of medical supplies. Current influenza surveillance techniques lack the necessary spatial and temporal fidelity to benefit local hospital systems. This dissertation helps correct that issue through three chapters. Chapter one identifies an approach to augment social media data using the Digital Interaction Program (DIP). DIP uses application program interfaces to digitally converse with and seek social media users' participation in an online questionnaire. This questionnaire is designed to collect spatial and temporal data and augment social media data, such as demographic information. Chapter two uses DIP to identify where and when influenza tweets posted across New York City and London at fine spatial and temporal scales. It was found that on average influenza tweets tend to occur closer to a user's home ZIP Code, in comparison to those users' non-influenza tweets. Therefore, this information suggests that influenza tweets can predict influenza cases at a finer geographic scale than current research suggests. Influenza tweets are most often posted when a user is experiencing peak symptoms, not symptom onset. Finally, Chapter three of this research tests if, when, and to what degree influenza tweets can predict local hospital ER admissions. It was found that most hospitals can use influenza tweets to predict influenza ER admissions on average of about eight days

advanced. Chapter three speculates that influenza tweets have the potential to identify influenza propagation between the different age groups in New York City. Therefore, Twitter has the spatial and temporal potential to provide a more timely and spatially accurate influenza surveillance system that is focused on local hospital systems.

Dedication to
My family, Mrs. Terri Vertalka, Ms. Alaina Vertalka, Ms. Alexis Vertalka, and Ms. Catherine Terpstra
My Family and Friends

ACKNOWLEDGEMENTS

At long last! I once more have the privilege of acknowledging my gratitude towards my family, friends, and colleagues and I'm even more grateful than ever before.

To my mother, *Terri Vertalka*, thank you for always asking about my 'magnificent' project. Though at times it was very frustrating to me, I am grateful you remained curious about how my dissertation was going and how I was doing. To my two sisters, *Alaina* and *Alexis Vertalka,* it is an absolute privilege to have you as my sisters. I couldn't imagine anyone else!

To my amazing partner in crime, *Catherine Terpstra*, thank you for patiently waiting for me to complete this dissertation. Despite the many frustrations a dissertation brings forth, you were at my side. Thank you! I can't wait to see what happens next.

I would like to thank my *Uncle Bob and Aunt Judy Vertalka* for being supportive family in the area. It was great having the many diners and lunches together!

Next, I would like to thank my advisor, *Dr. Eva Kassens-Noor*. All things that are innovative require entering a world of uncertainty. Thank you for not being too frustrated with my meandering through this uncertain world.

To *Dr. Ashton Shortridge*. Thank you for sitting down and talking with me about a number of different things from the dissertation to employment advice. Those conversations meant a lot to me!

I want to next thank *Dr. Igor Vojnovic* for keeping me focused on adding more theory to my dissertation. I also want to thank Igor for always enthusiastically teaching class.

To *Dr. Theresa Bernardo*. Thank you for providing the medical knowledge needed to complete this dissertation.

To the entire *Geography Department*. You are my academic family, one that will be greatly missed! Thank you for all the great times and memories!

My *dearest friends,* the joys, frustrations, and pitfalls we have shared together in the last few years will never be forgotten.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The primary research question this dissertation seeks to answer is, what is Twitter's role in predicting influenza ER admissions in New York City and London? This dissertation contains three chapters that each focus on a separate but interlocking theme to help answer that question.

Chapter one proposes and develops a new approach that augments social media data sources by altering the way scholars interact with social media (SM) users. Unclear SM data, including Twitter, are being used in scientific research. Rather than using SMs' application program interfaces (API) solely to gather SM data, as current research employs, Chapter one introduces a new approach that uses SMs' API to send messages to users and ignite digital conversations between the scholar and the user. The digital conversation invites users to complete an online questionnaire with questions pertinent to the research objective. Therefore, this data is used to augment missing social media data, such as demographic information and help alleviate the undirected nature of users posting content. Chapter one focuses on gathering intellectual input from Volunteered Geographic Information (VGI) contributors, in this case SM users. Traditionally, SM as VGI involves SM users solely acting as passive sensors, gathering and reporting information of the world but not actively contributing to the scientific process. Chapter one provides an approach that gathers intellectual input from SM users, thus enhancing their role in VGI.

Chapter One has two theoretical contributions. First, it provides a link between the traditional and modern world of surveys. Traditionally, surveys have been conducted through the mail or phone. As time passed, the invention of internet and mobile technology decreased mail and phone survey response rates and increased alternative survey modes. Most often the findings from these alternative approaches were not centered on the betterment of scientific knowledge. Chapter One represents a shift in the manner surveys are conducted with a very mobile population that creates very unclear data since it views technology as not a barrier but rather a portal to communicate with people. Second, this Chapter's theoretical focus centers on the role of the VGI contributor as it relates to scientific knowledge. The role of VGI contributors continues to evolve given that new undefined VGI data sources are constantly emerging. For instance, VGI was initially defined as untrained citizens collecting and sharing spatial data through OpenStreetMap. However, given the explosive growth of digital data, VGI definitions were quickly altered to include the different ranges of data users made. Chapter One seeks to expand the role of SM users in the scientific process. As evident below, several Twitter users provided survey-based insight in addition to their tweet for a scientific purpose. bu

Chapter Two focuses on identifying the space-time path differences between Twitter users experiencing influenza and users not experiencing influenza. Twitter represents a new data source to record activities of individuals through space and in time. Traditionally, measurements of space-time paths required individuals to record a diary of their spatial location at specific times and what activity they were doing. Chapter two uses the DIP framework to augment the use of Twitter data to identify when and where users' influenza tweets are posted to those same users' non-influenza tweets in New York City and London. Research has yet to identify where influenza tweets tend to occur in relation to the author's residence. Prior research has generally assumed that users tweeted about their influenza symptoms from home. This assumption has yet to be scientifically explored. The findings of Chapter two suggests this assumption is correct as users tend to tweet about their symptoms closer to their home ZIP Code and Postal Code meaning that, Twitter can be used to predict influenza cases at neighborhood scales. Scholars have also assumed that Twitter users are more apt to tweet about their symptoms early in the infection period. However, Chapter two finds that users tend to tweet about their influenza symptoms during periods of heightened symptoms. This also suggests that research using Twitter data to analyze space-time paths needs to account for the user's state of health as it will likely alter their typical space-time activity behavior.

Chapter three focuses on contributing to the digital disease detection realm. Current digital disease detection has been spatially limited to broad regional or metropolitan scales. This broad unit of analysis paints a timely picture of national, state, or city-wide influenza activity but provides little benefit to individual hospitals at the local scale. Chapter three builds on the findings of Chapter two by testing if and to what spatial and temporal ability influenza tweets can predict influenza ER admissions at local hospitals in New York City (NYC). Not surprisingly, the relationship of influenza tweets predicting influenza ER admissions varies from hospital to hospital. Influenza ER admissions at a majority of the hospitals in NYC are best correlated with influenza tweets when influenza tweets are posted eight days before actual ER admissions. In some cases, influenza tweets are best correlated with ER admissions when influenza tweets occur three weeks before ER admissions. For other hospitals, influenza ER admissions proceed influenza tweets. No research has examined the relationship between influenza tweets and local influenza ER admissions. While the findings did not show a strong relationship, it is clear that there is some potential for Twitter to act as an augmenting data source to traditional influenza surveillance.

The different correlation relationship between influenza tweets and influenza ER admissions represents a possible contagion effect of the influenza virus between different age groups. While Chapter three does not test the contagion effects between when a user posts influenza tweets and when local ER influenza admissions occurs it does provide a framework for future research to test.

CHAPTER 1
MORE THAN A HASHTAG: AUGMENTING SOCIAL MEDIA DATA THROUGH APPLICATION PROGAM INTERFACES AND ONLINE QUESTIONNAIRES

Abstract

Large quantities of public members on social media (SM) outlets discussing a wide range of topics popularized it as a data source for a variety of research topics. Much of this research treats SM users solely as passive sensors that gather information about the world around them much like weather stations or other nonhuman monitors. This research introduces and develops a method, called the Digital Interaction Program (DIP), that invites SM users to actively provide intellectual input for scientific purposes. DIP engages SM users into a digital conversation and augments their passively collected data (e.g., content, geographic location, time of posting). DIP first identifies potential research participants, then uses Application Program Interfaces (APIs) to recruit research participants to complete an online questionnaire. The online questionnaire asks research participants questions that augment the traditional sensory data provided by SM users. As a result, SM users become more than just sensors: they become contributors to a number of different applications including research, urban planning, disaster response, and other applications utilizing SM data. To showcase DIP, an example of its implementation is provided.

Introduction

As of 2016, over 65% of internet using adults are on one or more social media platforms such as Twitter, Facebook, Instagram, and Pinterest (Pew Research Center, 2016). Users are encouraged to post timely descriptions about themselves, others, and the world around them through text, pictures, videos, maps, and other communication medias. Users post these descriptions in real-time and with self-reported location information or geographic coordinates provided by GPS enabled devices. In essence, these users are social sensors through space and time (Elwood, 2008; M. F. Goodchild, 2007b; Sakaki, Okazaki, & Matsuo, 2010). For these reasons, researchers pair SM data with crowdsourcing tools (classification algorithms, natural language processing, regressions, and other prediction or classification tools) for event predictions, modeling trends, and monitoring activities (Daume & Galaz, 2016). Table 1.1 lists some examples of applied SM research.

Table 1.1: Different Research Topics using SM Data

Monitoring Activities	Prediction and Trends
● the 2007 Virginia Tech shootings using Flickr, Facebook, Myspace, Second Life (Palen, Vieweg, Liu, & Hughes, 2009) (● 2007 California Wildfires using , Flickr, personal blogs, and Twitter (Sutton, Palen, & Shklovski, 2008) ● 2009 Red River floods Oklahoma City Fires using Twitter (Vieweg, Hughes, Starbird, & Palen, 2010) ● 2010 Haiti Earthquake using Twitter (Sarcevic et al., 2012) ● 2011 Egyptian uprisings using Twitter (Starbird & Palen, 2012) ● 2011 UK riots using Twitter (Denef, Bayerl, & Kaptein, 2013) ● 2011 Great Japan Earthquake used Sina (the parent company to Sina Weibo, a Twitter like service) (Yang, Wu, & Li, 2012) ● 2012 Hurricane Sandy used Twitter, Nixle, and Facebook (Hughes, St Denis, Palen, & Anderson, 2014) ● 2013 Boston Marathon Bombing used Twitter (Starbird, Maddock, Orand, Achterman, & Mason, 2014)	● stock market with Twitter (Bollen, Mao, & Zeng, 2011) ● earthquakes through Twitter (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013) ● soccer matches with Twitter (Yu & Wang, 2015) ● presidential elections through Twitter (Tumasjan, Sprenger, Sandner, & Welpe, 2010) ● crime with Twitter (Gerber, 2014) ● digitally detecting diseases Twitter (Bodnar & Salathé, 2013; Salathe et al., 2012) ● Advertising legacies with Twitter (E. Wright, Khanfar, Harrington, & Kizer, 2010)

This research begins by providing an overview of relevant literature on the uses and shortcomings of Volunteered Geographic Information and the participatory role of citizen scientists. Then SM Application Program Interfaces are described followed by a detailed description of Digital Interaction Program (DIP). To showcase the process, application, and challenges of DIP, an example is provided that applies DIP to augmenting influenza infected

Twitter users in New York City and London. Finally, this paper discusses how DIP changes the participatory role of SM citizen scientists, DIP's role in a range of SM applied research, limitations, and future progress of SM augmentation.

Background

Survey methods in the last Century can be binned into three different eras. The first era involved the basic components of survey design and collection methods. The second era was rapid development in survey use as the federal government begin using survey to help gather data on infrastructure investment. During the first two eras, surveys were mostly done through mail or phone calls with very high response rates. The third era of surveys began when technological advancements created the internet and mobile communication. These inventions caused a decline in survey response rates and weakening of sampling methods but caused a growth in unique forms of data collection and the collection of continuous data through technological mediums (Groves et al., 2011). For instance, Netflix, a popular online video streaming site, does not use a mail or phone survey to gain insight into which videos users enjoy. Instead, they focus on capturing users video preferences when a user ‘rates’ a video through a 1-5 system. This is similar to the Likert Scale where users gauge how well they enjoyed the video; where a 1 represents the video was not enjoyable, 3 is average enjoyability, and 5 represents a very enjoyable video. The backbone of this paper rests on a digital survey that collects data from nontrained individuals for a scientific purpose.

The collection, production, and dissemination of spatial data has traditionally been a top-down approach in which experts created datasets using highly precise and expensive measuring instruments. The accuracy, bias, precision, and error of this data are scientifically tested using a number of approaches (root mean square error, ground truth, cross-validation, etc.) (Feick &

Roche, 2013). In the end, this scientific process produced robust data, but only about very specific phenomenon within a small study area or very broad phenomenon in a large study area. An example of the latter case includes Census data production. Consequently, individuals could neither afford nor had the knowledge to create these datasets. Instead, data creation (field studies, surveys, measurement samples, etc.) responsibilities fell almost exclusively upon the scientific community or those considered experts in data construction (Connors, Lei, & Kelly, 2012; Haklay, 2013; Sui, Elwood, & Goodchild, 2012).

Technological advancements, including the internet and telecommunications, introduced a bottom-up approach of spatial data creation by non-expert volunteers. This type of data has been referred to as Volunteered Geographic Information (M. F. Goodchild, 2007a). Affordable GPS enabled devices, such as smartphones allow non-experts to record spatial data about a number of different things. For instance, a non-expert might purposely collect the location and color of every fire hydrant on their street by using their smart phone. The non-expert, referred to as a citizen scientist, then contributes this data onto a community repository server that stores and shares VGI data with other citizen scientists. This process can be done with little to no scientific knowledge or data collection background (Elwood, 2008; M. F. Goodchild, 2007b; Sakaki et al., 2010).

Data quality issues arise, however, when citizen scientists are generating VGI. Since citizen scientists have little to no experience in data collection, the data they do collect may be prone to error. Citizen science driven data errors, however, are often corrected through crowdsourcing, or consensus of the masses. For example, one user may incorrectly label the location of a fire hydrant, but other VGI contributes will correctly record the fire hydrant’s location, correcting any mislabeling. In some instances, VGI has been as accurate or nearly as accurate as traditional authoritarian data sources (M. F. Goodchild & Li, 2012).

The advent and popularity of social media outlets established a need to revisit the definitions of VGI in late 2000s and early 2010s. Early definitions of VGI centered on an army of individuals volunteering their time to collect the location and attributes of earth’s features, mostly to create mapping-based projects, such as OSM. Eventually, SM outlets became a popular spatial data source for scientific research (see above in Table 1.1). SM data, on one hand, aligns well with traditional definitions of VGI. For instance, SM data (with its geographic attributes) is generated by non-expert users. Additionally, some social media outlets, such as Twitter, have an option to enable location data sharing. When such an option is enabled, users are volunteering their spatial data. For these reasons, SM may represent VGI-like data. However, unlike OSM data generated by users, SM users are unaware of their role in scientific research. For example, Twitter users unknowingly tweet about a variety of events which are then being captured (unbeknownst to Twitter users) for scientific studies in a variety of fields (see Table 1.1). This shift represents a divergence in defining VGI. The original VGI definition involved citizen scientists consciously producing spatial data for a collaborative scientific goal (OSM, Panaramio, iNaturalist). After the rise of SM, a new type of VGI definition was needed, one where volunteers contribute spatial data but lack a scientific direction and reason for contributing. This latter type of data is referred to as implicit VGI and traditional VGI approaches, such as OSM, as explicit VGI (Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2017).

Citizen scientists provide varying degrees of input into VGI data creation, similar to Arnstein's ladder of public participation in urban planning. Arnstein's ladder is composed of several different rungs that describe the degree to which the public engages in the planning process. For example, public members who have *delegated power* to implement planning decisions have an increased participatory role when compared to public members who only provide *input* into a planning decisions (Arnstein, 1969). More engaged public members, such as the former case, are placed at a higher rung in Arnstein's ladder. Comparable to Arnstein’s ladder, Haklay (2013) describes a four rung ladder for citizen scientists participating in the VGI process. Stage one is the most *basic* level of participation, where citizen scientists record the world around them, but provide no intellectual contribution to science. In this case, the citizen scientist acts solely as a sensor. Stage one is where SM data is located. Stage two, *distributed intelligence*, involves citizen scientists learning basic approaches to collecting and interpreting data. Often, the citizen scientist is tested on knowledge learned, as produced quality assurance measure. Stage three refers to *community science* where the citizen scientists define the problem and collect the data in collaboration with experts but, data analysis is left solely to experts. The fourth is *extreme citizen science*. In this stage, the citizen scientist is involved in all aspects of scientific discovery and production.

Stage one VGI data, in comparison to other stages, is the noisiest data type. Stage one citizen scientists are treated as if they are solely sensors of the world, cataloging their observations but for undirected and unknown scientific purposes. This is unsurprising that SM outlets are placed on this rung considering they were created with the intent to be an online platform for users to socially interact and not as data centers for scientific studies. Therefore, it is not surprising that SM data is littered with noise when it is used for scientific purposes since users are unaware of their inclusion in a scientific study. Below discusses how the noisy nature of SM data influences scientific studies.

Contextual Shortcomings: SM is unregulated and has little to no consequences for posting false messages. For instance, in the frantic hours and days following the April 15, 2013 bombing of the Boston Marathon, investigative units and news outlets sought suspect information and narrative stories from Twitter and Facebook. These outlets, however, were spreading misinformation about the Boston Marathon bombing (BMB) (Gupta & Kumaraguru, 2012; Starbird et al., 2014).Rumors on SM referenced two individuals as the potential suspects. These rumors were believed to be true until the FBI released the Tsarnaev brothers' names as the actual suspects. Though some newly opened Twitter accounts were suspended for spreading misinformation, no financial or otherwise imputing repercussions occurred. Furthermore, misinformation spread about Hurricane Sandy included photoshopped images, fake reports of the NYC stock exchange flooding, and sharks swimming the streets of NYC (Hill, 2012). News agencies began to report on one user purposely spreading false information, leading to the user issuing a public apology (Gross, 2012). Some of the aforementioned SM messages, such as sharks swimming the streets of NYC, are obviously false. Others are more difficult to identify, such as misidentifying the BMB suspect, requiring more intensive investigation.

Spatial Shortcomings: Location data provided by SM outlets is sparse (approximately 1%) and present quality and reliability issues. First, users may provide factitious location information, such as "Candyland" or "The Matrix". Second, users may select ambiguous location information. For example, a user sets their location to "London" which may reference London, Canada or London, United Kingdom. Third, location settings can be geographically over-generalized such as "Michigan" or "United States". Fourth, users may post SM content outside of their declared location. For instance, a user may have declared Detroit as their home city but are posting content in New York City as they visit for a weekend. Fifth, location data derived from 'check-ins' is spatially biased (Hasan, Zhan, & Ukkusuri, 2013). Check-ins refer to when a user discloses their location based on an immediate feature such as, a business, landmark, park, or other point of interest. More popular features will represent higher check-in rates and therefore give the impression of a densely populated area.

Temporal Shortcomings: The posting time of SM content also presents data quality issues. A user can post material at any time. For example, several research articles discuss how SM data can predict the occurrence of a disease before traditional surveillance methods (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011; Boyle et al., 2011; Santos & Matos, 2014). Digital disease detection research only estimates the likelihood of when a user experienced a disease or assumes that user has the disease at the time of posting. Contracting a disease, however, largely occurs in four sequential stages: incubation, initial onset of symptoms, peak symptoms, and dissipation of symptoms. Users may post their symptoms anytime during the latter three of these stages. Furthermore, it may be difficult to assign the SM post to any one stage based on context. For example, a user stating, "I want this flu to go away!" may be logically assigned to any influenza stage. Knowing exactly when the user experienced their initial or peak influenza symptoms would be useful knowledge in digital disease detection, but this information is not directly communicated through SM or a user's SM message. Therefore, in some cases, the time of a post cannot be definitely linked to a historical, current, or future tense.

Demographic Shortcomings: SM outlets may collect demographic data from its users for purposes outlined in the user term agreement policies. The data collected by the outlets, however, are not publicly available. Therefore, much of the data collected from SM, contains no demographic variables about the user.

Privacy Shortcomings: Privacy concerns have been documented with geospatial surveillance data and VGI. Crampton et al., (2013) notes that geo-surveillance techniques have potentially ushered in a society in constant fear of being watched by authoritarian groups, similar to prisoners being watched in Bentham's panopticon. Furthermore, there are scenarios where VGI participants might have good intentions to volunteer valuable and confidential information, such as an address, during a disaster (M. F. Goodchild, 2008). However, the public release of this data could result in the user experiencing future harassment. This problem becomes further complicated when VGI data is collected about personal or protected data such as, location information and medical wellbeing of users (Goranson, Thihalolipavan, & di Tada, 2013; J. F. Jones, Hook, Park, & Scott, 2011).

Dangers of Misinformation: The potential dangers of unverified SM data will vary according to the research topic. For instance, wrong Ebola treatments were being spread on SM (Oyeyemi, Gabarron, & Wynn, 2014), although no scientific study was conducted on mortality rates caused by this misinformation, it begs wondering. Misinformation also spread about the Zika virus (Venkatraman, Mukhija, Kumar, & Nagpal, 2016) which may have introduced unnecessary public panic, but this claim has yet to be studied. Other SM research assumes a degree of misinformation and strictly rely on SM's natural crowdsourcing structure to alleviate this problem (Gao, Barbier, Goolsby, & Zeng, 2011). Below discusses current solutions to the above problems.

Current Contextual Solution: Chicago Health officials in 2013 launched a privately developed application to track food poisoning episodes through tweets. Users expressing food poisoning symptoms in a tweet were sent a questionnaire asking them where they ate. The program was able to collect roughly 200 surveys indicating incidences of food poisoning at restaurants, 16%

of which subsequently failed a health inspection (“Twitter helps Chicago find sources of food poisoning | Reuters,” 2014).

Current Spatial Solution: Sparse and misrepresented location data can be augmented by regional dialect within a user's post (Chang, Lee, Eltaher, & Lee, 2012; Cheng, Caverlee, & Lee, 2010; Han, Cook, & Baldwin, 2013), the way users interact with other users, (Chandra, Khan, & Muhaya, 2011; Cheng et al., 2010) and, by the relationship types within a SM users’ social network (Jurgens, 2013). However, these approaches rely on indirect methods that can only predict the location of a user within 10km.

Current Temporal Solution: Current Twitter research is able to identify the tense of the Twitter post as either historic, current, or future. Adding tense introduced more accurate prediction of users experiencing influenza-like symptoms (Lamb, Paul, & Dredze, 2013). However, their research is unable to match the tense of a post to when the author actually experienced influenza symptoms.

Current Privacy Solution: Protecting confidential VGI can be difficult if not impossible. Currently, VGI does not have any agreed upon metadata standards including those for data protection. In some cases, VGI is considered public data despite that it may contain sensitive location information. Depending on the SM outlet, location data must be enabled for the spatial data of a user to be obtained. In such cases, the user is elected to protect or not protect their location information. Likewise, SM outlets allow users to select if their account is considered public or private. As this option implies, public accounts are viewable by any person and private accounts can only be viewed by the account's associated 'friends' or 'followers'.

Current Demographic Solution: Online web services gather individuals' demographics, date of birth, contact information, and email and other commonly used web-based data by scraping web content from an individual's digital footprint. Internet users will surf various websites and become members of a select few that, in essence, comprise their digital footprint. During this process a user may disclose various pieces of personal information such as their age, gender, or contact information to one or many websites that they frequent. Several online sites 'scrape' this information to create a database of web demographics. The data richness of this database is dependent on the size and detail of an individual's digital footprint. However, the data within the database may be incomplete or outdated. For instance, the contact information of a individual may change several times in a short period. This information, however, is not updated on any website; therefore, the scrapped data may not be reliable. Other sources, such as the Pew Research Center periodically surveys SM outlets to gain user insight into membership, usage, and demographics. Pew Research Center's profile snapshots of SM outlets leads to unsurprising demographic conclusions; SM users are young, college educated, affluent, and reside in non-rural areas ("Social Media Update," 2016). However, as their own research has reported, the demographic profile of these outlets is quickly changing, potentially causing their surveys to be quickly outdated.

Despite the current approaches to gaining value from SM data, new methods need to be introduced to augment the continued permissive nature of online data (Wilson & Graham, 2013). This research proposes a method for SM users to provide scholars with data to supplement current SM data fields and data that augments SM users' personal insight into a variety of themes. However, the scholars do not provide these users with scientific expertise. While most research focuses on using SM users as social sensors, this research introduces a method to increase the participatory nature of SM users by placing them between rung one and two of Haklay’s (2013) ladder of citizen science participation. This method is called the Digital Interaction Program (DIP). DIP is a flexible approach for digitally interacting with SM users to collect intellectual data from SM users. In turn, this data can supplement and augment SM data. First, this research provides an overview of the DIP concept including APIs and traditional SM data collection processes. Then this research provides an example of the DIP by identifying Twitter users experiencing influenza-like illness. Finally, this research discusses the utility, shortcomings, and future research of augmenting SM data.

Digital Interaction Program Concept

Application Program Interfaces

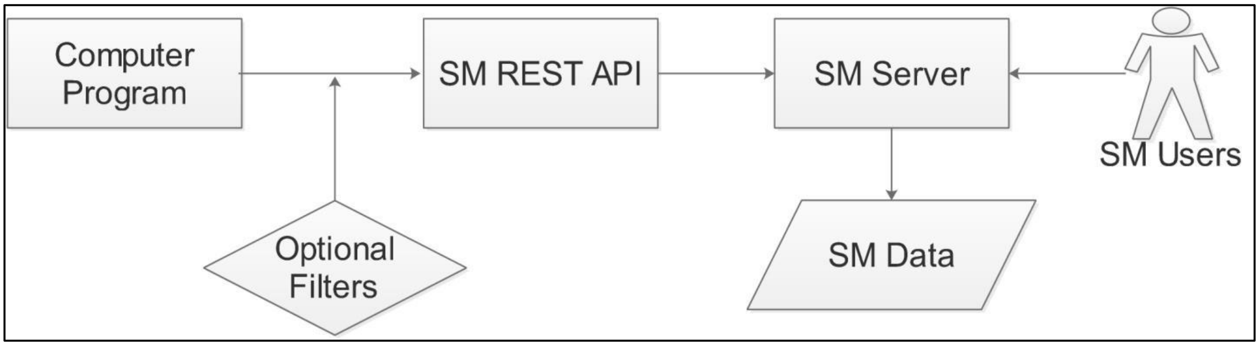
APIs are a set of protocols and tools that help people develop software applications that interact with computer servers. API functions and purposes are broad. In this research through, APIs are viewed as conduits for the transfer of data and messages between servers and personal computers. In this case, the server is designed to accept and respond to API requests from personal computers. The server then delivers data to or receives data from a personal computer. A program language (C++, Python, R, JavaScript, etc) is used to 'read’, ‘write’, or ‘edit’ data and messages across the API to the source computer server.

Broadly speaking, SM APIs are used towards the commercialization of products or for collecting data. Commercial APIs require payment but in return, purchasers are afforded more flexibility to interact with the SM's server. Most often this freedom is geared toward purchasers advertising their services and products in creative ways. For example, 1-800-Flowers.com provides a Facebook application called "Gimme Love" where users can send flowers to other Facebook users (Kaplan & Haenlein, 2010). Compared to commercial APIs, data capturing APIs are usually free of charge, but offer less flexibility to interact with the SM server. Both commercial and data capturing APIs restrict upload and download rates. API technical description, including bandwidth limitations are often documented online, such as Twitter(https://dev.twitter.com/overview/documentation).

Traditional SM Participation

SM data is traditionally collected retroactively or in real-time. Figure 1.1 describes real-time SM data collection. Here, SM users are posting content from their personal computers to the SM server concurrently as a computer program on a third-party computer receives SM data from a server via an API. In this traditional approach, the SM user is unaware that their data is being captured by a third party. This approach suggests that SM users are strictly sensors that gather information of the world around them and contribute that data for unknown scientific purposes (Haklay, 2013).

Figure 1.1: Technical Architecture of Real-Time Approach to Collecting Social Media Data



Digital Interaction Program (DIP)

Traditional approaches to SM research underutilize the intellect of SM users and potential of SM APIs. SM users have valuable insight into local anomalies, personal circumstances, and a host of other themes that may not be shared in an SM post. To gather SM users' valuable insight, this research suggests digitally interacting with them through APIs. APIs have three main functions: read, write, and edit data on a server but current SM research ignores API write and edit functions. Using the API to write data to the server presents an opportunity for researchers to digitally connect with SM users and then recruit them into their research process. DIP allows researchers to gain knowledge from SM users by following four steps outlined in Figure 1.2:

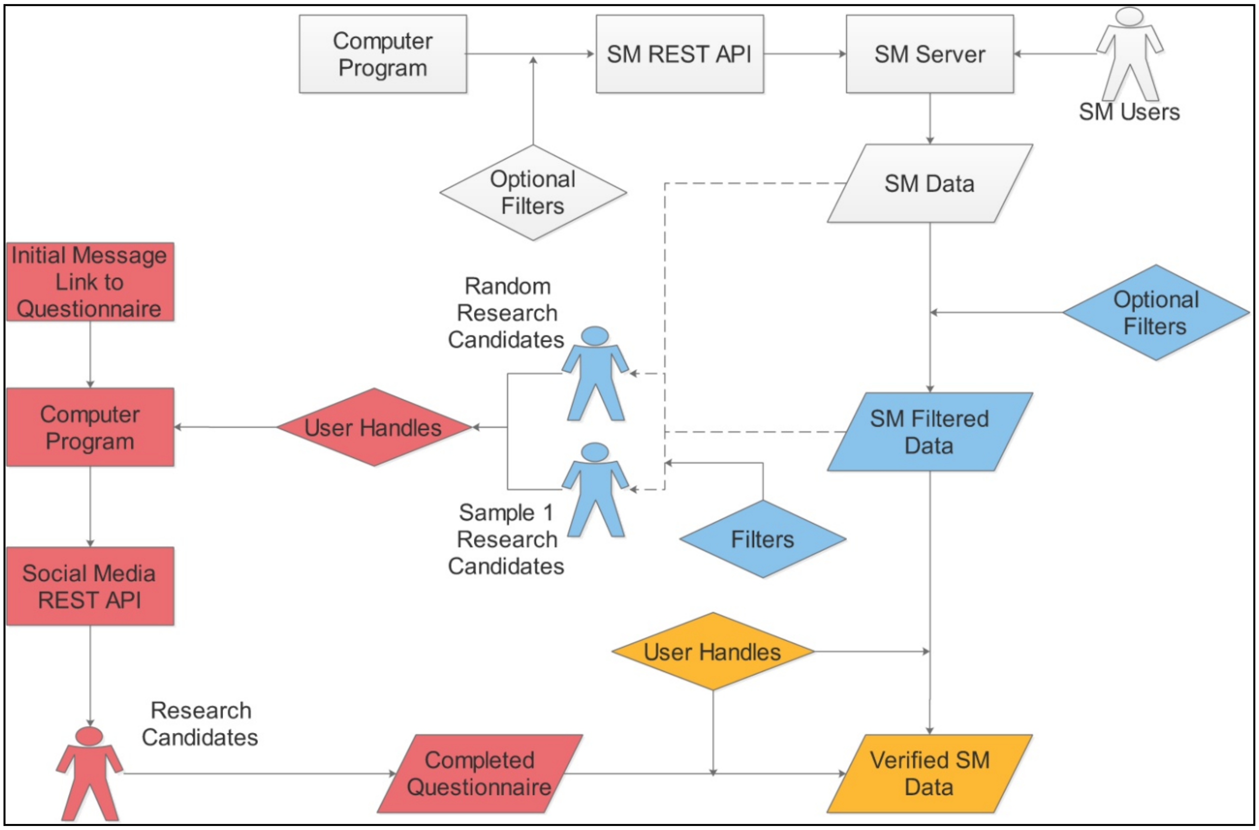
Step 1. Capture SM data (Grey).

Step 2. Select research participants (Blue).

Step 3. Digital Conversation and Questionnaire Design Data Merger (Red).

Step 4. Data Merger and Verification (Orange).

Figure 1.2: Overview of DIP's Technical Architecture



Step 1: Capture SM Data

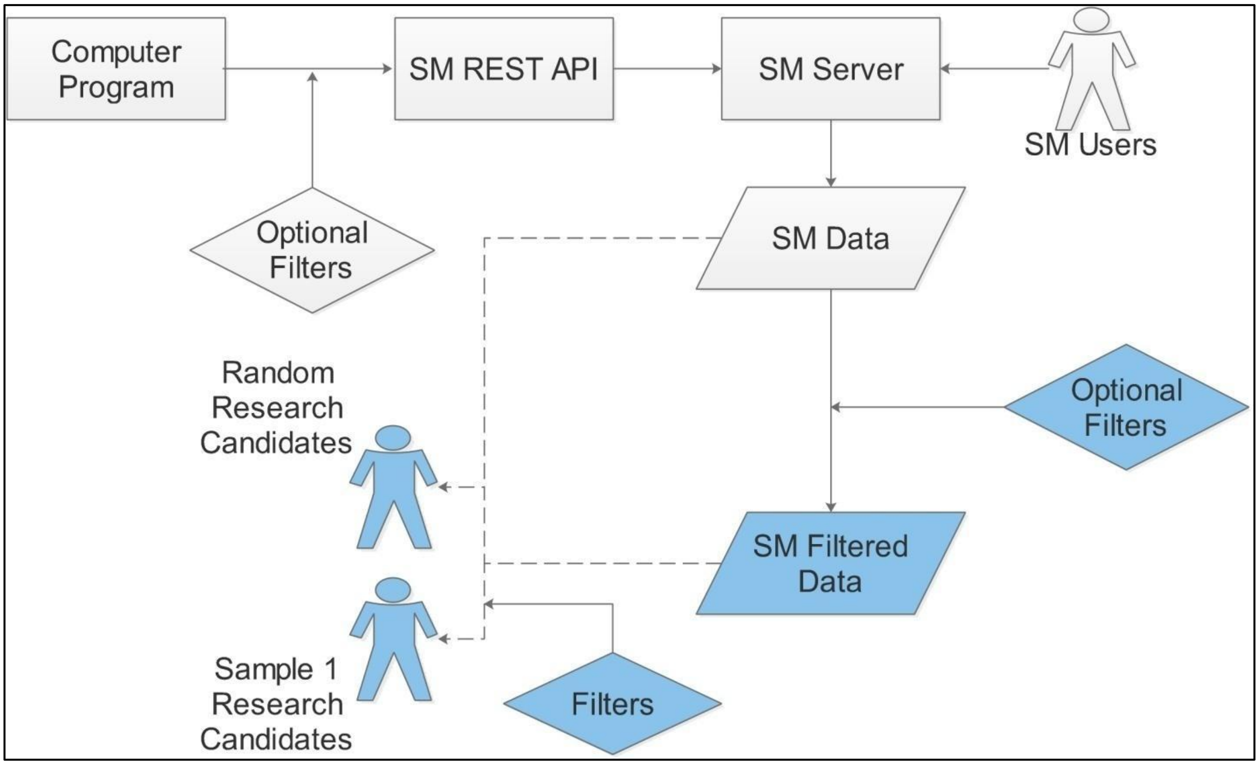
This step collects SM data in a manner that is identical to Figure 1.1. SM data is collected in real-time through a computer program accessing the SM server via an API. The computer program may contain a subroutine that filters data by location, keywords, user handles (screen names), or other criteria. Each SM API will offer different types and ranges of freedom when it comes to applying API filters. For example, Twitter's API permits developers to filter tweets by language, keywords, location, and user. Facebook's API, on the other hand, restricts filtering to topical trends, such as 'Detroit Pistons Basketball' or 'The Grammys''. As previously mentioned, SM outlets document their API protocols and filter limitations online.

Step 2: Select research participants

Figure 1.3, outlines the selection of research participants. After the SM data is collected, a computer program filters through the data to identify research participants fitting the research criteria. Basic filters include context (keywords, phrases, hashtags, external links, emojis), location (bounding box of an area, cities, states, countries, regions), and time of posting (before, during, or after a specific time, within a period of time, etc) can be applied as criterions for selecting research participants. More advanced filters may include searching for specific pictures, videos, and reply messages. Filter options are limited to the availability of data variables found on SM outlets. For example, some SM outlets do not contain location information, others limit the amount of text in a post. No SM outlet provides demographic variables of their users.

The above filtering options can identify research participants through a variety of sampling techniques. For example, researchers can use systematic sampling (every 10th user or every other user to post at a certain location), stratified sampling (100 random users from each of the five Buroughs of New York City), cluster sampling (all collected users from three random Buroughs in NYC), and random sampling. Clearly, the sampling method should align with the research's data needs.

Figure 1.3: Overview of Technical Architecture for Selecting Research Candidates

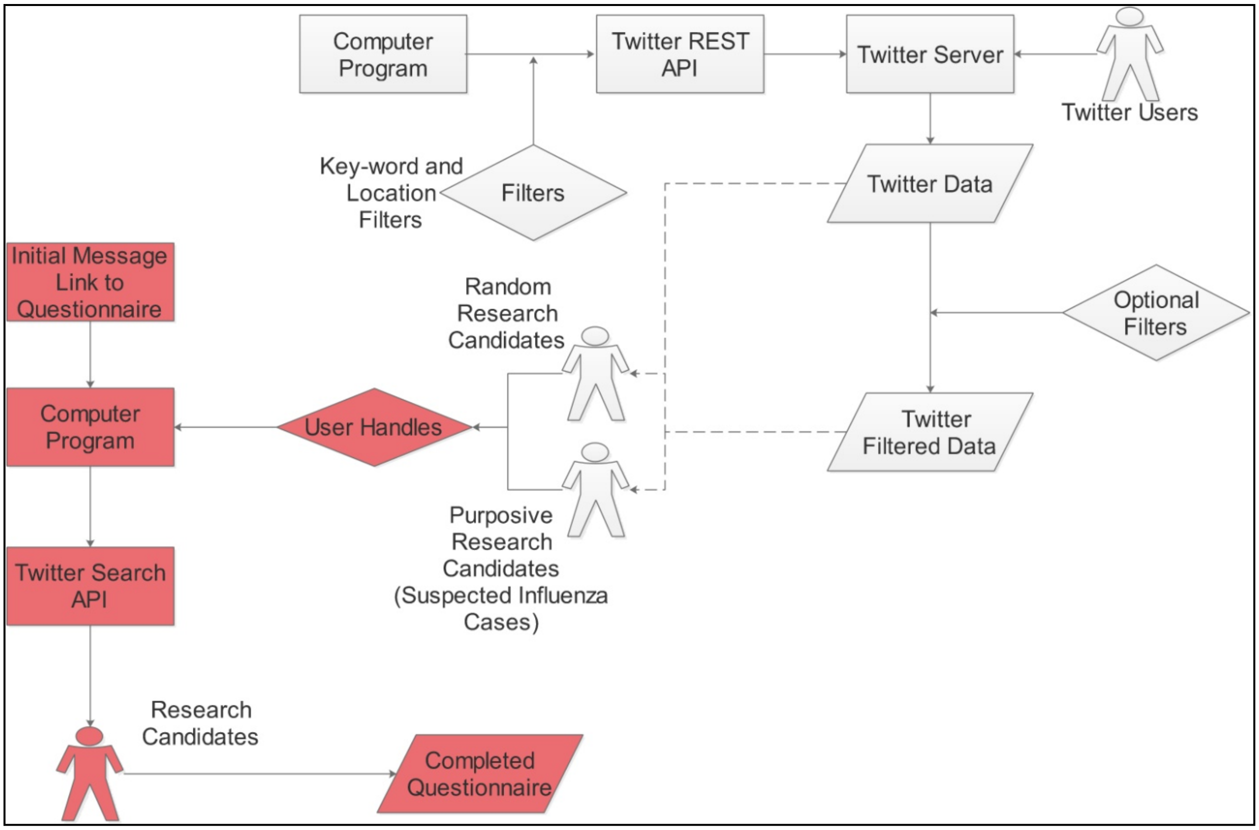


Step 3: Digital Conversation and Questionnaire Design

A digital conversation is initiated with the research participants identified in step 2, using the write function of an SM’s API. This process is outlined in Figure 1.4. First, user names (or handles) of the research participants are collected. Second, a computer program sends a short message to research participants via the API. The message contains an external link to an online questionnaire and a request for the participant to complete said questionnaire. After receiving the message, the research participant has one of four options: 1.) ignore the message's request, never complete the questionnaire, and never engage in a digital conversation 2.) read the message's request, complete the questionnaire, and never engage in a digital conversation 3.) read the message's request, never complete the questionnaire, and engage in a digital conversation 4).read the message's request, complete the questionnaire, and never engage in a digital conversation. The online questionnaire acts as a database to record SM users’ intellectual input.

The questionnaire can be designed to supplement and clarify research assumptions or well-known data inaccuracy issues such as, vague context, users' spatial whereabouts, and confirming when users post material. Additionally, it can be designed to augment SM data gaps, such as a user's demographic information. However, the SM user's handle (screen name, Twitter name, etc.) must be collected in the questionnaire since it is used as a common field to merge the questionnaire data to the user's SM data. At this stage of research, the questionnaire can be designed to provide SM users with scientific knowledge about why their data is collected and how their data is involved in scientific processes or theory.

Figure 1.4: Overview of Technical Architecture for Digital Conversation



Step 4: Data Merger and Verification

Step four merges the questionnaire dataset and SM dataset into a single dataset using the research participants' username. The merged dataset represents the research participants intellectual input (via the questionnaire) and their SM data.

Figure 1.5: Overview of Technical Architecture for Data Merger and Verification

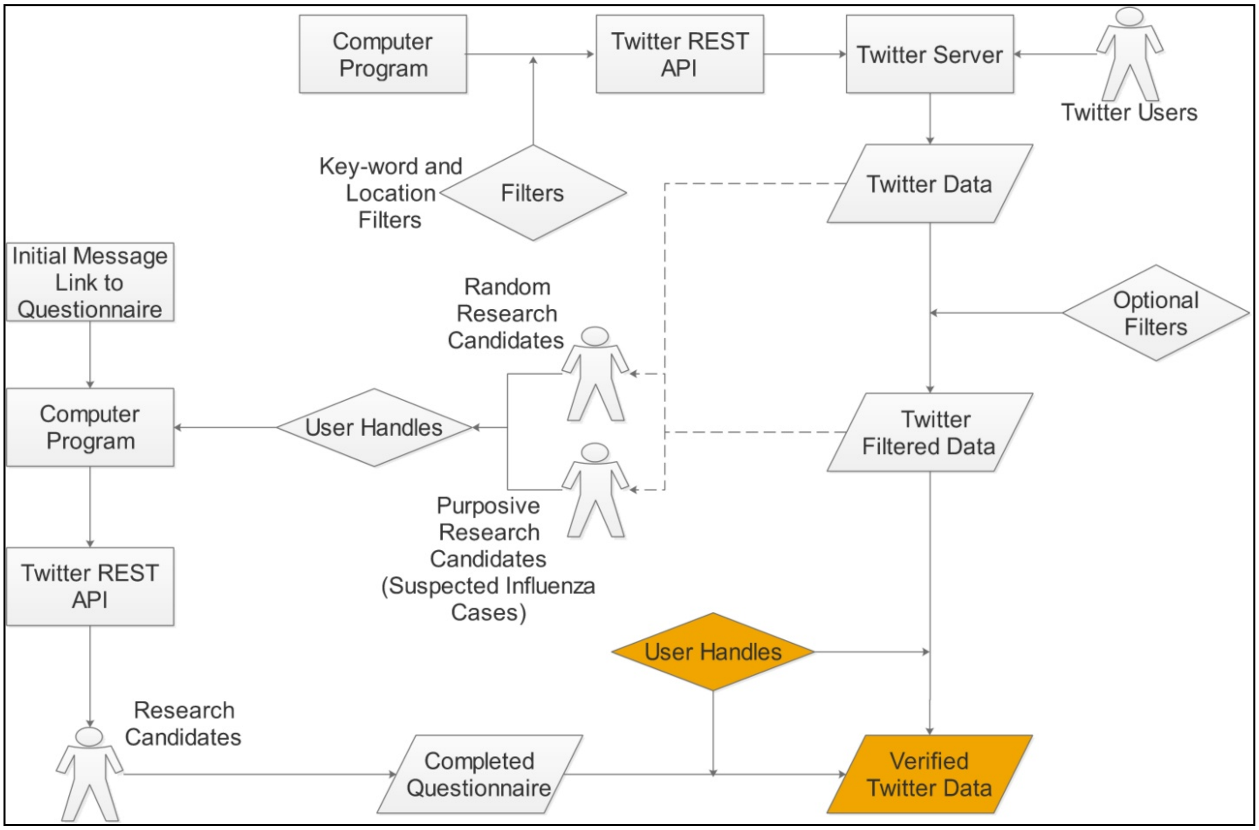
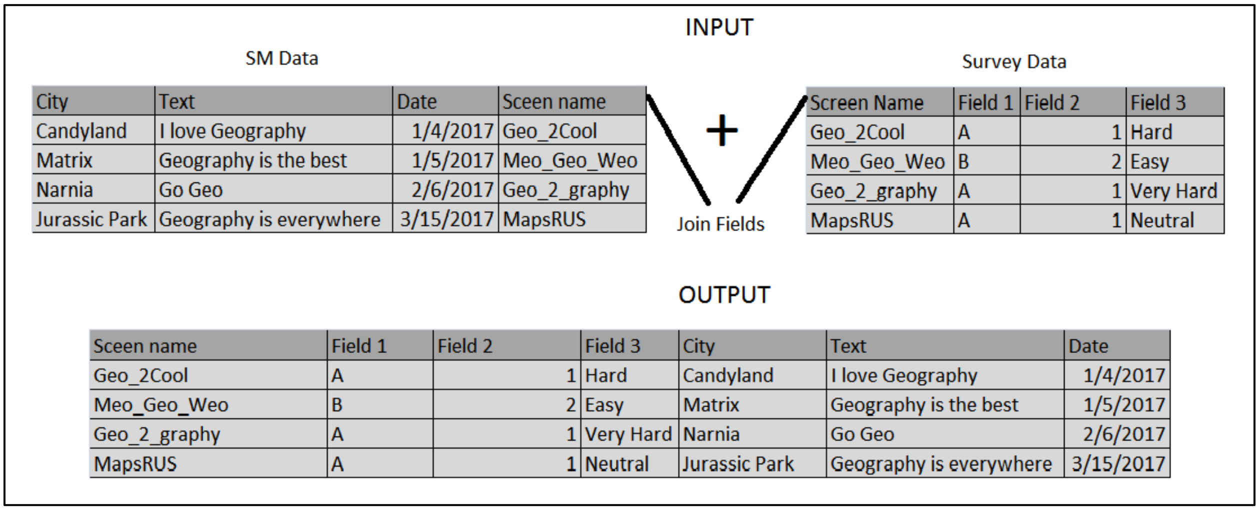


Figure 1.6: Example of Data Merge


INPUT
SM Data
Survey Data
+
Join Fields
OUTPUT

City	Text	Date	Sceen name
Candyland	I love Geography	1/4/2017	Geo_2Cool
Matrix	Geography is the best	1/5/2017	Meo_Geo_Weo
Narnia	Go Geo	2/6/2017	Geo_2_graphy
Jurassic Park	Geography is everywhere	3/15/2017	MapsRUS

Screen Name	Field 1	Field 2	Field 3
Geo_2Cool	A	1	Hard
Meo_Geo_Weo	B	2	Easy
Geo_2_graphy	A	1	Very Hard
MapsRUS	A	1	Neutral

Sceen name	Field 1	Field 2	Field 3	City	Text	Date
Geo_2Cool	A	1	Hard	Candyland	I love Geography	1/4/2017
Meo_Geo_Weo	B	2	Easy	Matrix	Geography is the best	1/5/2017
Geo_2_graphy	A	1	Very Hard	Narnia	Go Geo	2/6/2017
MapsRUS	A	1	Neutral	Jurassic Park	Geography is everywhere	3/15/2017

Application of DIP

An example of DIP's implementation is showcased below followed by a discussion of its benefits and weaknesses. In this example, DIP is applied to identify Twitter users likely experiencing influenza-like symptoms in NYC and London. Influenza is most prominent during fall and winter months from the beginning of October to the end of March with symptoms including fever, cough, sore throat, headache, vomiting, and diarrhea (CDC, 2016).

Step 1: Data SM Capture

Step one the DIP process involves capturing SM data. This research downloaded a 1% sample of Twitter activity using Twitter's API from November 26, 2014 to March 17, 2015 with keyword filters applied within the bounding-box location of London and NYC. Keywords included 'flu', 'cough', 'sore throat', and 'headache'. These words show a 95% correlation with national influenza statistics (A Culotta, 2013; Aron Culotta, 2010). This search resulted in roughly, 14 million tweets each from NYC and London. However, this sample may not be random or representative of the population (Morstatter, Pfeffer, Liu, & Carley, 2013).

Step 2: Select Research Candidates

Step two of the DIP process involves selecting research participants conducive for the study. This research selected both random and purposive research participants. The inclusion of random candidates rests on the possibility that users may experience influenza-like symptoms, but do not express their symptoms through a tweet. Purposive candidates include those only located within NYC or London and mentioned keywords related to influenza-like symptoms (A Culotta, 2013; Aron Culotta, 2010). Twitter's API loosely applies filters to its data searches in Step one. Therefore, location and keyword filters are again applied in Step two. A total of 8,696 purposive and 4,906 random research candidates were matched to the above criteria for NYC and 7,756 purposive and 2,792 random research candidates for London.

Step 3: Digital Conversation and Questionnaire Design

Step three of the DIP process is designed to engage users in a digital conversation and then recruit them into participating in an online questionnaire. Research participants identified in Step two received a tweet containing an external link that directs users to an online close-ended questionnaire. This initial tweet is written as follows: "@*username* My name is Josh plz help researchers at MSU save lives from the flu by taking this short survey *survey link*". This tweet was sent through Twitter's API write function and acts as a digital ice-breaker between the research participants and the researcher. It sparked polite, rude, sarcastic, and host of other responses from the research participants. Table 1.3 shows some of the responses. Many users were eager to provide additional input to the scholar including: homeopathic remedies to the flu, a link to the user's blog about the flu, and how the flu has affected them personally. Research participants received informative and polite replies from the researcher. Replying to research participants messages likely decreased their perception of the scholar’s intentions as being fake or fraudulent.

Due to limitation on how many tweets can be sent over Twitter's API, research participants were not sent reminders to complete the questionnaire. Once the questionnaire was completed, no additional tweets were sent to that research participant asking them to complete the questionnaire. Research participants only received additional tweets to complete the questionnaire if they continued to tweet influenza-keywords.

Table 1.2: Examples of Tweet Responses

- @MSUFluResearch I filled out the survey but I don't live in New York so my zip code was not on the list. Its 06010 for CT. - do you mind if I don't . Sorry - no - Will do the survey ..in alittle later Thank You - I'll do it !! - We've done your survey. But others might like to... - my name is Jeff - can't do it	- ok no problem then 👌 - Again with the survey? I already told you I CAN'T READ!!! - completed best of luck. - i am the flu - Done :) - no I'm ok mate - F*** off. - Is water wet? - Plz go f*** urself thanks pal bye bye

Of the 13,602 NYC research participants, 275 completed the questionnaire (2%) and of the 10,548 London research participants, 295 completed the questionnaire (3%). The questionnaire was designed to determine if a user experienced influenza, identify which symptoms were experienced, when those symptoms peaked, the user’s home ZIP Code, and basic demographic data. It also asks for the user's handle name. The questionnaire was powered by Google Forums and can be viewed in detail at http://goo.gl/ko3qvx and http://goo.gl/NPemrp for NYC and London, respectively.

Step 4: Data Merger and Verification

This research linked the Twitter dataset to the online questionnaire through the common field of username. This linked dataset represents the augmented social media dataset.

Discussion

The objective of this research is to highlight the significance of DIP as a flexible approach to supplement and augment SM data by gaining additional insight and knowledge from SM users. This process represents a shift in participatory level of SM citizen scientists producing VGI. Haklay described four rungs on a ladder in which each rung ranks the participation level for citizen scientists creating VGI. Rung one, the least participatory, involves citizen scientists solely acting as sensors that gather information about the world around them with an unbeknownst role in the scientific process. This has been the traditional approach of using SM data in research. This research proposes an alternative approach called the Digital Interaction Program (DIP). DIP invites SM users to digitally engage scholars and complete an online questionnaire. As shown above, the digital interaction between users and the scholar ranges from humble discussions about the research topic to frightening insults. This type of interaction suggests that the citizen scientist participation ladder described by Haklay (2013) needs an additional rung situated between rung one and two. Here, citizen scientists are not viewed solely as sensors, but instead provide intellectual contributions through digitally conversing with scholars and through their answers on the online questionnaire. These intellectual contributions help provide context to a user's SM data and the any research questions. The DIP process does not provide VGI contributors with knowledge on how to collect, record, analyze, or disseminate data, as suggested by Haklay’s step two. Therefore, the DIP process suggests that an additional rung needs to be placed between the first and second rung in Haklay's ladder.

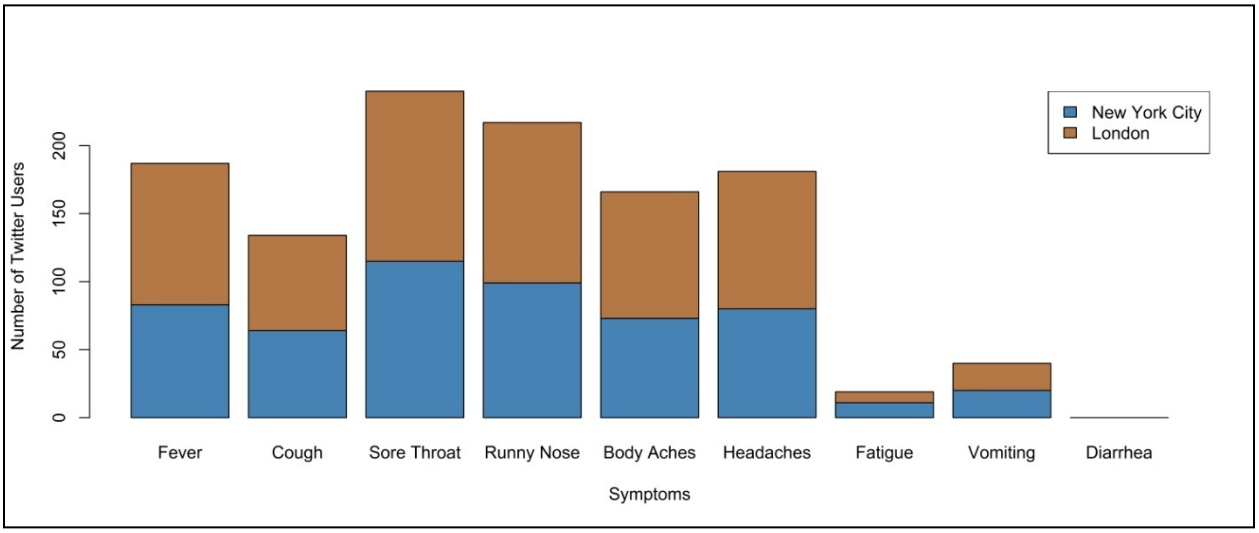
Benefits

Supplementing and Augmenting SM Data

SM data contains contextual, spatial, temporal, and demographic shortcomings (Elwood, 2008; M. F. Goodchild & Li, 2012; Sui et al., 2012; Tulloch, 2014). The DIP system helps identify and correct these shortcomings. DIP collects intellectual input from SM users through an online questionnaire. The questions should be designed to clarify the context and spatial-temporal content the SM user's post. However, the questionnaire can ask participants any question including those relating to demographics, socio-economic status, and other data and topics not available through SM websites. DIP can also be executed at a variety of spatial scales but only if the SM API allows data collection at different spatial scales. For instance, Twitter's API allows searches to occur at a variety of spatial scales, some including global, national, state, county, and local. This research applied the DIP process to the metropolitan areas of New York City and London. However, some SM APIs do not have filter options for selecting geographic areas.

As an example, this research used the DIP system to identify Twitter users experiencing influenza-like symptoms in NYC and London. Figure 1.7 shows the frequency of influenza-like symptoms experienced by research participants in NYC and London. Many of the research participants experienced multiple symptoms. In most cases, Twitter users self-diagnosed themselves without describing their symptoms but instead focused on stating they ‘had the flu’. The additional information presented in Figure 1.7 would likely not be possible to identify unless interaction with SM users occurred.

Figure 1.7: Variety and Distribution of Influenza Symptoms Experienced by New York City and London Twitter Users



The example provided above also demonstrates users' willingness to engage with the scholar. As previously mentioned, Twitter users were excited to share approaches to lessen the likelihood of becoming infected with influenza, home remedies to decrease recovery time from infection, and links to other influenza documentation. It was not this research's goal to examine the credibility of their ideas but the nature of people sharing ideas through DIP. At this time, it cannot be determined if this is a common side effect of DIP or not.

Ranging Application

DIP is an approach that can be used for a wide range of disciplines including researchers, demographic surveyors, urban planners, disaster managers, disease outbreak specialists and other fields that need to interact with SM users. For example, urban planners might find DIP useful because it is an approach that can be used to gain input from SM users discussing the urban environment. For instance, urban residence might be expressing their frustration with parking opportunities in a city through SM. DIP could be implemented so that context is added in order to identify the specific reasons why SM users are frustrated with parking in the city. Is it too expensive? Is there not enough parking? Are parking garages (structures) located in unsafe areas? DIP provides the framework to probe SM users so that these questions can be answered.

DIP could also be an alternative approach for disaster managers to identify the circumstances of potential disaster victims in a secure and centralized manner. During a disaster, disaster managers try to ascertain as much information about a disaster victim as possible; medical needs, location, accessibility, and other important pieces of information. Traditionally, SM provides a platform for disaster victims to share this information through a public post, tweet, or other message but users may reluctantly do so considering it can be viewed by anyone. Therefore, the privacy of sensitive information during a disaster crisis is important. DIP can collect and store sensitive data such as, address, medical needs, or other personal information on a private database. This may lead to more disaster victims sharing their sensitive data. Before and during a disaster, disaster managers seek to efficiently share the necessary data to mitigate negative effects of disaster. DIP stores SM users' intellectual input in a centralized and easily accessible database. Therefore, disaster managers can quickly access and search the database for relevant data, rather than waiting for the data to be shared.

DIP can also be a method applied by disease outbreak specialists mapping the near real-time spread of disease outbreaks. Disease outbreak specialists seek data sources that describe the when and where of a disease outbreak. Most current disease outbreak surveillance systems are not in real-time and offer a coarse spatial resolution either at the city or state scale. However, a more informative approach would be to describe the disease outbreak in near or real-time and with specific whereabouts of disease spread. The DIP approach can be applied in near-real time with survey questions that center on gaining clarification about users' current state of health.

Similar to disaster managers' data privacy and storage predicament, disease outbreak specialists need to securely store and quickly access sensitive data. As previously mentioned, DIP provides these needed services.

Privacy of Data

Privacy issues with VGI have been well documented, as described above. The digital conversations initiated by DIP will be public if the engagement occurs through direct messages, such as tweets. The conversations will be considered private if the engagement occurs through a private message. However, the survey where SM users provide their intellectual input maybe stored privately or openly available to the public. Obviously, the level to which information from the questionnaire is shared will depend on the research at hand and the data content. The example provided in this paper privately stored Twitter users' responses that can only be accessed by the author of this paper. Privately storing questionnaire responses is especially important for the collection and storage of sensitive data, such as a user's influenza health status.

Even though the questionnaire is stored privately does not mean it is immune to hacking efforts. Hackers could access the data by hacking into the services that provide them. Depending on the sensitivity of the data, any data breaches could lead to the publicly leaked data. This obviously presents ethical issues that future research can address.

Limitations

Questionnaire Limitations

The self-completed questionnaire used in the DIP process presents limitations. First, users completing the questionnaire may report false information. Furthermore, it is difficult to impossible to ensure that any reported information is in fact true. Second, the length of the survey, like many survey methods, may be a topic of concern. In the above example, some Twitter users described the length of the survey as too long, others thought it was too short (See Kaplowitz, Hadlock, & Levine, 2004; K. B. Wright, 2005 for more information for online participation rates.) Lastly, a questionnaire asking SM users to disclose personal data may decrease participation rates or may lead the participant to skip that question. Alternatively, though, SM users may be more willing to provide personal information during a disaster as a good Samaritan act.

Surveys often contain response bias since survey propensity increases from users who have interest in the survey variables. This phenomenon is present with or without incentives. The online survey in DIP is also likely to contain a response bias since users completing the survey are likely to have some vested interest in the survey variables. Depending on the implementation of DIP, this may or may not be an issue (Groves, 2006). For instance, using DIP to identify parking issues in a large city is likely to have a survey propensity composed of people who excessively dislike or favor city parking. On the other hand, DIP might be employed to gather information on what teachers think of the education system. Since the target audience is teachers it is less likely to have a response bias.

API Bandwidth Limitations

The application of DIP is dependent on the limits of a SM's API write and read functions. The write function of a SM’s API allows personal messages or general posts to be sent to users. Often these limitations are based on a fixed rate, total number, or a combination of both. For instance, an API limits writing functions to 20 messages per hour, 400 messages per week, or a total of 1,000,000 message. Exceeding these limits stops the sending of messages.SM outlets may not document their API bandwidth limits for the write functions. While this is unfortunate, it helps stop SPAM messaging on SM. If the limits were published, a program (script) could be developed to throttle the rate of messages being sent through the API without exceeding the bandwidth limits. Therefore, not knowing the bandwidth limits makes automating DIP difficult, because the rate of messaging research participants can quickly outpace the API write function limits. In comparison to the write function, the read function of SM APIs are well documented but still have rate limits.

API updates

SM platforms periodically change their API's functions and permissions, sometimes without warning. Depending on the social media platform, the API functions and protocols may change frequently and/or drastically. A minor change, for instance, would involve the SM organization adding a new filter function to the API. On the opposite end of the spectrum, the SM outlet could change their opensource APIs to become proprietary therefore eliminating the availability of free data. Depending on the API changes, the DIP process might need to be altered or in some cases may become futile.

Digital Dialogue Challenges

Ideally, research participants complete the questionnaire after receiving the initial message from the researcher. However, this is usually not the case. Users often identified the initial message as automated SPAM, sent by a ‘robot’ SM account. Their thinking is reinforced by two notions. First, users receive a generic message from an unknown account, which immediately gives the impression of SPAM or 'junk' mail. Second, that initial message asks research participants to click on an unknown link that directs them to an unknown external website, which may increase their suspicion. It appeared that research participants were more likely to complete the questionnaire when their suspicion is eased. Digitally conversing, on a one-on-one basis, with concerned research participants helps build trust between them and the researcher. Questions presented by the research participant should not be answered in an automated way.

Message viewing

Some SM outlets may strictly structure their API's write function to send messages solely under ‘private’ viewership. For instance, Twitter’s API allows messages to be sent to a private inbox or through a public tweet. Facebook’s API, however, only allows messages to be sent to private inboxes. As the names suggests, the messages sent to a user’s private inbox are not considered public and therefore can only be viewed by the user. Private messaging through a SM’s API warns the recipient that the incoming message is outside of their network, which requires the user to 'accept' the message before it can be viewed. This acceptance process adds another hurdle to users viewing the message, potentially decreasing the chances of their participation in the online questionnaire.

Conclusion

The abundance of SM data coupled with the advent of SM APIs brings new opportunities to digitally interact with diverse SM users for a variety of research themes. As a response, the past decade has seen a prolific increase in research and applications centered on social media data (Kaplan & Haenlein, 2010). However, current research views SM APIs as a one-way street, whose only purpose is to gather data. Instead, social media APIs can digitally connect, converse, and recruit SM users to complete an online questionnaire. The questionnaire is used as a platform for users to provide their intellectual insight to supplement and augment SM data. When users complete the questionnaire they are participating as more than just sensors, suggesting they are not situated on the first rung of Haklay’s ladder for citizen science participation. However, DIP does not situate SM users on Haklay's second rung because SM users do not directly contribute to the scientific process of the research. Rather, applying DIP to SM data situates SM users between rung one and two. In the example provided, Twitter users provided their intellectual input as users self-diagnosed themselves with influenza-like symptoms. Even though many of the users are not medical doctors, their intellectual input helps provide context to their current influenza situation. Without DIP, this added context would be absent.

Future research needs to focus on applying DIP to crisis management and other disease surveillances. The DIP process is intended to capture intellectual data from SM users in addition to filling in missing SM data gaps. However, during a crisis event, the chances for DIP to capture misinformation may increase since survivors might overstate their current situation, such as exaggerating medical needs.

As the media covers a particular event, SM media discussions are likely to increase, potentially causing an increase in the spread of misinformation. In cases such as this, DIP may become a process that helps filter out misinformation. On the other hand, DIP may become a repository for misinformation as participants purposely and incorrectly answer questions on the online questionnaire. The DIP process needs to be tested on a subject that has gained large media attention.

The questionnaire response rate was about 2-3%, which is low compared to traditional questionnaire response rates. Given that users receive SPAM-like messages in the DIP process, it is not surprising that the questionnaire response rate was low. Research needs to focus on methods to increase the questionnaire response rate.

TRANSITION

The previous chapter discussed an approach, called the Digital Interaction Program (DIP). DIP is designed to augment and/or supplement social media data by increasing the participatory role of SM users. Traditionally, social media users are viewed solely as social data 'sensors'; posting pieces of information of the world, their daily activities, and other events. This is evident in the numerous studies conducted about the role various SM outlets play in predicting certain events and its ability to create situational awareness about disaster scenarios. However, there are contextual shortcomings of SM acting as predictive and awareness data sources. Furthermore, most research utilizing SM has not fully understand ways to include SM users into the scientific process. DIP was designed so that SM users are not just unbeknownst data sensors for scientists but instead transform into a data sensor that has a more involved role in Science. Their contributions rest on providing additional data or intellectual input to scientists by completing an online questionnaire. This questionnaire seeks SM users' formal input about a scientific topic which helps scholars augment current and supplement any missing SM data. This dissertation uses Haklay’s (2013) ladder of citizen participation and Groves (2006 and 2011) survey methods as a framework to build the aforementioned theories.

The next chapter uses DIP to spatially and temporally augment influenza related tweets. Chapter two compares a user's space-time distributions of their influenza tweets against their non-influenza tweets, correcting two shortcomings. First, current research assumes the time an influenza tweet is posted indicates when the author experienced symptom onset. However, this is not always the case, as a user may tweet about their influenza symptoms days after initial symptoms. Second, digitally detecting influenza cases on Twitter is applied to a spatially broad national, state, and metropolitan levels. This Chapter identifies the location of influenza tweets in relationship to the user’s home ZIP Code. Chapter two produces two findings. First, Chapter

Two suggests that research using Twitter data for space-time paths research will be dependent upon the health status of the user. This is an important aspect to consider as space-time path related datasets are growing in variety and velocity given the advent of Web 2.0 and the popularity of cellular devices. For example, using cellular location data to identify popular travel networks in a city might have very different results if a travel-heavy neighborhood is bedridden with sickness versus when that neighborhood is healthy. Second, Chapter Two suggests that influenza tweets might be a viable data source to help predict local influenza ER admissions since influenza tweets tend to occur at users' home ZIP or Postal Code. Moving forward, space-time paths research needs to examine how people interact with different urban design spaces as it relates to a variety of different health issues. Do such research could have important implications on disease outbreak mitigation strategies such as, identifying corridors of disease propagation and areas likely to experience high concentrations of sick individuals.

CHAPTER 2
THE GEOGRAPHY OF INFLUENZA TWEETS

Abstract

Research on the use of Twitter to detect influenza outbreaks has generally assumed an influenza tweet is posted when initial symptoms develop. Furthermore, such research has generally been applied on spatially broad scales (nation, regional, or metropolitan) since the fine-scale spatial context of influenza tweets is unknown. This chapter sheds light on these spatial and temporal assumptions. This research uses Vertalka's (2017) Digital Interaction Program as a tool to digitally ask New York City and London influenza-ridden Twitter users when they experienced peak influenza symptoms and their ZIP or Postal Code of residency during the 2014-15 influenza season. Users expressing influenza symptoms tend to tweet about their symptoms during periods of heightened symptoms (days after symptom onset), suggesting the posting of influenza tweets succeed the occurrence of actual infection infections. Furthermore, tweets expressing influenza symptoms tend to originate within users' home ZIP or Postal Codes, implying influenza detection through Twitter can occur at finer geographic scales than current research indicates.

Introduction and Background

People’s daily activities occur in space and time. A straightforward example involves common activities such as work, school, social events, and rest that people experience during any given week (Hanson & Huff, 1988). All of these activities occur in a given space during a period of time and for the most part are fairly predictable. For example, many of us work in a location that is separate from where we sleep. Likewise, many of these activities do not overlap; we cannot sleep and work simultaneously. Many of us also know not to call anyone after a certain hour of the night, as that person is likely to be asleep and it would otherwise be rude to wake them. These activities are not consistent though as peoples’ schedules change. For instance, someone may be too sick to work during their scheduled work time, thus they are not in their work location during their work time. Instead, during their work time they are likely in their home location resting or recovering.

This research starts by discussing space-time paths then discusses Twitter's role as a data source in understanding people's space-time path in an urban setting. Twitter's role in detecting influenza cases is also discussed followed by stating two shortcomings of digitally detecting influenza cases. To correct these shortcomings, this research employees the Digital Interaction Program (DIP) which is briefly discussed below. Finally, the findings of this research are stated and its implication on space-time paths and digitally detecting influenza cases.

The when and where people's activities have been referred to as the space-time trajectory or paths of individuals. Space-time paths are, geometrically speaking, composed of a 2D planar coordinate space with a temporal dimension as a third dimension. Points are established within the spatial field and the time is recorded when an activity occurred at any particular point (Hägerstraand, 1970). From this, a timeline can be built that categorizes where a person was and what activity occurred. In turn this timeline can characterize transportation networks, parks, shopping districts, and other urban features. Individual’s space-time path behavior will vary from day-to-day and week-to-week and consequently their spatial footprint will reflect that change. These changes maybe subtle, such as stopping at a new coffee shop on the way to work or more dramatic such as, moving to a new city. Therefore, long term studies into the space-time path of individuals is preferred as it provides a larger mosaic of an individual’s behavior and finer detail of events along numerous space-time points. However, individuals need to provide enough data points throughout a long enough time period for high spatial-temporal detail of the urban environment to be reached. When there are too few space-time data points of an individual, it becomes difficult to determine if a user is behaving in a normal capacity and consequently has little interaction with the built environment (P. Jones & Clarke, 1988) or abnormally for whatever reason (Bayarma, Kitamura, & Susilo, 2007). Likewise, the more individuals included in space-time path studies, the easier large-scale patterns can be identified.

Traditionally, this type of research relied on recruiting individuals to keep detailed diaries of where and when they partook in daily activities (Janelle, Goodchild, & Klinkenberg, 1988). This approach was expensive and only provided a small glimpse into the space-time path of individuals. The advent of space-time data embedded into social media outlets presented a shift in the amount and variety of space-time paths data. The last decade has seen a rise of real-time web-based data streams such as, social media outlets and search engine queries, that are acting as an augmented and implicit data sources for traditional and authoritarian data sources (M. Goodchild, 2009; Sui et al., 2012). Twitter, a micro-blogging website that limits users to 140-character messages or tweets, has been a dominant data source for this type of research as it relates to urban form and function. Urban form references the city’s physical shape and size of buildings, streets, and other features. Urban function refers to activities that occur at but within different spaces (Crooks et al., 2015). From these descriptions, urban form and function are interrelated. For instance, Birkin & Malleson (2012) discuss the space-time path of Twitter users as a means to identify complex space-time path behavior of individuals. Their research focuses on when and where tweets are posted as it relates to land use patterns and travel behavior. They had several important findings. First, they identified that users tend to travel an average of 2.5 km on any given day. Second, those that tweet most often do not tweet in residential areas. Third, the context of tweets does not always match their spatial location. A glaring drawback of their research rests on the fact that their data contained no ground truth. Therefore, the researchers were forced to make assumptions about where users reside. Cooks et al., (2015) identify that Twitter can be used as a data source to monitor public activities across different spatial and temporal scales. They make note of tweets at voicing the diurnal nature of activities where users tweet about work during the day and night-life activities at night. Their research suggests that Twitter can be used as a proxy to help identify the form and function or urban areas. Furthermore, Golder & Macy (2011) identified changing moods of happiness as a result of diurnal and seasonal changes.

Twitter has also emerged as one of the dominant digital data sources for scholars to detect influenza outbreaks. Scholars use the content in tweets to determine if users are experiencing influenza-like illness (fever of greater than 100°F, cough, sore throat, muscle aches, headaches, runny or stuffy nose, fatigue). The tweet's embedded timestamp and geographic information (user enabled location data or GPS coordinates) provide data on when and where a user is experiencing influenza-like illness. Ultimately, these embedded data fields allow scholars to use Twitter to predict the occurrence of influenza cases in certain locations. For instance, Broniatowski, Paul, & Dredze (2013) demonstrate that influenza related tweets can predict national and local influenza rates two weeks before traditional surveillance. Paul, Dredze, & Broniatowski (2014) found a high correlation between Twitter users experiencing influenza-like illness and government influenza data for 10 English speaking countries. Achrekar et al (2011) found influenza tweets to be highly correlated with national influenza rates for age groups of 5-24 and 25-49. They note, however, that this correlation likely reflects the demographic age range of Twitter users. Research further distinguished self-reported influenza cases on Twitter from users that are merely discussing the presence of influenza or describing an influenza infected person other than themselves by using machine learning algorithms (Lamb et al., 2013). Culotta

(2010) identified an ensemble of Twitter keywords that predict influenza rates while minimizing false-positives.

All the aforementioned research does not account for the space-time path of influenza tweets as it relates to where and when users tweet about their flu symptoms. This is clearly important as Twitter users may post about their influenza symptoms from anywhere at anytime. Lamb et al., (2013) partially help overcome this issue by classifying influenza tweets based on whether the author is referencing their influenza symptoms in a present or past tense. For instance, the first and second sentences describes a user's past and current experiences with influenza, respectively.

1.) That was a terrible flu! Those symptoms were awful!

2.) I am sick with the flu! These symptoms are awful!

Their approach rests on the idea that adding temporal context to influenza tweets results in more accurate digital disease detection. While their approach is straightforward, it does not entirely correct the research assumption that Twitter users post their influenza symptoms at symptom onset. It is clear that tweet one is written in past tense but it is unclear of what historic time period the author's influenza tweet is referencing (yesterday?, a week ago?, etc.). Furthermore, users can discuss their current influenza symptoms using temporally vague language. For example, a user tweeting that they "Never want to have the flu again" could be referencing their current or past influenza symptoms. Therefore, it is difficult to accurately determine when a Twitter user may have experienced influenza symptoms based solely on the tweet's content.

Not only have scholars yet to characterize the temporal timing of when Twitter users tweet about their influenza symptoms, but they have yet to identify where influenza ridden users tend to tweet. Digitally detecting influenza cases on Twitter has occurred on broad spatial scales including national, state, and occasionally city levels. Research has not examined finer spatial scales of where influenza tweets tend to occur along a user’s space-time path. It is important to quantify this because digitally detecting influenza has yet to derive the spatial context of tweets relating to influenza-like illness.

Therefore, this research seeks to answer two questions:

> Question 1: Where do Twitter users prefer to tweet about their influenza symptoms in comparison to their home ZIP or Postal Code?
>
> Question 2: At what stage of influenza infection do influenza-ridden Twitter users tweet about their influenza symptoms?

This research seeks to answer these questions by introducing findings about where and when users tweet about their influenza symptoms compared to when and where those same users post non-influenza tweets. This research uses the Digital Interactive Program (DIP) to digitally engage and then encourage suspected influenza-ridden Twitter users to complete an online questionnaire (Vertalka, 2017). This questionnaire probes users about their spatial-temporal tweeting behavior. As a comparative, this research tests the above questions by engaging Twitter users from two global English speaking cities, New York City and London. Both of these cities offer a rich collection of geolocated tweets (City Metric, 2014) making them ideal in this type of study. This research starts by briefly describing the spaces in which Twitter usage occurs. Next, Vertalka's (2017) DIP is used to capture the spatial and temporal tweeting behavior of Twitter users experiencing influenza-like illness. Then this research analyzes the results from DIP to identify when and where influenza tweets commonly occur. Finally, theories are built about the spatial-temporal tweeting pattern of influenza-ridden Twitter users.

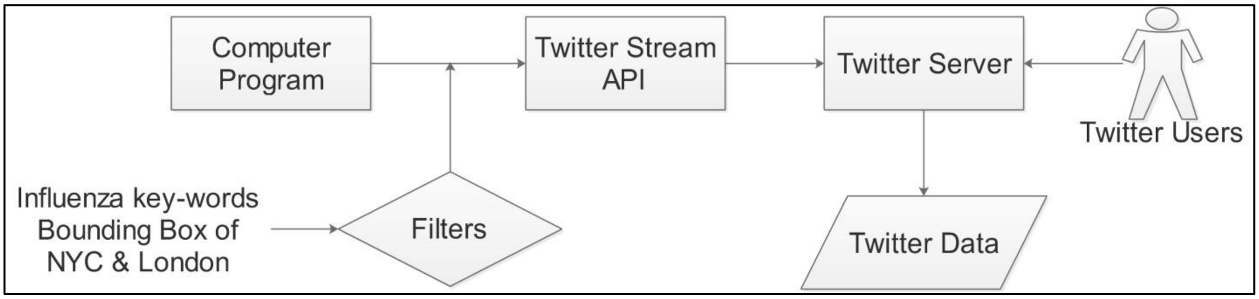
Data Collection

This research uses Vertalka's 2017 four step approach to digitally interact with social media users through Application Program Interfaces (API). As described by Vertalka (2017), social media APIs can be viewed as a conduit that connects Twitter users to the scholar and vice versa. Therefore, APIs are used to digitally communicate with social media users. During the digital conversation, SM users are encouraged to discuss their influenza symptoms, when symptoms peaked, and their home location (ZIP/Postal code) in an online questionnaire. Below, the four steps are discussed in more detail. This research also added an additional step to Vertalka’s (2017) approach (discussed in detail below) which adds more tweets the study.

Step 1: Data Twitter Capture

Step one of Vertalka's (2017) approach involves collecting SM data. Using Twitter's Stream API, this research captured over 14 million tweets (a 1% sample) within the bounding box coordinates of each NYC and London from November 26, 2014 to March 17, 2015. Tweets were captured based on keywords relating to influenza-like illness: 'sore throat', 'cough', 'headache', and 'fever' (Cullota, 2010). However, the captured tweets may not be a representative sample of Twitter activity (Morstatter, Pfeffer, Liu, & Carley, 2013). Additionally, it is likely not representative of the population in NYC or London.

Figure 2.1: Technical Architecture for Data Capture of Influenza Tweets in NYC and London



Step 2: Select Research Candidates

Step two of Vertalka's DIP approach centers on selecting groups of SM users that fit a research objective. In this research, purposive Twitter users were selected based on whether they tweeted within NYC and London and mentioned influenza-like symptom key-words 'flu', 'cough', 'sore throat', and 'headache’. These keywords show a 95% correlation with national influenza statistics in the US (Aron Culotta, 2010, 2013). A total of 8,696 and 7,756 purposive research candidates were matched to the above criteria for NYC and London, respectively. Random research participants were not selected as this group represents those that did not tweet about any influenza symptoms.

Step 3: Digital Conversation and Questionnaire Design

Step three of Vertalka's DIP approach centers on engaging in a digital conversation with identified research participants. Research participants identified in Step two were sent a tweet containing an external link to an online close-ended questionnaire. These tweets were sent through Twitter's API. A countless number of Twitter users engaged in a conversation with the scholar by tweeting back and forth. Many of these users asked for clarification of the research project and were curious about the authenticity of the scholar's account. All questions were met with appropriate responses.

The questionnaire was designed to identify which symptoms, if any, were experienced by the research participant, when those symptoms peaked, and the user’s home ZIP Code. It also asks for the user's handle name. The questionnaire did not ask for a user’s address as this is likely too personal of a question which will reduce response rates or increase the number of incomplete surveys. The questionnaire was powered by Google Forums and can be viewed in detail at http://goo.g/ko3qvx and http://goo.gl/NPemrp for NYC and London, respectively. Research participants were sent a link to the questionnaire three to five days after posting their influenza symptom since users may have tweeted about their influenza symptoms but have yet to experience peak influenza symptoms. Due to restrictions on how often Twitter's API can be accessed, research participants were not sent reminders to complete the questionnaire. Research participants only received additional tweets to complete the questionnaire if they continued to tweet influenza keywords. This was done because users might have experienced re-infection. Once the questionnaire was completed by the user, no additional request tweets were sent to that user. Of the 8,696 NYC potential research participants 275 completed the survey (2%) and of the 7,756 London potential research participants, 295 completed the survey (3%).

Step 4: Add Twitter Timeline Data

Since Twitter's Stream API returns only a 1% sample, this research used Twitter's Timeline API to gather historical tweets posted by each research participant completing the questionnaire. Collecting this data increased the tweet coverage of each user that is otherwise missed by the Stream API 1% sample. However, this approach can also introduce too much data since a

Twitter user's timeline tweet data could reference years’ worth of tweets and consequently multiple episodes of influenza. This issue is corrected as stated in the analysis section below.

Step 5: Data Merger

Stream and Timeline tweets from each research participant were linked to their questionnaire answers by their username.

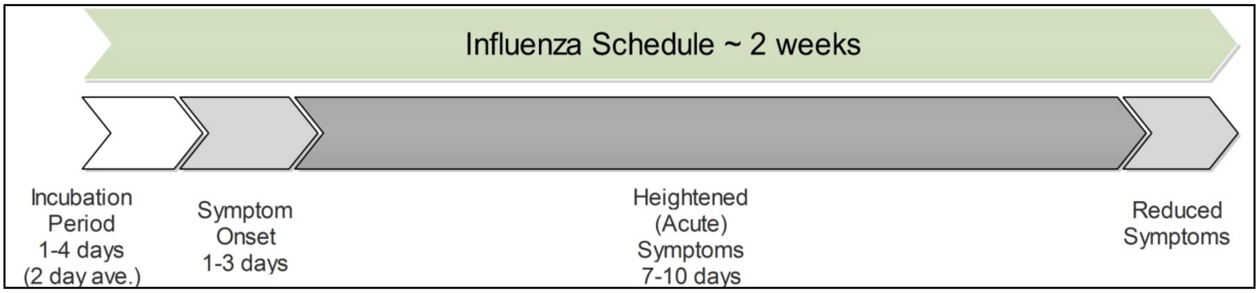
Analysis

This research used Vertalka's (2017) approach to add spatial and temporal context to Twitter users in NYC and London. This is accomplished by merging Twitter's Stream and Timeline data of identified research participants with their completed questionnaire. The below analysis is broken into a time and space component.

Time

The intention of this research is to identify the timetable of when influenza-ridden Twitter users most often tweet about their influenza symptoms. The questionnaire asked research participants when their influenza symptoms peaked. From this information, a timetable of when users tweet about their influenza can be built based on the schedule of influenza infection, shown in Figure 2.2. Upon infection, the virus will undergo a 1-4 day (mean: 2 days) incubation period (Fiore et al., 2008). After incubation, symptoms will quickly develop after mild symptoms are experienced from 1-5 days (mean: 2 days) to where acute symptoms can last for 7-10 days (Taubenberger & Morens, 2008). After this period, symptoms tend to quickly subside.

Figure 2.2: Timeline of Influenza Symptoms


Influenza Schedule ~ 2 weeks
Incubation Period 1-4 days (2 day ave.)
Symptom Onset 1-3 days
Heightened (Acute) Symptoms 7-10 days
Reduced Symptoms

Source: (Fiore et al., 2008 and Taubenberger and Morens, 2008).

A similar timetable was constructed for research participants and their influenza symptoms. By comparing the time a user posted their first influenza tweet to when they self-reported the time of their peak symptoms, an infection timetable can be identified for each user. The time period the first influenza tweet was posted was used since, it suggests the earliest possible indicator of influenza infection. This research identified 58 New York City and 48 London Twitter users that tweet about their influenza symptoms no more than 15 days prior to reporting when their peak influenza symptoms. Fifteen days was set as the cutoff because this number represents the duration of influenza infection; three days of symptom onset, plus the period of acute symptoms (7-10 days), plus two days of symptoms residing (Taubenberger & Morens, 2008). Influenza related tweets occurring before 15 days of self-reporting peak influenza symptoms may suggest a different type of illness, multiple infections, or the rare possibility of reinfection.

Space

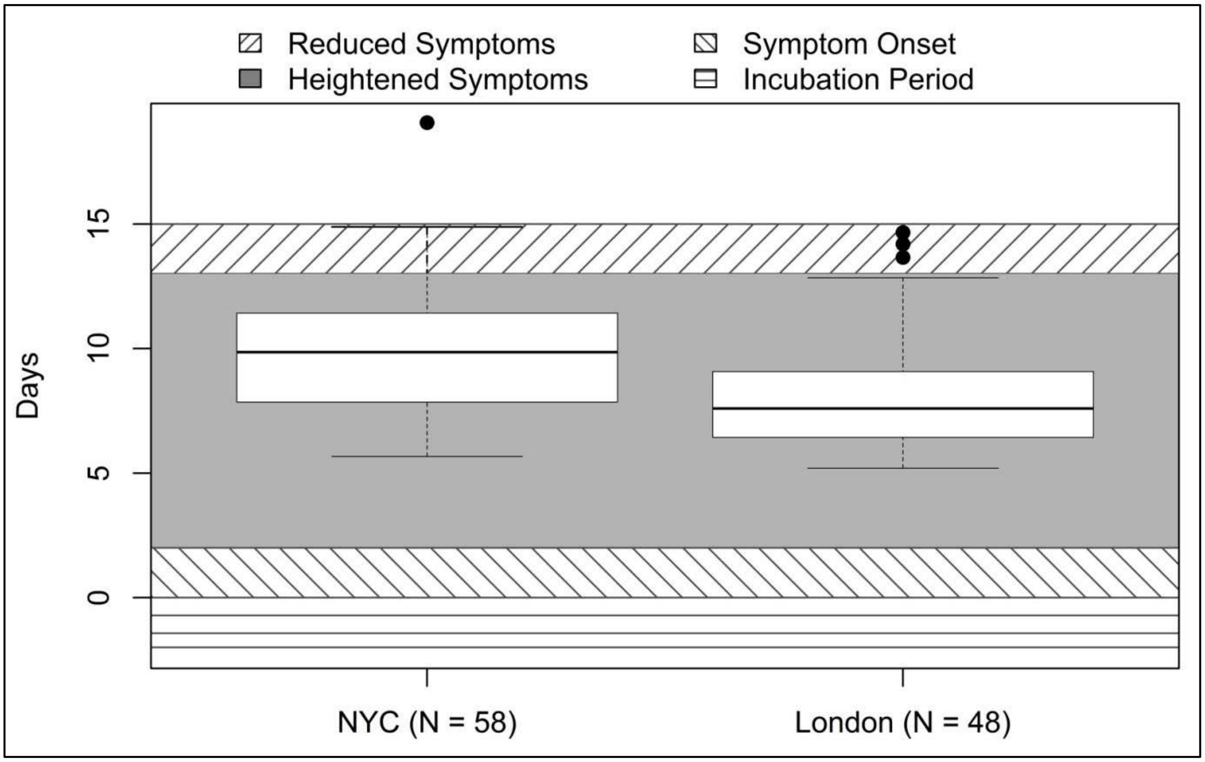
The survey asked research participants to select their home ZIP or Postal Code. This question is asked to cross-compare the research participants declared home ZIP or Postal Code in the questionnaire with the geotagged location data of the user's influenza and non-influenza tweets. This research identified 20 NYC and 53 London influenza-ridden Twitter users that reported influenza symptoms, provided their home ZIP or Postal Code in the questionnaire, and had geotagged enabled influenza and non-influenza tweets. Similar to the above, these users might have tweeted multiple times about their influenza sickness during a single influenza episode, which would bias any analysis towards the spatial tweeting behavior of a user. To account for this, this research used a user's first influenza tweet posted at no more than 15 days before users experienced peak symptoms. The identified research participants also posted multiple non-influenza tweets. This research kept only non-influenza tweets posted three months before and three months after the user's initial influenza tweet. Two metrics are used to compare the spatial distribution of influenza versus non-influenza tweets of the research participants. The first metric identified the percentage of influenza and non-influenza tweets originating within a research participant's home ZIP or Postal Code, outside their ZIP or Postal Code, and outside of the NYC or London Study area. To test for statistical differences between the mean rate of where influenza and non-influenza occur an Exact Test with a Poisson distribution was conducted. However, given that the occurrence of non-influenza tweets maybe dominated by only a few users, a second metric was used to gauge where influenza tweets tend to occur. The second metric identified the spatial pattern of where influenza tweets occur in comparison to non-influenza tweets. For metric two, Euclidean distance was measured from research participants declared home ZIP or Postal Code in the questionnaire to the geotagged location of the user's influenza and non-influenza tweets. The spatial distribution of influenza and non-influenza tweets for each user were then averaged. Averaging the spatial distribution of influenza and non-influenza tweets based on each research participant corrects the issue of any single research participant potentially dominating the distribution of where influenza tweets occur.

Results

Time

Figure 2.3 is a box and whiskers plot of the temporal distribution of when NYC and London users tend to tweet about their influenza symptoms. The box represents the days when 50% of users tweet about their symptoms. The solid line that divides the box represents the median time users tweet about their influenza symptoms. The whiskers or dashed lines above and below the box represent the upper and lower quartile, respectively. Points represent outliers. Nearly all NYC and London research participants tweet about their influenza during the period of heightened symptoms, as shown in Figure 2.3. In comparison to London research participants, NYC research participants tend to tweet later in the course of their influenza sickness, as indicated by the upper whisker's location in the reduced symptom boundary. Only one NYC user tweeted about their influenza symptoms after experiencing the worst of their symptoms. This user also represented an outlier.

Figure 2.3: Temporal Distribution of When Influenza Tweets were Posted



Space

Roughly, 76% of NYC and 72% of London influenza-ridden Twitter user's first tweet influenza symptoms in their self-reported ZIP or Postal Code. It cannot be determined if ZIP or Postal Code originating influenza tweets represents a user's home address or some other location the user frequents within their home ZIP or Postal Code. Of these same users, only 68% and 64% of their non-influenza tweets originate in their ZIP and Postal Code, respectively. In general users in both New York City and London tend to tweet about their influenza symptoms close to home, in comparison to their usual spatial pattern of tweeting. The Exact Test with a Poisson distribution suggests that there is a statistical significance in the difference between where NYC and London users tweet influenza and non-influenza tweets.

Table 2.1: Spatial Locations of Where Influenza and Non-Influenza Tweets are Posted

City	Non-Influenza Tweets			Influenza Tweets			Poisson Results
	Outside of Study Area	Outside ZIP/Postal Code	Within ZIP/Postal Code	Outside of Study Area	Outside ZIP/Postal Code	Within ZIP/Postal Code	
New York City	630 (4%)	4659 (28%)	11625 (68 %)	0 (0%)	7 (24%)	22 (76%)	p-value < 2.2e-16 Distribution is different
London	463 (2%)	6489 (33%)	12596 (64%)	0 (0%)	14 (28%)	36 (72%)	p-value < 2.2e-16 Distribution is different

Both Figure 2.4 and 2.5 display the average spatial distributions of users' geolocated influenza tweets and their non-influenza tweets in NYC and London, respectively. The average user tends to tweet about their influenza symptoms closer to their self-reported home ZIP or Postal Code, in comparison to their non-influenza tweets. For both NYC and London, influenza tweets tend to occur within a one-kilometer radius of a user’s home ZIP or Postal Code. Whereas, these same users tend to tweet about their non-influenza symptoms around a two-kilometer radius of their home ZIP or Postal Code as suggested by the large spike of non-influenza tweets in Figure 2.4 and Figure 2.5. This reinforces Birkin and Malleson (2012) findings where tweets tend to occur at about 2.5 km from the user's home. The large spikes in non-influenza tweets could be a result of many different urban forms and functions such as, public transportation stops, mean distance to work, and other common urban activity spaces that are near population clusters. They could also indicate the spatial relationship between residential and commercial areas. The spike near kilometer zero represents residential areas whereas the spike near two kilometers represents a commercial district where users congregate and frequently tweet (Birkin and Malleson, 2012).

Figure 2.4: Spatial Distributions of Where Influenza and Non-Influenza Tweets are Posted from Home ZIP Codes in NYC

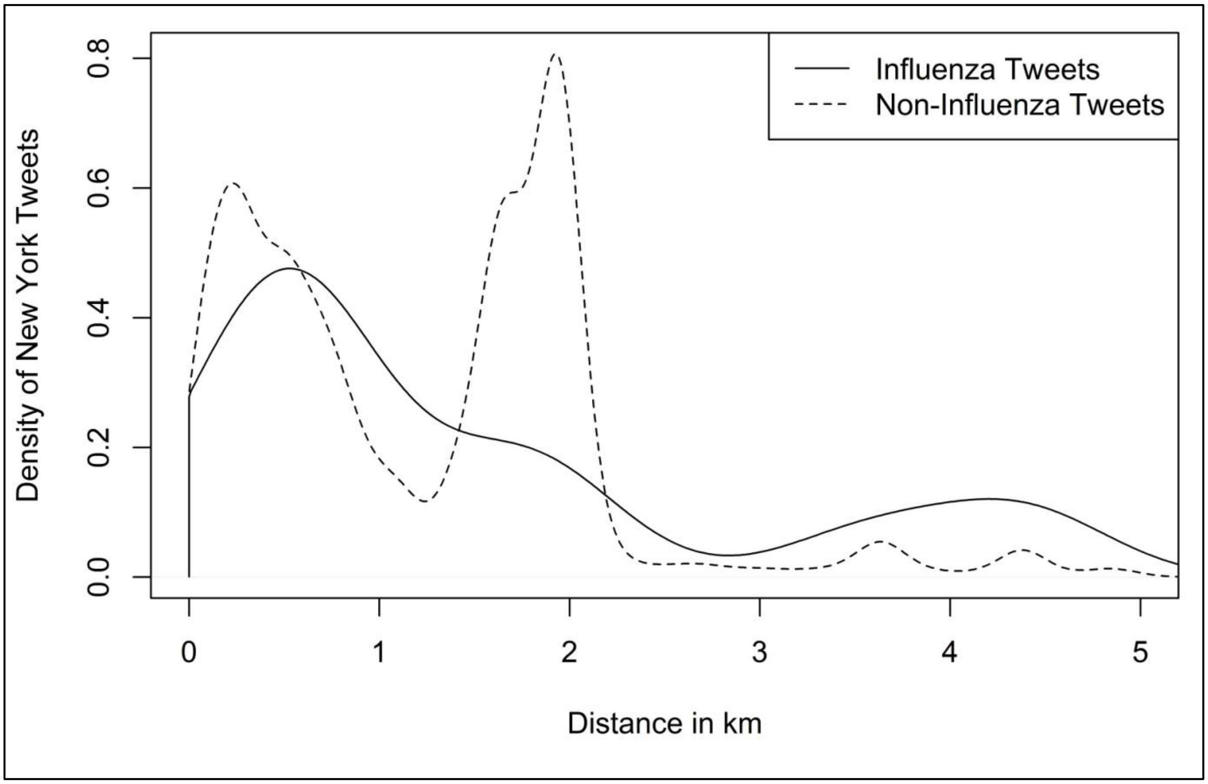
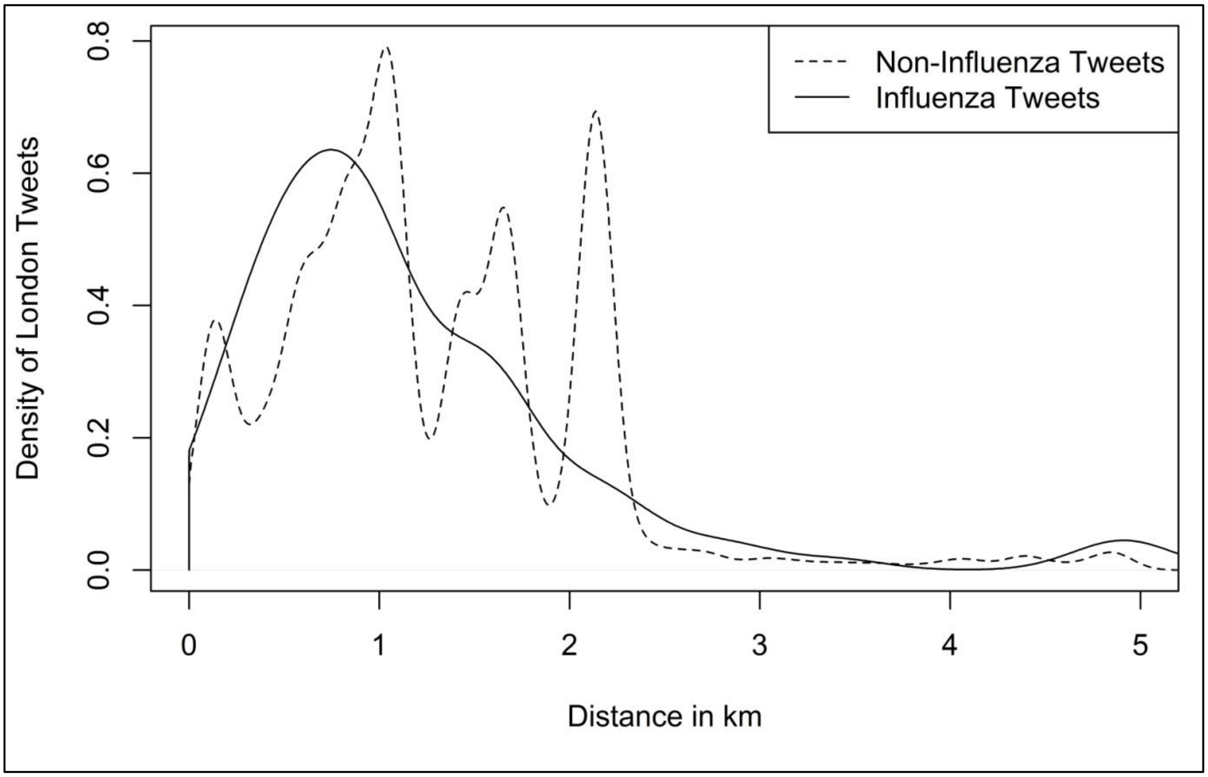


Figure 2.5: Spatial Distributions of Where Influenza and Non-Influenza Tweets are Posted from Home Postal Codes in London



Discussion

The goal of this research was to identify how Twitter user's space-time path alters when experiencing influenza-like symptoms. This was accomplished by using Vertalka's (2017) approach for augmenting social media data. Traditionally, space-time paths of individuals were recorded by that individual keeping a detailed diary of where and when they did different activities. This process was time consuming and expensive for researchers. Twitter, and other social media outlets, are a new data source were space-time path data can be collected and analyzed inexpensively and quick.,

Prior research focusing on space-time paths of Twitter users does not account for the user's state of health. As this research has shown, space-time paths of Twitter users are likely altered when they are experiencing influenza-like illness. The user’s state of health is an important factor to consider because their day-to-day activities change which effects their spatial-temporal path. For instance, Twitter users who are cancer victims are likely to possess a unique space-time path. On one hand, these Twitter users will have a limited range around their residency as they maybe too tired to travel for daily errands or activities of live considering their health state. However, these Twitter users are likely traveling large distances to seek medical care for their cancer. In this case, cancer has a very different space-time path when compared to influenza.

Twitter users who tweet about their symptoms often do so when experiencing acute symptoms. One possible explanation to this may rest on the user experiencing maximum frustration or misery, and once that point is reached it is then expressed through a tweet. Another possible explanation is that a Twitter user is bored at home and has nothing better to do than tweet about their symptoms. It was identified that 76% of NYC and 72% London influenza-ridden tweets originate within the author's home ZIP or Postal Code. When not experiencing influenza symptoms 68% of NYC and 64% of London tweets originate in the authors home ZIP or Postal Code. Furthermore, on average influenza tweets are more concentrated near the user's self-reported home ZIP or Postal Code in comparison to the spatial pattern of a user's non-influenza tweet, as shown in Figure 2.4 and 2.5. This indicates that a user's space-time path during influenza infection has less spatial variation. These influenza-ridden users might be experiencing a frustrating level of influenza symptoms and consequently are staying home. Users tweeting about their symptoms outside of their home ZIP or Postal Code suggests users are experiencing influenza symptoms at work, school, doctor's office, or other urban activity spaces that is not in their home ZIP or Postal Code.

The space-time path difference between New York City and London Twitter users may rest on several factors. Space-time paths are dependent on people's interaction with the natural or built landscape. Therefore, the design or spatial arrangement of the different urban forms and functions of cities will influence where people are likely to tweet. For instance, research has already identified how population, job density, stores within a certain radius, and distance to the central business district, affect the way people interact with the urban landscape (Bhat & Misra, 1999; Ettema, 2005; Yamamoto & Kitamura, 1999). Furthermore, the distance to which people interact with the urban environment decreases as the city becomes more compact, has higher degrees of mixed landuse, more options for public transportation, and is more connected (Boarnet & Crane, 2001; Handy, 2005; Khattak & Rodriguez, 2005; Kockelman, 1997). This is because individuals have more opportunity to interact with the spaces around them and do not need to travel for such opportunities. However, this relationship may also be symbiotic as individuals not only interact with the built landscape but mold it to their preferred needs and desires so their space-time paths become more convenient (Crooks et al., 2015). As mentioned in the above literature, Twitter has been used a proxy for mapping out the space-time paths of users. Therefore, it is not surprising that the location of influenza and non-influenza tweets in New York City and London differ, as they would differ between cities of varying urban form and function. Perhaps though future research can identify if a generalized spatial-temporal distribution pattern exists between cities that are composed of similar types of culture, economics, design, or built by similar planning theory and practices.

Understanding a deeper context to the spatial-temporal pattern of influenza tweets corrects two assumptions of digitally detecting influenza cases through Twitter. The first assumption assumes that Twitter users experiencing influenza symptoms post such material as symptoms start.

However, this assumption is not entirely correct. This research identified that Twitter users experiencing influenza-like symptoms first tweet during periods of acute symptoms or2-12 days after symptom onset. Second, digital detection of influenza cases through Twitter assumes users tweet about their influenza symptoms at a coarse spatial scale, such as national or local scale. However, Twitter users stay closer to home during periods of infection, suggesting digitally detecting influenza cases through Twitter can occur at spatial scales finer than national, state, and metropolitan level.

As shown in this research, the space-time path of individuals is affected by their influenza infection. Other health issues, such as cancer, diabetes, might also affect the space-time paths of individuals. While it remains unstudied in the fullest of scopes, researchers who use Twitter as a data source for understanding space-time paths should account for the health status of the user as it is likely to change the space-time path pattern of that user.

Conclusion

This research discusses significance and importance of examining the health status of a Twitter user when examining their space-time path. Twitter has been used as a data source to distinguish between the diurnal activities of people and identify urban form and function such as, users preferring to tweet in commercial areas within 2.5 km from their home. However, no prior research has identified how a Twitter user's state of health will alter their space-time path. This research addresses that shortcoming and identifies two significant contributions to digitally detecting influenza cases. First, traditional digital influenza detection research assumes the time a user tweets about their symptoms reflects when they first experience symptoms. This research demonstrated that Twitter users tweeting about their influenza symptoms do so when their symptoms are acute. Second, this line of research does not include any literature for where an influenza tweet originates. This research concludes that influenza-ridden Twitter users tweet about their symptoms in a more spatially confined area in comparison to users' typical tweeting area. Additionally, the location of influenza tweets tends to be within or very near a user's home ZIP or Postal Code. Those tweeting about their influenza symptoms outside of their home ZIP or Postal Code, tend to do so within three kilometers of their home ZIP or Postal Code.

The findings of this research should not be universally applied to all cases of digitally detecting influenza cases. The tweeting behavior of individuals may vary based on many criteria some including health state of the user, demographics, cultural identify of user, availability of reliable internet, and other social-economic factors. Future research needs to focus on identifying factors that influence the spatial and temporal distribution of where and when influenza-ridden Twitter users tweet. One approach to solve this problem is to apply Vertalka's (2017) Digital Interaction Program (DIP). DIP allows scholars to digitally engage and recruit a wide variety of social media users (from diverse demographics, culture, etc) as research participants in scientific research. In turn, the spatial and temporal distributions of different populations can be identified. Furthermore, the severity of influenza varies from season to season as symptom intensity varies from person to person. It is expected that the Twitter user will remain more geographically confined during influenza outbreaks that present severe symptoms, as users are more likely to stay at home. However, this theory remains untested. Lastly, future research can be undertaken which examines how space-time paths are affected by different health issues within similar and different urban layouts, transportation networks, cultures.

This research has several limitations. First, DIP questionnaire participants could report false information about their home ZIP or Postal Code and when they experienced their worst symptoms. Second, the findings presented in this research are not intended to be a generalized rule. Rather, influenza tweets in different cities and influenza seasons are likely to create different spatial-temporal distributions. Third, this research contained a small number of research participants. Having a larger sample size would produce a more robust distribution of when and where influenza occur.

TRANSITION

Chapter two identified spatial and temporal shortcomings of digitally detecting influenza cases through Twitter. Research on the digital detection of influenza through Twitter has traditionally focused on comparing influenza tweets with national or regional scale influenza prevalence. Research has assumed Twitter users tweet about their symptoms at the moment of onset and from their residency. Chapter two used the Digital Interaction Program (see Chapter one) to correct these assumptions by identifying the spatial and temporal distributions of influenza tweets. New York City and London Twitter users tend to tweet about their influenza symptoms from locations closer to their home ZIP or Postal Code, in comparison to where these same users tweet about their daily lives. People in both New York City and London tend to tweet about their influenza symptoms when they experience acute symptoms. It is expected that the results of Chapter two will vary according to city, influenza season, demographics, culture, and a host of other reasons.

DIP not only asked questions relating to when users experienced peak influenza symptoms, home ZIP or Postal Code, and which influenza symptoms they experienced, but also asked about demographics, gender, age, and if the user received an influenza vaccine. Unfortunately, DIP had a low participation rate and therefore could not provide enough cases to stratify the findings presented in Chapter two by these factors.

Chapter three builds on the findings presented in Chapter two. Influenza-ridden Twitter users tend to tweet about their influenza symptoms near their home ZIP or Postal Code. This suggests that correlating influenza tweets to actual influenza cases can occur at intra-metropolitan spatial scales. Chapter three explores that possibility by correlating influenza tweets with influenza Emergency Room (ER) admissions for each hospital in New York City. Chapter two also suggests that the occurrence of influenza tweets succeeds the actual time point of infection. This may indicate that influenza tweets tend to be posted before the actual occurrence of influenza cases. Therefore, Chapter three explores when peak correlation occurs between influenza tweets and influenza emergency room admissions at each hospital in New York City. Being able to predict influenza cases at a finer geographic scale introduces the possibility of a local influenza surveillance approach for hospitals, which is a novel outcome in disease surveillance. This research is not intended to replace current surveillance approaches but rather to provide hospitals with added knowledge about local influenza prevalence and its impacts.

The capabilities of Twitter as a reliable data source for influenza detection at individual hospitals varies. At some hospitals influenza tweets are able to predict influenza ER admissions and at other hospitals the opposite is true. Best cases involved influenza tweets predicting influenza ER admissions over two weeks in advance. Unfortunately, the correlations coefficients where relatively low. It is unknown as to what hospital characteristics contribute to influenza tweets predicting influenza ER admissions and correlation values. Chapter three uses regression methods to dig deeper into identifying and understanding what socio-economic factors influence when influenza tweets predict influenza ER admissions.

The analysis in Chapter three could not be conducted with the London dataset. Transferring the required data from London to the United States is near impossible since the United Kingdom Courts recently increased the security requirements of trans-Atlantic transfers of sensitive data. In fact, as of writing this dissertation, no academic institution has successfully applied to receive sensitive data from the United Kingdom since the increased security requirements. This dissertation was able to obtain influenza data for London, but it was a dataset that only included the number of broadly defined influenza patients entering a hospital on a specific date. For instance, eight people entered hospital 'X' on December1st, 2014. The diagnosis classification for patients entering a hospital are coded as 'ears, nose, and throat'. This coding scheme covers a wide spectrum of diagnoses outside of the intended influenza diagnosis. However, there was no available option to receive patient data with more detailed diagnosis. Furthermore, the dataset provided suppressed daily hospital admissions with fewer than five patients. As a comparison, the New York City dataset does not suppress the raw data but the data use agreement requires the researcher to suppress the data when disseminating the results. Therefore, in its raw but suppressed state, the London dataset arrived without the necessary detail needed for a strong analysis. Additionally, the London dataset was missing patient Postal Codes, a necessary component to analyze the data. This data piece is important because it is used as spatial weights to identify the geographic extent of where influenza patients for each hospital reside. It is not feasible to generalize the spatial weights based on the catchment area of influenza patients for New York City hospitals.

CHAPTER 3
USING TWITTER AS A LOCAL INFLUENZA SURVEILLANCE SYSTEM

Abstract

Current research correlates influenza tweets to influenza cases at broad national or regional scales. This research harnesses Twitter as a contributed data source to identify when influenza tweets best correlate with influenza Emergency Room admissions at different local hospitals in New York City. Tweets were downloaded based on influenza keywords and geotagged within New York City. An ensemble of machine learning algorithms classified tweets based on the context of who is infected with influenza. Optimal lag and lead periods were identified of when self-diagnosed influenza tweets best correlate to influenza Emergency Room admissions. On average influenza admissions were predicted about 8.5 days in advance. This research also identified demographic and transportation factors having significant influence on when influenza tweets best correlate to influenza Emergency Room admissions in New York City. This research introduces a more spatially appropriate indicator to detect potential influenza Emergency Room admissions. Which in turn may potentially provide warning to individual hospitals to prepare their facilities for incoming influenza patients.

Introduction/Background

Pandemic and seasonal influenza outbreak occurrences are expected to further stress hospital Emergency Rooms (ERs), despite modern intervention strategies and surveillance programs (Derlet, Richards, & Kravitz, 2001; Dugas et al., 2012; Trzeciak & Rivers, 2003). These stresses include: shortages of medical staff and supplies, rendering ERs less effective at patient treatment. In some cases, ambulances are diverted from ERs operating at or over patient capacity (Schull, Morrison, Vermeulen, & Redelmeier, 2003). Anti-viral interventions intended to limit the severity and scope of influenza outbreaks may not be an effective remedy for seasonal or pandemic influenza, and ultimately influenza ER admissions (Lessler, Reich, & Cummings, 2009). For instance, influenza antiviral drugs administered in 2009 and 2014 were 56% and 13% effective, respectively, at preventing influenza ("Center for Disease Control and Prevention," 2016). Given these shortcomings, national and local surveillance programs are established to monitor influenza incidence and prevalence. For example, the Center for Disease Control (CDC) monitors influenza through five metrics: virology, healthcare outpatient illnesses, mortality, hospitalizations, and geographic spread as briefly defined below.

Virology: Over 300 labs report the total count of tests for influenza and the number of respiratory specimens positive for influenza.

Outpatient Influenza-like Illness Surveillance Network (ILINet): The number of patient visits to healthcare facilities where the patient presents influenza-like illness symptoms (headache, fever, cough, sore throat, vomiting, diarrhea).

Mortality: Tracking influenza mortality through the National Center for Health Statistics and Influenza-Associated Pediatric Mortality Surveillance System.

Hospitalizations: Monitoring laboratory confirmed influenza hospitalizations for adults and children through the Influenza Hospitalization Surveillance Program.

Geographic Spread: Reporting on the state-wide geographic spread of influenza as either *no activity*, *sporadic, local, regional*, or *widespread* by state health departments.

Results from these surveillance programs are published 1-2 weeks after the occurrence of influenza cases (Broniatowski et al., 2013; Ginsberg et al., 2009). Spatially speaking, the CDC reports their findings at the state level which provides little information to individual hospitals that typically serve immediate surrounding communities.

Some metropolitan areas also employ their own surveillance program. For instance, New York City's Department of Health and Mental Hygiene (NYCDOHMH) conduct their own version of influenza surveillance. However, their approach is somewhat similar to that of the CDC's as NYC monitors the following programs:

Outpatient Influenza-like Illness Surveillance Network (ILINet): ILINet is monitored by the CDC, as discussed above.

Syndromic Surveillance: Emergency Departments send electronic data to NYCDOHMH regarding influenza cases. This is used to analyze changes in the trend of influenza cases.

ED ILI Visits versus ER ILI Admissions: NYCDOHMH identifies the proportion of influenza cases that require admissions. This is used to identify the severity of influenza.

Laboratory Confirmed Influenza Cases: NYCDOHMH identifies laboratory confirmed influenza cases for more accurate determination of influenza severity.

Traditional data sources, such as those by the CDC, are collected, analyzed, and disseminated by scientific experts. Health practitioners and researchers have looked at alternative sources by untrained people to monitor influenza activity including, Google FluTrends and Healthmap.com. In 2008, Google correlated influenza search queries with CDC influenza cases in an application called Google FluTrends (GFT). GFT provided influenza surveillance data at the state level and had begun testing at the city level (Ginsberg et al., 2009). For the better part of five years the application seemed accurate. In 2013, however, GFT over-estimated influenza prevalence due to sustained spikes of influenza media coverage, causing an increase in public concern. This in turn created an increase in influenza search queries and ultimately GFT overestimating influenza prevalence (Butler, 2013). Since the report of this miscalculation, GFT has remained offline.

Despite its hiccup, GFT has not published the necessary geographic detail of influenza cases to create local influenza insight for hospital ERs. Healthmap.com is another online surveillance program but unlike GFT, it offers high geographic resolution about the prevalence of several communicable diseases (Healthmap.com). This is accomplished by sophisticated algorithms identifying content found in media, social media, and official reports from international organizations associated with any given disease. However, the site has yet to correlate the occurrence of a disease to respected ER admissions.

Given the shortcomings of GFT and Healthmap.org, researchers have sought to digitally detect influenza cases with Twitter. Twitter is a micro-blogging website that allows users to post messages (tweets) of up to 140 characters. Twitter users may post about a variety of topics that occur in their daily lives. Some tweets contain 'hashtags' (#) to indicate trending topics or emojis that act as ideograms. Twitter records the time a tweet is posted and, if user enabled, where the tweet was posted. As of 2016, roughly 303 million tweets are posted each day (Business Insider, 2016).

Alternative data sources such as, Twitter and other SM outlets, represent a different type of data that is permissive in nature. This type of data has been referred to as implicit Volunteer Geographic Information (VGI) (Senaratne et al., 2017). VGI refers to non-experts collecting spatial data about the world and uploading that data onto a community computer. Early VGI involved non-experts to contribute data to create OpenStreetMap or Wikimapedia (Sui et al., 2012; Tulloch, 2014). This type of VGI data has an explicit role. For instance, users who upload data on OpenStreetMap do so with an intended purpose of being active contributors. Twitter and many SM data outlets, however, are not driven by the scientific contribution purpose of OpenStreetMap or Wikimapedia. Therefore, SM data outlets used in scientific research are often more littered with irrelevant data observations when compared to early forms of VGI. Despite its permissive nature, Twitter has been effective at predictive research including earthquake detection (Sakaki et al., 2010), stock market trends (Bollen et al., 2011), grass fires in western US (Vieweg et al., 2010), and use or misuse of antibiotics (Scanfeld, Scanfeld, & Larson, 2010). Twitter has even been used for detecting influenza cases. In such cases, research has focused on identifying keywords associated with influenza tweets (Aron Culotta, 2010), linking influenza related tweets to national news stories (Chew & Eysenbach, 2010), correlating influenza tweets with national influenza rates (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011) or regional influenza rates (Bodnar & Salathé, 2013), and even correlating influenza rates in New York City with influenza tweets (Broniatowski et al., 2013). The scholars detecting influenza cases were able to do so with correlation rates between 0.70 and 0.97 and have focused on national or metropolitan scales. However, Vertalka (2017) identified that influenza tweets tend to occur in spatial pockets near a user's home ZIP or Postal Code, suggesting that digitally detecting influenza outbreaks can occur at finer spatial scales, such as local hospitals.

This research seeks to identify when and where influenza tweets can best correlate with influenza ER admissions in New York City and then examines the factors that influence the correlation values. This article first discusses the process of collecting and cleaning hospital and Twitter data located in New York City. Next, it temporally finds the optimal time to which influenza tweets are best correlated with influenza ER admissions. This research uses multiple regression methods to then identify demographic, economic, and transportation factors that influence the correlations. The results of this paper indicate that Twitter may timely and local insight for several hospitals in NYC.

Hospital Data

This research collected inpatient and outpatient ER hospital admissions in New York City from Statewide Planning and Research Cooperative System (SPARCS) from November 26th, 2014 to March 17th, 2015. Inpatient and outpatient influenza cases were selected based on ICD-9-CM codes that are indicative of influenza-like illness (Appendix 3.1) (Marsden-Haug et al., 2007). Family doctors and house-call doctors treating influenza patients were not selected since these types of clinics do not offer medical emergency services. Over 66,000 influenza ER admissions occurred in NYC during this period. Roughly 90% of these admissions occurred as outpatient cases. As expected, the burden of influenza cases affected the very young and very old, as shown in Table 3.1.

Table 3.1: Age Adjusted Influenza Cases per 1,000 People

Age Range	Total Influenza Cases (Age Adjusted rate* per 1,000 People)
Persons Less Than 5	13,577 (23.25)
Persons Between 5 and 18	14,928 (11.41)
Persons Between 18 and 39	17,365 (5.69)
Persons Between 40 and 64	13,485 (4.82)
Persons Over 64	6,909 (6.06)

*denominator represents population total of each age range

Twitter Data

This research downloaded roughly 14 million tweets from Twitter's Application Program Interface (API) with location filters based on the bounding box of New York City and its surrounding Boroughs from November 26th, 2014 to March 17th, 2015. Twitter's stream API returned a 1% sample of all tweeting activity within this area and time period.

It should be noted that captured tweets may not be a representative sample of all Twitter activity (Morstatter et al., 2013). Additionally, it may not representative of the population at hand. Roughly 25% of adults use Twitter but that percent is predominantly young, affluent, educated, and non-rural (Pew Research Center, 2015). These statistics can be further broken down based on the number of Twitter accounts versus who frequently tweets. For example, 20 and 30-year olds represent over 50% of all Twitter accounts but makeup less than 40% of all Twitter activity. However, teenagers account for about half of all Twitter activity but makeup roughly 15% of all Twitter accounts. Furthermore, there are more males than females on Twitter. Females, however, tweet more frequently. While there are more white Twitter users, black Twitter users are overrepresented (Murthy, Gross, & Pensavalle, 2016). The demographic over and underrepresentation of Twitter users will spatially vary (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). The demographic skewness is likely to affect the outcome of this research. However, the goal of this research is to use influenza tweets as a proxy for understanding influenza activity for individual hospitals and not to predict whether a person tweeting about influenza will enter an ER.

Methods

Prior research was limited to applying influenza tweets to predict influenza cases at broad scales. The purpose of this research is to identify the role Twitter plays in acting as proxy to understand ER influenza activity for ER hospitals in NYC. To accomplish this, a four-step data analysis process was developed. Part one centers on cleaning the downloaded Twitter data to identify tweets focusing on users self-diagnosing themselves with influenza-like symptoms. Part two discusses when influenza tweets are best correlated with influenza activity in NYC. Part three focuses on when and where influenza tweets are best correlated with influenza ER admissions at individual ER hospitals. Part four discusses the socio-economic factors that influence the correlation value of where and when influenza tweets coincide with influenza ER admissions.

Part 1: Tweet Data Cleaning

Part 1 of the methods portion contains five steps. Step one removes retweets. Step two identifies tweets with keywords associated with influenza-like symptoms. Step three examines the content of an influenza tweet to identify self-diagnosed influenza case. Step four describes a lapse period where influenza infected Twitter users are not allowed to have two influenza episodes within a specified window. Finally, Step five removes all tweets that are not GPS enabled and located outside of NYC.

Step 1: Data Cleaning - Removing Retweets

Retweets, or reposts, were removed from the dataset to eliminate unnecessary tweet redundancy. For example, international music pop star Justin Bieber may have tweeted about his influenza sickness, “The show can’t go on. I’m sick w/ the flu!! Sorry fans” and subsequently a firestorm of retweets might follow given his popularity on Twitter. Therefore, removing retweets also helps eliminate over counting influenza cases.

Step 2: Data Cleaning - Identifying Tweets with influenza key-words

This research used two approaches to ensure accurate selection of influenza related tweets. First, this research selected tweets containing the following key-words: influenza, flu, headache, cough, and sore throat as these keywords show a 95% correlation with influenza activity (A

Culotta, 2013; Aron Culotta, 2010). However, selecting tweets based solely on influenza keywords would indicate that the following two tweets are equally important:

1.) "Well, I have a sore throat, headache, and cough....flu?"

2.) "I guess the flu is bad this year...at least according to the CDC."

These two tweets are contextually different but contain at least one influenza keyword. Tweet one is a primary influenza case where the user has self-diagnosed themself with the flu. Tweet two, suggests that a user is posting a public service announcement potentially warning other users about flu activity.

Step 3: Data Cleaning – Identifying self-diagnosed influenza tweets based on context

Using influenza key-words does not account for contextual differences between tweets. To correct this issue, the contextual differences between different tweet context are identified using an ensemble of machine learning algorithms (MLA).

An ensemble of machine learning algorithms (MLA) was used to distinguish between the above self-diagnosed tweets, public service announcement tweets, and non-influenza tweets. An ensemble of MLAs provides more robust analytical results compared to running individual MLAs. MLAs are a type of artificial intelligence where computers learn how to perform a task or make a decision. For instance, linear regression is a machine learning method where the computer learns the trend of a dataset and then using that trend can make future predictions. MLAs have been used to classify a Twitter user's political affiliation (Pennacchiotti & Popescu, 2011) and sentiment (Go, Bhayani, & Huang, 2009). This research uses R Statistical Software's "RTextTools" package (Jurka, Collingwood, Boydstun, Grossman, & van Atteveldt, 2013) to train Support Vector Machine, Logit Boosting, Bagging, Random Forests, Artificial Neural Net, and Decision Tree models on the three tweet categories (non-influenza, self-diagnosed influenza, and public service announcement). Each of these models then independently classifies a tweet based on what the model has learned about the context of the different tweet categories. After all the models classify a tweet, a democratic vote occurs between the different models to reach a consensus on the classification of a tweet as either a self-diagnosed case, PSA, or non-influenza tweet.

For the MLAs to classify different tweets, an independent training dataset from Twitter's API was coded to detect the difference between self-diagnosed influenza tweets, influenza public service announcement (PSA) tweets, and non-influenza tweets. Over 2,000 tweets were read and coded by the author and a colleague in a double-blind method; 1,286 for self-diagnosed influenza tweets, 802 for secondary influenza tweets, and 103 for PSA. Any coding discrepancies between the two coders were either corrected or eliminated from the study if consensus could not be reached. The training dataset is built on the co-occurrence of non-sparse terms

$$\begin{array}{lll} x_{11} x_{12} \cdots & x_{1p} \cdots & y_1 \\ x_{21} x_{22} \cdots & x_{2p} \cdots & y_2 \\ x_{n1} x_{n2} \cdots & x_{np} \cdots & y_n \end{array}$$

where x equals a single variable term and y equals the classified code for each tweet. The ensemble of MLAs was calibrated and validated using the above training dataset and K-folds cross validation. K-folds cross validation is a method where the training dataset is split into K number of parts. One of those parts is used to train the MLAs and the other K-1 parts are used to test the MLAs. This process is repeated until every single split has trained the MLAs and predicted the K-1 remaining parts. This research used ten folds to detect influenza tweets (Bodnar & Salathé, 2013; Kohavi, 1995).

Step 4: Data Cleaning - Syndrome Elapse Time

According to the CDC (2016), influenza-like symptoms may occur for two weeks and consequently users may tweet about their influenza symptoms within this period (initial symptoms, peak symptoms, missing work/school/social activity, visiting a doctor, and recovery). See Table 3.2 for examples. While influenza tweets are often posted during periods of heightened symptoms (Vertalka, 2017), continuous tweeting of influenza-symptoms by an individual will overestimate influenza activity. To account for this behavior, a six day window is typically created around the account holder's first influenza tweet (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2011). However, this research used a 14-day window because influenza symptoms typically last up to 14 days. Any influenza tweet within that window is discarded, except for the user's first influenza tweet. Any tweet outside of that window is treated as a separate influenza case for that individual.

Table 3.2: Influenza-like Tweets According to Syndrome Elapse Timeline

Influenza Activity	Example Tweet
Initial Symptoms	"Headache and a sore throat…this might be the flu."
Peak Symptoms	"I feel like death. Everything hurts. #IHateFlu"
Missing Activity	"Calling into work sick….again b/c of flu."
Visiting a Doctor	"Coughing in the Doctor's waiting room is like coughing on a cough"
Recovery	"Finally starting to feel better. I never want the flu again!!!!"

Step 5: Data Cleaning - Select tweets that are posted within NYC

Step 5 spatially selects tweets that are posted within NYC and its five Buroughs using the GPS coordinates of the tweet and not the user declared location data. The difference between these two is important. GPS coordinates of tweets represents the location of where the tweet originated from. On the other hand, user declared location data is supposed to represent where a user lives. This latter case presents several issues as users can falsely describe their location (declare their home is in NYC but actually live in London), fictitiously describe their location (Candyland, Jurassic Park, etc), or broadly define their location (New York State instead of Brooklyn). Given these shortcomings only GPS enabled tweets posted in NYC were used in this research.

Part 2: Optimally Identifying when Influenza Tweets Predict ER Admissions in NYC

Previous research has shown that influenza related tweets can predict the actual occurrence of influenza cases by 14 days (Aramaki, Maskawa, & Morita, 2011; Broniatowski et al., 2013; Aron Culotta, 2010). Pearson's correlation coefficient is used to identify when influenza tweets best correlate with influenza cases in NYC. Where X equals all ER influenza cases in NYC, Y equals daily influenza tweets as a proportion of daily tweeting activity (Lamb et al., 2013) for all of NYC, and r equals the correlation value between 0 and 1. Normalizing influenza tweets by overall tweeting activity corrects issues of Twitter activity varying on certain days of the week, month, or year.

$$r = \frac{\Sigma\mathrm{xy} - \frac{\Sigma\mathrm{x}\Sigma\mathrm{y}}{\mathrm{N}}}{\sqrt{(\Sigma x^2 - \frac{(\Sigma x)^2}{N})} + \sqrt{(\Sigma y^2 - \frac{(\Sigma \mathrm{y})^2}{N})}}$$

There are many advantages to using this method such as, the ease of interpretation. However, one of the drawbacks with this method, and any correlation method, is that it does not indicate causality. Therefore, the occurrence of influenza tweets may not directly influence influenza ER admissions. While this may be true, it is not the intention of this research to directly link influenza tweets to influenza ER admissions. Rather this research's goal is to identify when influenza tweets best correlate with influenza activity at local ERs. Pearson's correlation coefficient was repeated to identify the maximum correlation value to when either influenza tweets best predicted influenza ER admissions or vice versa. In economics, this process is often referred to as lag and lead indicators. Lagging indicators look back at some number of rows by temporally shifting the data to down or to the right, which moves it forward in time. Leading indicators look ahead some number of rows by temporally shifting the data to the down or to the left, which moves it back in time. Table 3.3 shows an example how lag and leads shift.

Table 3.3: Lag and Lead Shifting Example

Time	Value	Lag Value	Lead Value
10/01/2017	100	NA	99
10/02/2017	99	100	98
10/03/2017	98	98	97
10/04/2017	97	97	NA

This research shifts influenza tweets to the right (lag) to identify when influenza tweets best correlate with influenza ER admissions and shift influenza tweets to the left (lead) to identify when influenza ER admissions correlate with influenza tweets.

Part 3: Identifying when Influenza Tweets Optimally Predict Influenza ER Admissions on a Hospital by Hospital Basis

This research identifies a finer geographic scale of where and when influenza tweets best correlate with influenza ER admissions for each hospital in NYC. The time influenza tweets optimally correlate with influenza ER admissions is likely to be different for each hospital. This research uses two steps to spatially and temporally correlate when influenza tweets will optimally coincide with influenza ER admissions.

Step 1 - Create Spatial Weights

ER admissions are geographically influenced (Peköz et al., 2003; Wennberg, 1979). Many hospitals operate under a 'catchment' type of system, where their services influence patient arrivals by geography, hospital size (number of beds, resources, and other measures), and proximity to competing hospitals (Delamater, 2013). Large hospitals near patients but far from competing hospitals are likely to experience more patients in comparison to nearby hospitals that compete for patients. In the latter example, people will likely seek medical attention at the closest hospital. The hospital data provided by New York State’s Department of Health SPARCS includes data fields of the patient's ZIP Code of residence and name of the ER facility the patient entered. Given these two data fields, the spatial distribution of where influenza patients reside can be calculated for each hospital.

The proportion of influenza patients in each ZIP Code by hospital indicates the catchment area respective to each hospital. This is calculated based on ranking the distance of influenza patients home ZIP Code centroid to the hospital of interest. This was done for every hospital in NYC. The cumulative sum of the proportion of influenza ER admissions is then calculated based on the ranked distance of ZIP Codes for each hospital. This proportion of influenza patients in each ZIP Code for each hospital represents the spatial weights for that hospital, which is applied in the below Step two.

This research also determines if influenza tweets are spatially concentrated near a hospital. Similar to the above approach, this identifies the proportion of influenza tweets in each ZIP Code of NYC. Then this research ranked all ZIP Codes based on the distance of their centroid to each hospital. Cumulative sum of the proportion of influenza tweets is then calculated based on the ranked distance of ZIP Codes for each hospital.

Step 2 - Identify Lag and Lead Values for Each Hospital

Similar to Part 2, this step uses Pearson's correlation coefficient to identify when influenza tweets best correlate with influenza ER admissions for each hospital. Where X equals all influenza ER admissions at a single NYC hospital, Y equals daily influenza tweets as a proportion of daily influenza activity (Lamb et al., 2013) and spatially weighted by the ZIP Code proportion of influenza cases for a single hospital, and r equals the correlation value between 0 and 1. The below formula is executed on a hospital by hospital basis. The spatial weights identified in the above step are applied to influenza tweets. This spatially adjusts influenza tweets to be aligned with the spatial catchment area of influenza ER admissions and consequently a more robust temporal comparison can be made between these two datasets. (see results section).

$$r = \frac{\Sigma \mathrm{xy} - \frac{\Sigma \mathrm{x} \Sigma \mathrm{y}}{\mathrm{N}}}{\sqrt{(\Sigma x^2 - \frac{(\Sigma x)^2}{N})} + \sqrt{(\Sigma y^2 - \frac{(\Sigma \mathrm{y})^2}{N})}}$$

Step 3 - Identify Lag and Lead Values for Each ZIP Code

This research used inverse distance weighted (IDW) method to spatially interpolate lag and lead values between the location of hospitals. Here, $\hat{V}$ is the value to be estimated. V_i is the known lag or lead value at a hospital. $d_i^o \ldots . d_n^o$ is the distances from the hospitals to the power of p of the point estimate.

$$\hat{V} = \frac{\sum_{i=1}^{n} \frac{1}{d_i^p} V_i}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

The distance to which IDW operates will dictate the interpolation values within each ZIP Code. A distance based approach was used that roughly mimics the spatial clustering of influenza ER admissions near a hospital.

Part 4: Identify Variable that Influenza Lag and Lead of Influenza Tweets Predicting Influenza ER Admissions

This research used multiple regression methods to identify and understand variables driving the lag and lead values of ZIP Codes in NYC. Data from the 2015 Census including: race, age, education, median household income, and travel behavior data in addition to influenza ER admissions, influenza tweet frequencies, and tweet frequencies were examined as predictors for determining a ZIP Code's lag or lead value. All independent variables were aggregated at the ZIP Code level (N=229) and normalized by ZIP Code population. This research also examined the possibility of any spatial relationships amongst the independent variables associated with the dependent variable but found no relationship.

Three regression models were fitted where Y equals the lag or lead value in each ZIP Code, X equals the independent variables normalized by population, and b equals the parameter estimates of each variable.

$$Y = a + b_1 + X_1 + b_2 + X_2 \dots + b_p + X_p$$

All of the following models presented had fairly normal distributions, independence among the error terms, and possessed little multicollinearity.

Results

Part 1: Tweet Data Cleaning

Step 1: Data Cleaning - Removing Retweets

Of the 14 million tweets captured, nearly 400,000 retweets were removed (about 2.8%).

Step 2: Data Cleaning - Identifying tweets with influenza keywords

Once retweets were removed, 439,796 tweets were identified as containing influenza keywords of flu, sore throat, cough, and headache (Aron Culotta, 2010).

Step 3: Data Cleaning - Identifying influenza tweets based on tweet context

Step 3 involves using MLAs to classify influenza keyword tweets into three categories: self-diagnosed influenza tweets (primary tweets), non-influenza tweets, and PSA tweets. The ensemble of MLA models produced a classification accuracy of about 93% (95% C.I. - 92.5% to 93.7%). Table 3.4 summarizes the cross-validation classification of the ensemble of MLAs.

Table 3.4: Cross-Validation Summary of Ensemble MLAs

	Primary Flu Tweets (N=1286)	**Non-Influenza Tweets (N=802)**	**PSA Tweets (N=103)**
Sensitivity	93.4%	92.8%	90.5%
Specificity	93.5%	99.2%	96.1%
Positive Predictive Value	95.7%	98.6%	21.1%
Negative Predictive Value	90.4%	95.7%	99.8%

For all three influenza categories the sensitivity, specificity, and negative predictive values are high. Sensitivity refers to how well the MLAs correctly classify the tweet when it is positive (true positive rate). Specificity refers to how well the MLAs correctly classify the tweet when it is negative (true negative rate). For example, the sensitivity in Figure 3.3 indicates that on average the ensemble of MLAs correctly classified about 93% of the primary flu tweets. Positive and negative predictive values reference the probability that a tweet is positive when the MLAs indicate positive, and the probability that the tweet is negative when the MLAs indicate negative (Bradley, 1997). The positive predictive value, for example, would state that there is about a 93% probability that a tweet is a primary flu case when MLAs say it is a primary flu case. These performance metrics are widely used in MLAs with predictive classification.

Positive predictive values are high for all influenza categories except for public service announcement influenza tweets. However, this is expected as the training sample size for PSA influenza tweets is smaller than the sample size of the other two influenza categories. This is calculated according to the proportion of PSA tweets versus all positive outcomes. It is also likely that the MLAs had a difficult time distinguishing between self-diagnosed influenza tweets and influenza public service announcement tweets because the content of these two tweet categories are similar.

It should be noted that MLAs' accuracy is dependent on the uniqueness of the categories in the training dataset. The MLAs used above had a high degree of accuracy, suggesting that it can successfully distinguish the difference between all three flu cases. However, there are tweets that are misclassified by the ensemble of MLAs. Therefore, the presence of unwanted false-positives will be found. Likewise, the ensemble of MLAs will discard some influenza related tweets (false-negatives). The inclusion of false-positives and exclusion of false-negatives may skew the below research results. However, given the high accuracy of the ensemble of MLAs as a classifier, it is unlikely that any misclassified data will significantly change the results presented in this paper.

About 18,000 tweets were identified as self-diagnosed influenza cases using the above training dataset and MLAs.

Step 4: Data Cleaning - Syndrome Elapse Time

Of the 18,000 tweets identified as self-diagnosed influenza tweets, roughly 8,700 tweets were removed as users tweeted about their symptoms multiple times in a two-week window. Therefore, about 9,300 tweets were unique cases of influenza.

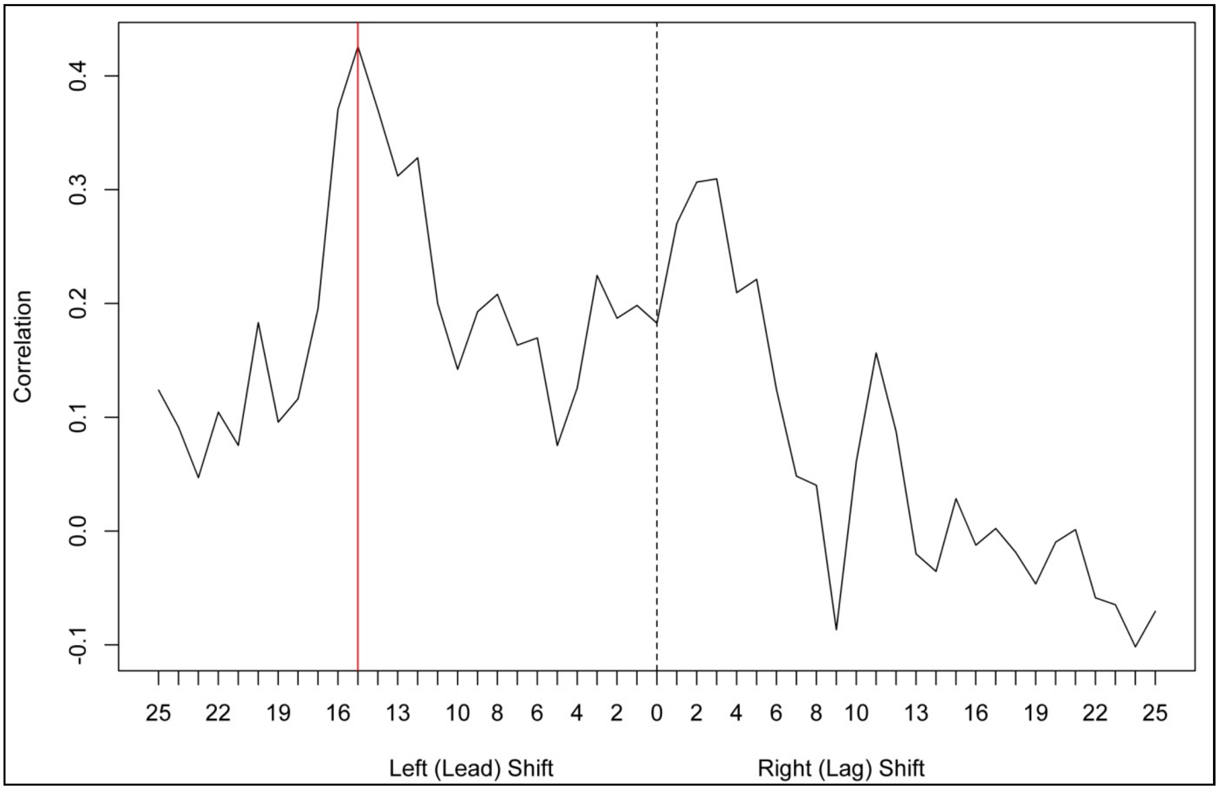
Step 5: Data Cleaning - Remove tweets outside of study area

The last step is to ensure that single episodes of self-diagnosed influenza cases were occurring in NYC. Any tweet that did not contain GPS enabled coordinates or was posted outside of NYC was removed. This resulted in 2,447 tweets that contain influenza keywords, are self-diagnosed influenza cases, and were posted in NYC.

Part 2: Optimally Identifying when Influenza Tweets Predict ER Admissions in NYC

Step two identified the optimal time to when influenza tweets are best correlated with influenza ER admissions in NYC. Figure 3.1 displays the lag or lead days influenza tweets predict influenza ER admissions in all of NYC and each day's correlation value up to 25 days.

Figure 3.1: Correlations of Influenza Tweets Lags and Leads on ER Hospitalizations



The vertical dashed line at zero separates the lead and lag shifts of influenza tweets where each x-axis value represents the number of days influenza tweets lag or lead influenza ER admissions. The y-axis represents Pearson's correlation coefficient value. The correlation of influenza tweets for each shift is calculated against temporally stationary influenza ER admissions. Figure 3.1 indicates that influenza ER admissions are best correlated with influenza tweets when influenza tweets are temporally shifted to the left suggesting that influenza tweets succeed influenza ER admissions. The highest correlation value is over 0.40, which occurred at a lead value of 15 days. This indicates that influenza tweets precede influenza ER admissions by 15 days.

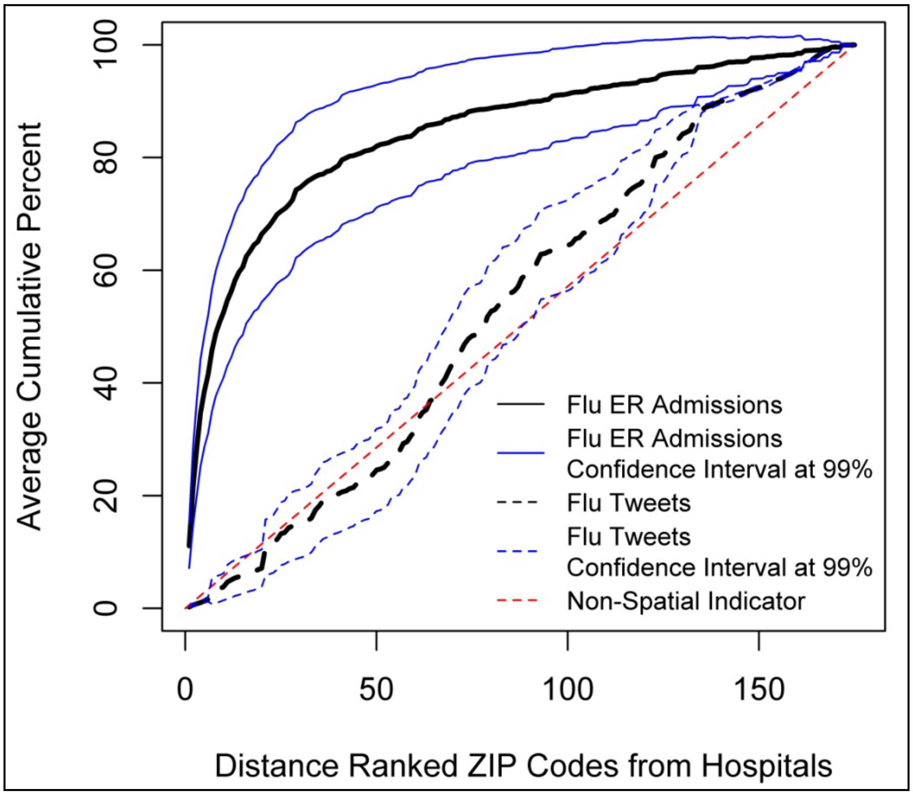
Part 3: Identifying when Influenza Tweets Optimally Predict Influenza ER Admissions on a Hospital by Hospital Basis

Step 1 - Create Spatial Weights

Figure 3.2 illustrates the average cumulative percentage of the proportion of influenza ER admissions and influenza tweets found in distance ranked ZIP Codes from NYC hospitals. The red dashed line represents a theoretical pattern where the average cumulative percentage of influenza ER admissions and influenza tweets are proportionally even in all NYC ZIP Codes. Lines above the red dashed line represent spatial clustering around the hospital since higher proportion of cases are occurring in shorter distances from the hospital. Lines below the red dashed line represent spatial clustering far from the hospital. In Figure 3.2, the black solid line represents the average cumulative percentage rate of influenza ER admissions. On average, nearly 75% of influenza ER admissions occur within 31 out of 229 ZIP Codes surrounding the hospital. This suggests that influenza ER admissions are spatially clustered around hospitals, considering the remaining 25% of influenza ER admissions occur in the next 32 through 120 distance ranked ZIP Codes. The solid blue lines represent confidence intervals at 99%. Confidence intervals were calculated because not all hospitals mimic this spatial relationship as some hospitals' influenza patients do not come from adjacent ZIP Codes.

Figure 3.2 also displays the average cumulative percentage of the proportion of influenza tweets in distance ranked ZIP Codes from the hospital, as shown by the dashed black line. This line follows a similar shape and slope to the red dashed line, suggesting that influenza tweets, on average, are not spatially clustered around hospitals. On average, 75% of influenza cases occur within the closest 118 ZIP Codes of a hospital in comparison to 75% of influenza ER admissions occurring within the closest 31 ZIP Codes of a hospital. Confidence intervals at 99% were also calculated for influenza tweets, as some influenza tweets may be more spatially clustered around certain hospitals.

Figure 3.2: Average Cumulative Percentage of Influenza ER Admissions and Influenza Tweets by Distance Ranked ZIP Codes from all Hospitals in NYC with 99% Confidence Intervals



Step 2 - Identify Lag and Lead Values for Each Hospital

Table 3.5, shows the optimal lag and lead values for each hospital in NYC ordered by when influenza tweets correlate with influenza ER admissions.

Table 3.5: Optimal Lag and Lead Values for Emergency Room Hospitals in New York City

Correlation	Days	Lag or Lead	Hospital
0.13	24	Lag	Mount Sinai Hospital - Mount Sinai Hospital of Queens
0.25	23	Lag	Metropolitan Hospital Center
0.15	23	Lag	St. Johns Episcopal Hospital So Shore
0.12	22	Lag	SBH Health System
0.15	20	Lag	Mount Sinai Beth Israel Brooklyn
0.24	20	Lag	Mount Sinai Brooklyn
0.29	18	Lag	New York Community Hospital of Brooklyn, Inc
0.25	18	Lag	University Hospital of Brooklyn
0.13	10	Lag	Lincoln Medical & Mental Health Center
0.24	10	Lag	North Central Bronx Hospital
0.22	9	Lag	Forest Hills Hospital
0.23	9	Lag	Jacobi Medical Center
0.27	9	Lag	Montefiore Medical Center-Wakefield Hospital
0.19	9	Lag	Mount Sinai Hospital
0.24	8	Lag	New York-Presbyterian/Lawrence Hospital
0.21	8	Lag	New York Presbyterian Hospital - Columbia Presbyterian Center
0.23	7	Lag	Staten Island University Hosp-North
0.26	5	Lag	Lenox Hill Hospital
0.21	3	Lag	New York Presbyterian Hospital - New York Weill Cornell Center
0.15	3	Lag	Queens Hospital Center
0.25	1	Lag	Bronx-Lebanon Hospital Center - Concourse Division
0.22	1	Lag	Brookdale Hospital Medical Center
0.24	1	Lag	Brooklyn Hospital Center - Downtown Campus
0.28	1	Lag	Coney Island Hospital
0.28	1	Lag	Elmhurst Hospital Center
0.20	1	Lag	Jamaica Hospital Medical Center
0.21	1	Lag	New York Hospital Medical Center of Queens
0.24	1	Lag	New York Methodist Hospital
0.34	1	Lag	New York Presbyterian Hospital - Allen Hospital
0.26	1	Lag	Richmond University Medical Center
0.25	1	Lag	Staten Island University Hosp-South
0.23	0	--	Lutheran Medical Center
0.16	2	Lead	Kings County Hospital Center

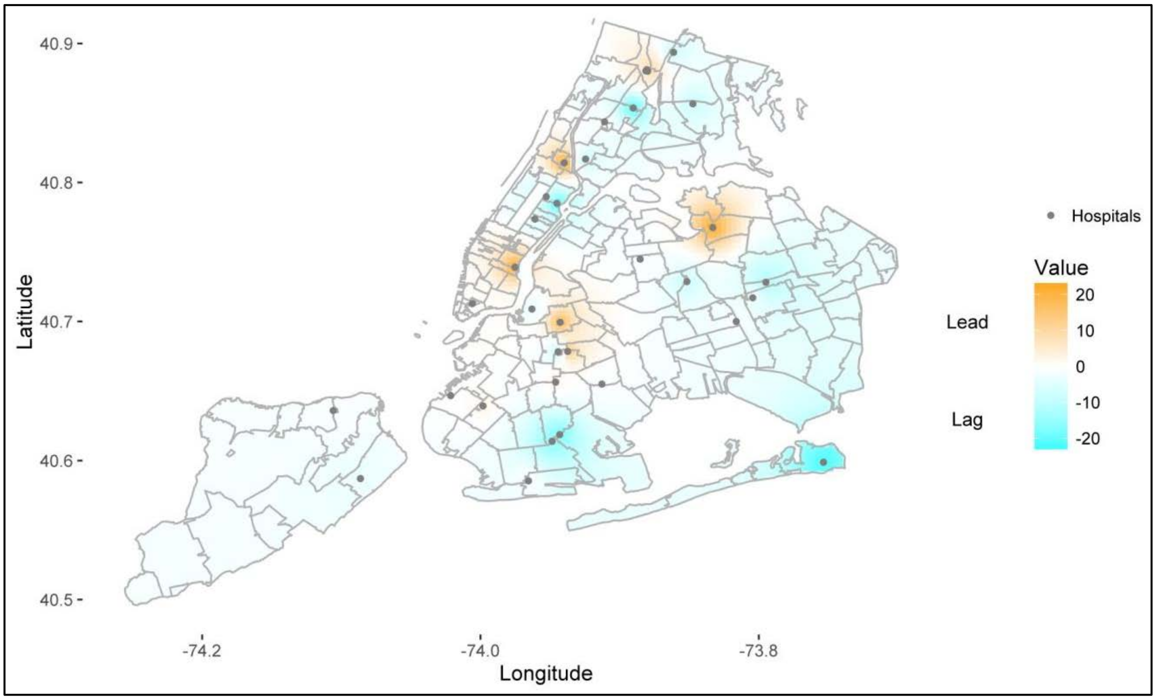
Table 3.5 (cont'd).

0.22	2	Lead	Maimonides Medical Center
0.30	12	Lead	Mount Sinai Beth Israel
0.11	13	Lead	Montefiore Medical Center - Henry & Lucy Moses Division
0.27	20	Lead	Kingsbrook Jewish Medical Center
0.33	20	Lead	Montefiore Med Center - Jack D Weiler Hosp of A Einstein College Division
0.19	21	Lead	Woodhull Medical & Mental Health Center
0.19	24	Lead	Bellevue Hospital Center
0.18	24	Lead	Flushing Hospital Medical Center
0.25	24	Lead	Harlem Hospital Center
0.27	24	Lead	New York-Presbyterian/Lower Manhattan Hospital

Table 3.5 shows when influenza tweets optimally correlate with influenza ER admissions. For example, the lag value for Mount Sinai Hospital of Queens (row 1) is 24 days. This means the highest correlation between influenza tweets and influenza ER admissions at this hospital occurred when influenza tweets were posted 24 days before ER admissions occurred. Influenza tweets correlate with influenza ER admissions at 74 of hospitals when influenza tweets are lagged one to 24 days (mean ~ 8.5 days). Conversely, influenza tweets correlate with influenza ER admissions at 25% of NYC hospitals when influenza tweets are leaded two to 24 days (mean ~ 17 days). All correlation values presented in Table 3.5 are relatively low.

Next, inverse distance weighted (IDW) function was applied to interpolate lag and lead values throughout NYC's ZIP codes. Figure 3.3 is the result of the IDW function applied to the lag and lead values of hospitals in NYC. Roughly 82% of IDW values represent lags (18% represent leads). The lag and lead values interpolated for each ZIP Code are used in the regression analysis below.

Figure 3.3: Spatial Representation of Lag and Lead Values for New York City



Part 4: Identify Variables that Influence Lag and Lead values of Influenza Tweets Predicting Influenza ER Admissions

Regression Model 1: Demographic Composition

This research used regression models to identify factors that influence influenza tweets temporally correlating with influenza ER admissions. First, this research regressed demographic, age, education, and median household income (independent) variables onto lag and lead (dependent) values for NYC. This model identifies key variables about the composition of people in NYC ZIP Codes and their influence on lag and lead values. Age, age by gender, race, and education attainment were used as independent variables because tweeting activity tends to be from young, affluent, and non-rural (“Social Media Update,” 2016) populations but influenza ER admissions are skewed toward populations that are either very young or old. Table 3.6 shows the output of this regression function.

Table 3.6: Results of Regression Model 1: Demographic Composition

Variable per 1,000 people	Estimates (CI at 99%)	Standard Error	T-Value	P-Value
Intercept	2.90 (0.66 to 5.14)	0.86	3.37	0.000890
Race: White	-0.0020 (-0.006 to 0.002)	0.002	-1.29	0.199
Race: Other	-0.006 (-0.015 to 0.003)	0.003	-1.82	0.07
Associates Degree	-0.05 (-0.11 to 0.01)	0.02	-2.3	0.02
Doctorate Degree	- 0.075 (-0.165 to 0.0143)	0.03	-2.18	0.03
Persons less than 18	-0.022 (-0.038 to -0.007)	0.006	-3.71	0.0002
Persons 18 to 39	0.014 (0.005 to 0.022)	0.003	4.19	0.00004

Adjusted R2: 0.19; Spatial Correlation of Residuals: 0.02; AIC: 1399

In Table 3.6, only four variables have significant influence (p-value at 0.05) at predicting the lag and lead values of when influenza tweets best coincide with influenza ER admissions. Examining the estimates of each variable sheds light onto that variable's relationship with the lag and lead values. Positive estimates indicate a decrease in time when influenza tweets best correlate with influenza ER admission, and vice versa. Examining the 'Persons less than 18' variable's estimate, for instance, indicates that an increase of one 'Person less than 18' per 1000 people will decrease the time when influenza tweets best correlate with influenza ER admissions by about 0.022 days. Therefore, ZIP Codes that have populations less than 18, have advanced education, and are composed of minority groups will increase the amount of time influenza tweets best correlate with influenza ER admissions. White populations were not significant in the model, this could be because Twitter is disproportionately black (“Social Media Update,” 2016). Persons 18 to 39 decrease the amount of time influenza tweets can best predict influenza ER admissions. The Persons 18 to 39 variable has the opposite relationship when compared to the

Persons less than 18. This could be a result of the young populations (less than 18) disproportionately tweeting more and older populations (18 to 39) disproportionately tweeting less (Pew Research Center, 2016). This model was able to explain roughly 20% of the variance of what drives lag and lead values in NYC ZIP Codes. The model's residuals were not spatially autocorrelated.

Regression Model 2: Transportation behavior

Second, this research examined travel behavior variables that determine when influenza tweets best correlate with influenza ER admissions. Travel characteristics were chosen as variables because influenza ER admissions are spatially clustered near the hospital. The mode of transportation may influence the lag and lead values since it represents people’s ability to physically access the hospital. Table 3.7 shows the output of several transportation factors role in predicting lag and lead values.

Table 3.7: Results of Regression Model 2: Transportation behavior

Variable per 1,000 people	Estimates (CI at 99%)	Standard Error	T-Value	P-Value
Intercept	0.5 (-1.2 to 2.2)	0.66	0.76	0.45
Drive Alone: Car, Truck, or Van	-0.02 (-0.03 to -0.01)	0.0044	-4.8	0.000002
Public Trans: Excluding Taxicab and Railroad	-0.003 (-0.066 to 0.061)	0.025	-0.11	0.91
Motorcycle	1.11 (-0.6 to 2.82)	0.66	1.68	0.09
Walked	0.0075 (-0.0091 to 0.024)	0.0064	1.17	0.24

Adjusted R2: 0.156; Spatial Correlation of Residuals: -0.005; AIC: 1410

The transportation model, similar to the demographic composition model, explains less than 20% of the variance that causes the temporal variation of influenza tweets on influenza ER admissions in NYC ZIP Codes. ZIP Codes with higher rates of people that drive alone to work tend to increase the amount of time by which influenza tweets can predict influenza ER admissions, as indicated by negative estimates. ZIP Codes with higher populations that drive a motorcycle suggest that influenza ER admissions increasingly precede influenza tweets, as indicated by positive estimates. These two variables showed statistical significance at the 0.05 level.

Regression Model 3: Combined Demographic and Transportation

Next, this research combined the demographic and transportation models but also included influenza ER admissions, influenza tweets, and overall Twitter activity as variables for influencing when influenza tweets best predict influenza ER admissions. These three variables were also normalized based on ZIP Code populations but did not add any explanatory power to the model and hence were not included in the final model. As indicated in Table 3.8, the combined model has 10-15% more explanatory power than previous models as indicated by an adjusted R-squared value of 0.28. Furthermore, the AIC values for the combined model are (marginally) lower than the two previous models. This suggests that the final model is a better fit for categorizing which factors influence when influenza tweets best correlate with influenza ER admission. This last model suggests that demographic and transportations variables are important.

Table 3.8: Results of Regression Model 3: Combined Demographic and Transportation

Variable per 1,000 people	Estimates (CI at 99%)	Standard Error	T-Value	P-Value
Intercept	3.21 (1.09 to 5.32)	0.81	3.94	0.00011
Total Housing Units	0.013 (0.003 to 0.023)	0.004	3.29	0.0012
Drive Alone: Car, Truck, or Van	-0.014 (-0.025 to -0.002)	0.004	-3.03	0.003
Motorcycle	1.46 (-0.15 to 3.06)	0.62	2.36	0.02
1st Graders	-1.27 (-2.55 to 0.018)	0.5	-2.56	0.01
Masters Degree	-0.033 (-0.06 to -0.006)	0.01	-3.12	0.002
Male Persons Greater than 64	-0.031 (-0.059 to -0.003)	0.011	-2.86	0.005
Female Persons less than18	-0.05(-0.08 to -0.022)	0.011	-4.43	0.000015
Race: Asian	0.0051 (-0.0015 to 0.0117)	0.003	2	0.047

Adjusted R2: 0.28; Spatial Correlation of Residuals: -0.005; AIC: 1372

Combining the demographic and transportation models lead to all variables, as shown in Table 3.8, to be significant at the 0.05 level. This model suggests that ZIP Codes with a population increase in males greater than 64, females less than 18, Masters degrees, 1st graders, and drive alone increase the time influenza tweets predict influenza ER admissions. When total housing units, Asian populations, and the number of people driving motorcycle to work increases, the amount of time influenza tweets predict influenza ER admissions decreases.

Discussion

Traditional approaches to influenza surveillance, such as the CDC, provide little temporal and spatial information to inform hospitals about nearby influenza cases. Likewise, current novel digital techniques to monitor influenza activity, such as GFT provide little benefit to individual hospitals given the lack of geographic detail. To fill this gap, this research uses Twitter as a potential real-time data source to identify the optimal time when influenza tweets best correlate with actual ER influenza admissions at local hospitals in NYC. This geographic scale of analysis is strikingly different and more beneficial than traditional approaches that analyze influenza activity at broad geographic scales (e.g., state or metropolitan) that provide little knowledge about local influenza activity to ERs. This research demonstrated that influenza tweets are correlated with influenza ER admissions on a hospital-by-hospital basis. Table 3.4, lists the optimal number days influenza tweets best correlate with each NYC hospital. Therefore, the novelty of this research may allow health officials a more timely glimpse into local influenza activity so they can prepare their staff and facilities for a possible surge of influenza patients.

Research shows that influenza tweets can predict influenza ER admissions up to 14 days in advance at national and metropolitan scales (Broniatowski et al., 2013; Paul et al., 2014; Signorini, Segre, & Polgreen, 2011). As Table 3.4 shows, the optimal correlation between influenza ER admissions and influenza tweets varies by hospital. For 31 hospitals (~75%), influenza tweets best correlate with influenza ER admissions 14 days (~mean of 8.5 days) in advanced. According to previous research, this range is expected and suggests that on average influenza tweets originate roughly a week before ER admissions occur. Influenza tweets correlate with influenza ER admissions at eight hospitals more than 14 days in advance. Six of these eight hospitals have influenza tweets correlating with influenza ER admissions about three weeks in advance. This is longer than previous research indicates. Through it remains untested, influenza propagation may explain this relationship. Influenza infection commonly starts in young school age populations then is transmitted to older populations (Hsieh, 2010). Tweets are commonly generated by younger populations (Murthy et al., 2016) who are usually the first infected with influenza. However, in 11 hospitals, influenza ER admissions correlate with influenza tweets on average of about 17 days in advance. This is suggesting that ER admissions are occurring before influenza tweets are posted. In this case, it could mean that the very young (under five years of age) are entering a hospital's ER considering they are often the first age group that is infected. From this point, the young child infects a person who frequently tweets. These active Twitter users tweet about their influenza symptoms on average 17 days after the young child's ER admission. Again though, this remains untested. These results are only applicable to these hospitals. This research found no generalized connection with lag and lead values and hospital characteristics such as, number of beds and estimated income.

Some of the differences between what drives the correlation between influenza tweets and ER admissions and vice versa rests on demographic, transportation, and economic variables as shown in the above multiple regression models. Demographic variables are important to consider because it is people who tweet and people who seek medical care. But not all people frequently tweet or seek medical care. Young, educated, non-rural, and affluent populations tend to do the majority of tweeting ("Social Media Update," 2016) while the very young and very old often seek medical care for their influenza symptoms (Hsieh, 2010). The demographic regression model highlighted this behavior as the coefficient parameter for young populations (18-39) tweeting about their influenza symptoms increased the amount of time influenza tweets can best predict influenza ER admissions. Populations 18 to 39 had the opposite effect on lag and lead values.

The final regression model presented in this research contains demographic and transportation variables. Demographic variables may represent a temporal propagation effect between the different age groups. There is an opposing relationship with age's ability to increase the time when influenza tweets can predict influenza ER admissions. This may indicate that lag and lead values for hospitals are a product of influenza propagation amongst the different age groups. ZIP Codes with younger female populations are going to increase the amount of time influenza tweets can predict influenza ER admissions, whereas older males decrease the amount of time. This may indicate that young females are tweeting about influenza symptoms before actual influenza ER admissions are occurring, indicating that people in these ZIP Codes become infected with influenza before influenza ER admissions occur. ZIP Codes with older male populations tend to have less time to when influenza tweets predict influenza ER admissions. This is not surprising considering initial influenza infection occurs in young children (Earn et al., 2012). While the regression results may indicate a demographic propagation affect influencing the lag and lead values of influenza tweets correlating with influenza ER admissions, it remains largely untested. Future research needs to be conducted that robustly tests this possibility.

As shown above, influenza ER admissions are spatially clustered near a hospital. Therefore, transportation variables were examined as predictors to lag and lead values because these variables represent the spatial mobility of users to access a hospital. The final model indicates the combination of demographic and transportation variables increase the model’s explanatory power compared to models that solely use demographic or transportation variables. Therefore, people's mode of transportation influences the lag and lead values found in each ZIP Code. The combined model only identified two significant transportation variables, 'populations that drive alone' and those that 'drive a motorcycle to work'. These variables have opposing relationships, such as the aforementioned young female's and old male's variables. ZIP Codes with higher proportions of people that drive a car alone to work tend to increase the time when influenza tweets can predict influenza ER admissions. However, ZIP Codes where people prefer to ride a motorcycle to work decrease the amount of time between influenza tweets and influenza ER admissions. One possible explanation to this opposing relationship is based on seasonality and transportation flexibility. Motorcycles are not typically used during cold winter months in NYC when influenza season peaks. During these colder periods motorcycle users are likely to use alternative transportation such as bus or subway lessening their transportation flexibility. Having less transportation flexibility during the winter months may influence motorcycle drivers to seek medical care more quickly than those with more flexible transportation options, such as a car.

The adjusted R-squared values in all the regression models are somewhat small indicating that there are unexplored variables that explain lag and lead values. Some of these unexplored variables could include average family size, family structure/network, type of employment, health insurance status, influenza vaccination rates, preexisting medical conditions, and other demographic and public health factors such as, propensity to washing hands, number of contacts, and exposure to sick populations. As with the above model, this model's findings need to be further explored with more robust research.

Several hospital variables did not influence lag and lead values but this research had limited hospital variables to test. Future research needs to test more hospital variables that influence lag and lead values. Identifying key hospital and social variables will create a more robust approach to understanding what influences lag and lead values without having to rely on comparing influenza tweets to influenza ER admissions.

Some of the differences presented in the lag and lead values of influenza tweets correlating with influenza ER admissions and the directionality of coefficients may be an indication of influenza propagation. However, this theory not directly tested in this research. Future research should explore this relationship in more detail. A possible approach would be to use Vertalka’s (2017) DIP approach but include questionnaire elements that ask about possible contagion effects.

Conclusion

The intention of this research was to examine when influenza related tweets best correlate with influenza ER admissions at specific hospitals in NYC. Prior research has demonstrated the utility of influenza tweets at predicting influenza infections up to 14 days in advance, but at broad geographical scales. This research introduces finer geographic scope of where and to what degree influenza tweets are associated with influenza rates at individual hospitals through four parts. Step one identifies Twitter users that are likely experiencing influenza-like illness through influenza keywords and by using MLA to distinguish non-influenza, self-diagnosed influenza, and public service announcement influenza tweets. Step two identifies the spatial dispersion of where influenza ER admissions and influenza tweets occur. Step three uses Pearson's correlation coefficient to identify the optimal lag or lead time when spatially weighted influenza tweets correlate with influenza ER admissions. Step four uses inverse distance weighted to interpolate the lag and lead values between hospital locations. Finally, Step fives uses multiple regression models to classify demographic, economic, education, and transportation variables that can predict the lag and lead values NYC ZIP Codes.

Using these steps, this research demonstrated, influenza tweets correlate with influenza ER admissions for over 75% of the hospitals in NYC. The average time influenza tweets are associated with influenza ER admissions is about 8.5 days. This figure aligns with other researchers' using influenza tweets as indicators for influenza cases at broader geographic scales. This paper also shows that in some cases hospitals experience influenza ER admissions before influenza tweets occur. Surveillance systems can use the above approach to monitor and characterize influenza activity for local hospitals. No prior research has examined how influenza tweets are associated with influenza ER admissions at local hospital scales. Examining this relationship is important as Twitter has the potential to be used as an additional surveillance data source that helps warn hospitals when local influenza activity is increasing.

The findings from this research suggest that there is potential for Twitter to act as a proxy for influenza ER admissions. As of right now, Twitter data should not be used to replace current influenza surveillance approaches. While current influenza approaches are outdated and provide little benefit to hospital systems, they still act as a source of truth for influenza activity. This research only tested the feasibility of Twitter as a proxy for influenza ER admissions in one city and for a single influenza season. Therefore, a generalized and well-built theory of Twitter's potential to be a proxy for influenza ER admissions remains largely unexplored.

While it remains untested, the potential for Twitter to act as a surveillance system that monitors the propagation of influenza activity may shed light onto the severity of influenza spread. However, understanding how influenza is shifting to and from different age groups by using Twitter is not feasible without comparing influenza tweets to influenza ER admission. From the comparison of these two datasets, it can be determined whether influenza tweets precede or succeed influenza hospital ER admissions. Not surprisingly, several hospitals could benefit from using Twitter as a data source to identify more localized influenza activity.

Future research needs to focus on several aspects. First, future research needs to use this approach to identify when influenza tweets best predict influenza ER admissions at hospitals in different cities. Most cities will possess similar demographic trends when it comes to tweeting activity and influenza ER admissions. Therefore, it is expected, but unknown, as to what demographic, economic, and transportation factors will be significant at predicting lag and lead values in cities other than NYC. While this research found no spatial influence for demographic, economic, and transportation factors predicting lag and lead values, changing the study site to another city might introduce a significant spatial component. Furthermore, future research needs to identify hospital characteristics that influence the lag and lead values of influenza tweets predicting influenza ER admissions. Doing so will give an indication as to whether the lag and lead values are strongly associated with a particular feature of a hospital. Second, this research predicted influenza ER admissions at hospitals during seasonal influenza. Future research needs to focus on predicting influenza ER admissions with influenza tweets outside of seasonal influenza and during pandemic influenza. Third, future research needs to address Twitter's ability to monitor and predict influenza propagation between different age groups. While this research did not directly test this relationship, there are findings presented in this research that may suggest such a relationship exists.

APPENDIX

Table A 3.1: Influenza-Like Illness ICD-9 Codes

ICD-9 code	Description
79.89	Viral infection NEC*
79.99	Viral infection NOS*
460	Nasopharyngitis, acute
462	Pharyngitis, acute
464	Laryngitis, acute, without obstruction
464.1	Tracheitis, acute, without obstruction
464.2	Laryngotracheitis, acute without obstruction
465	Laryngopharyngitis, acute
465.8	Infectious upper respiratory, multiple sites, acute NEC
465.9	Infectious upper respiratory, multiple sites, acute NOS
466	Bronchitis, acute
466.11	Bronchiolitis due to respiratory syncytial virus
466.19	Bronchiolitis, acute, due to other infectious organism
478.9	Disease, upper respiratory NEC/NOS
480	Pneumonia due to adenovirus
480.1	Pneumonia due to respiratory syncytial virus
480.2	Pneumonia due to parainfluenza
480.8	Pneumonia due to virus NEC
480.9	Viral pneumonia unspecified

CONCLUSION

This dissertation contains three chapters. The first chapter introduces and discusses a novel approach to engage social media users to participate in scientific research. Traditionally, scholars have viewed social media users as data sources that were unknowingly contributing to scientific discovers. This approach is apparent in research that highlights social media's ability to predict stock market trends and earthquakes. It is also evident in cases were social media data is used to gather situational awareness during disasters, such as the Boston Marathon Bombing. Chapter one focuses on building a more intimate relationship with social media users so that they become more than just social sensors cataloging the world around them. Instead, Chapter one discusses an approach so that social media users become sensors that are more involved in the scientific research process. This is accomplished by using Application Program Interfaces as a conduit to send messages to the social media user. These messages act a form of digital interaction where the research and social media user can engage in relevant scientific inquiry or discussions. In the case of this research, the digital interaction involved sending users a link to an online questionnaire which was designed to gather insight about their demographic and state of health. Completed questionnaires contained data elements that would otherwise be unobtainable through social media APIs. Therefore, research projects employing the DIP system can gain additional information about their social media subjects and the information they produce on social media.

Chapter two of this dissertation focuses on correcting temporal and spatial shortcomings of digitally detecting influenza cases through Twitter. Current research that digitally detects influenza cases through Twitter makes two assumptions. First, it assumes that influenza related tweets are posted when the author first experiences symptoms. Second, it assumes that influenza tweets are posted at the author's home. Chapter two uses the DIP approach, introduced and discussed in Chapter one, to gain additional temporal and spatial information about London and

New York influenza-ridden Twitter users. Chapter two of this dissertation found that influenza tweets are most often posted when the author experiences peak symptoms. This might be due to the author’s frustration with the flu or boredom as they recover at home without entertainment. Chapter two also found that influenza infected Tweets were posted closer to the author's home. This might be a consequent of the author experiencing a severe level of influenza symptoms causing them to be confined to the house or even bedridden. The fact that influenza tweets are occurring near a user’s home ZIP or Postal Code is suggestive that digitally detecting influenza cases can occur at local neighborhood scales.

Chapter three explores the possibility of using Twitter to detect influenza cases at a finer geographic scale. Research has typically focused on correlating influenza tweets with influenza cases at broad scales such as, Country, Regional, or Metropolitan levels. As identified in Chapter two, influenza tweets occur near the authors declared home ZIP or Postal Code. This suggests that an influenza tweet is spatially representative of an actual influenza case. Therefore, Chapter three focuses on predicting influenza ER admissions at individual hospitals in New York City. Conducting this analysis at the hospital scale provides medical professionals with timely information to prepare their facilities and staff for the potential of surge of influenza patients. However, not all hospitals are fortunate to have this insight as several hospitals’ influenza ER admissions proceed the occurrence of influenza tweets. Future research needs to examine what hospital factors influence the effectiveness of influenza tweets at predicting influenza ER admissions. For instance, do the number of beds in the ER influence the number of nearby influenza tweets? Furthermore, Chapter three introduces a possible theory of demographic propagation that might be influencing the different lag and lead values of influenza tweets correlating with influenza ER admissions. However, this remains largely untested. Future research needs to address this possible theory in more detail.

One of the main themes of this research is understanding the behavior of humans as it relates to their interaction with their cell phone and Twitter when experiencing influenza-like symptoms. The advent of the internet and mobile technology changed the way people interact with each other and their surroundings. Prior to this technological shift, people were confined to communicate through landline phones which also produced very limited data. Cell phones and social media outlets, on the other hand, produce terabytes to even petabytes of data every day. The insight that can be obtained from this vast amount of data is still largely unexplored and discussed among scholars and therefore, the possibilities of this data to build an understanding of human behavior remains largely unexplored. Future research needs to start exploring the possibilities of this data in helping solve not only undesired influenza ER surges but a host of other issues.

REFERENCES

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011). Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (pp. 702–707).

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B. B., Acuna-Soto, R., … Kohavi, R. (2011). Europe’s initial experience with pandemic (H1N1) 2009-mitigation and delaying policies and practices. *Annals of the Association of American Geographers*, *5*(1), 1. https://doi.org/10.1068/a130122p

Aramaki, E., Maskawa, S., & Morita, M. (2011). Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1568–1576).

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, *35*(4), 216–224.

Bayarma, A., Kitamura, R., & Susilo, Y. (2007). Recurrence of daily travel patterns: stochastic process approach to multiday travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, (2021), 55–63.

Bhat, C. R., & Misra, R. (1999). Discretionary activity time allocation of individuals between in-home and out-of-home and between weekdays and weekends. *Transportation*, *26*(2), 193–229.

Birkin, M., & Malleson, N. (2012). Investigating the Behaviour of Twitter Users to Construct an Individual-Level Model of Metropolitan Dynamics.

Boarnet, M. G., & Crane, R. (2001). *Travel by design: The influence of urban form on travel*. Oxford University Press on Demand.

Bodnar, T., & Salathé, M. (2013). Validating models for disease detection using twitter. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 699–702).

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

Boyle, J. R., Sparks, R. S., Keijzers, G. B., Crilly, J. L., Lind, J. F., & Ryan, L. M. (2011). Prediction and surveillance of influenza epidemics. *Medical Journal of Australia*, *194*(4), S28.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS One*, *8*(12), e83672.

Butler, D. (2013). When Google got flu wrong. *Nature*, *494*(7436), 155.

Center for Disease Control and Prevention. (2016). Retrieved October 17, 2016, from http://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm

Chandra, S., Khan, L., & Muhaya, F. Bin. (2011). Estimating twitter user location using social interactions--a content based approach. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on* (pp. 838–843). IEEE.

Chang, H., Lee, D., Eltaher, M., & Lee, J. (2012). @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)* (pp. 111–118). IEEE Computer Society.

Cheng, Z., Caverlee, J., & Lee, K. (2010). 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.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers*, *102*(6), 1267–1289.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating “big data”and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124–147.

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., … Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(5), 720–741.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing Twitter messages. *arXiv Preprint arXiv:1007.4748*.

Culotta, A. (2013). Detecting influenza outbreaks by analyzing Twitter messages. 2010. *arXiv Preprint arXiv:1007.4748*.

Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language Resources and Evaluation*, *47*(1), 217–238.

Daume, S., & Galaz, V. (2016). “Anyone Know What Species This Is?”–Twitter Conversations as Embryonic Citizen Science Communities. *PloS One*, *11*(3), e0151387.

Delamater, P. L. (2013). Spatial accessibility in suboptimally configured health care systems: A modified two-step floating catchment area (M2SFCA) metric. *Health & Place*, *24*, 30–43.

Denef, S., Bayerl, P. S., & Kaptein, N. A. (2013). Social media and the police: tweeting practices of british police forces during the August 2011 riots. In *proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3471–3480). ACM.

Derlet, R. W., Richards, J. R., & Kravitz, R. L. (2001). Frequent overcrowding in US emergency departments. *Academic Emergency Medicine*, *8*(2), 151–155.

Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., … Rothman, R. E. (2012). Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases*, *54*(4), 463–469.

Earn, D. J. D., He, D., Loeb, M. B., Fonseca, K., Lee, B. E., & Dushoff, J. (2012). Effects of school closure on incidence of pandemic influenza in Alberta, Canada. *Annals of Internal Medicine*, *156*(3), 173–181.

Elwood, S. (2008). Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal*, *72*(3), 133–135.

Ettema, D. (2005). Latent activities: Modeling the relationship between travel times and activity participation. *Transportation Research Record: Journal of the Transportation Research Board*, (1926), 171–180.

Feick, R., & Roche, S. (2013). Understanding the Value of VGI. In *Crowdsourcing geographic knowledge* (pp. 15–29). Springer.

Fiore, A. E., Shay, D. K., Broder, K., Iskander, J. K., Uyeki, T. M., Mootre, G., … Cox, N. J. (2008). Prevention and Control of Influenza Recommendations of the Advisory Committee on Immunization Practices (ACIP), 2008. Retrieved March 25, 2017, from https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5707a1.htm

Gao, H., Barbier, G., Goolsby, R., & Zeng, D. (2011). *Harnessing the crowdsourcing power of social media for disaster relief*. DTIC Document.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125. https://doi.org/10.1016/j.dss.2014.02.003

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, *1*(2009), 12.

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

Goodchild, M. (2009). NeoGeography and the nature of geographic expertise. *Journal of*

Location Based Services, *3*(2), 82–96.

Goodchild, M. F. (2007a). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Goodchild, M. F. (2007b). Citizens as sensors: web 2.0 and the volunteering of geographic information. *GeoFocus*, *7*, 8–10.

Goodchild, M. F. (2008). The use cases of digital earth. *International Journal of Digital Earth*, *1*(1), 31–42.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120.

Goranson, C., Thihalolipavan, S., & di Tada, N. (2013). VGI and public health: possibilities and pitfalls. In *Crowdsourcing geographic knowledge* (pp. 329–340). Springer.

Gross, D. (2012). Man faces fallout for spreading false Sandy reports on Twitter - CNN.com. Retrieved January 28, 2017, from http://www.cnn.com/2012/10/31/tech/social-media/sandy-twitter-hoax/

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 646–675.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Gupta, A., & Kumaraguru, P. (2012). Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media* (p. 2). ACM.

Hägerstraand, T. (1970). What about people in regional science? *Papers in Regional Science*, *24*(1), 7–24.

Haklay, M. (2013). Citizen science and volunteered geographic information: Overview and typology of participation. In *Crowdsourcing geographic knowledge* (pp. 105–122). Springer Netherlands.

Han, B., Cook, P., & Baldwin, T. (2013). A Stacking-based Approach to Twitter User Geolocation Prediction. In *ACL (Conference System Demonstrations)* (pp. 7–12).

Handy, S. (2005). Smart growth and the transportation-land use connection: What does the research tell us? *International Regional Science Review*, *28*(2), 146–167.

Hanson, S., & Huff, O. J. (1988). Systematic variability in repetitious travel. *Transportation*, *15*(1), 111–135.

Hasan, S., Zhan, X., & Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (p. 6). ACM.

Hill, K. (2012). Hurricane Sandy, @ComfortablySmug, and The Flood of Social Media Misinformation. Retrieved January 28, 2017, from http://www.forbes.com/sites/kashmirhill/2012/10/30/hurricane-sandy-and-the-flood-of-social-media-misinformation/#5d36cde7d967

Hsieh, Y.-H. (2010). Age groups and spread of influenza: implications for vaccination strategy. *BMC Infectious Diseases*, *10*(1), 106.

Hughes, A. L., St Denis, L. A. A., Palen, L., & Anderson, K. M. (2014). Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1505–1514). ACM.

Janelle, D. G., Goodchild, M. F., & Klinkenberg, B. (1988). Space-time diaries and travel characteristics for different levels of respondent aggregation. *Environment and Planning A*, *20*(7), 891–906.

Jones, J. F., Hook, S. A., Park, S. C., & Scott, L. M. (2011). Privacy, Security and interoperability of mobile health applications. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 46–55). Springer Berlin Heidelberg.

Jones, P., & Clarke, M. (1988). The significance and measurement of variability in travel behaviour. *Transportation*, *15*(1–2), 65–87.

Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM*, *13*(13), 273–282.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2013). RTextTools: A supervised learning package for text classification. *The R Journal*, *5*(1), 6–12.

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

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly*, *68*(1), 94–101.

Khattak, A., & Rodriguez, D. (2005). The impact of neo-traditional developments on traveler behavior. *Transportation Research A*, *39*(6), 481–500.

Kockelman, K. (1997). Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area. *Transportation Research Record: Journal of the Transportation Research Board*, (1607), 116–125.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating Fact from Fear: Tracking Flu Infections on Twitter. In *HLT-NAACL* (pp. 789–795).

Lessler, J., Reich, N. G., & Cummings, D. A. T. (2009). Outbreak of 2009 pandemic influenza A (H1N1) at a New York City school. *New England Journal of Medicine*, *361*(27), 2628–2636.

Marsden-Haug, N., Foster, V. B., Gould, P. L., Elbert, E., Wang, H., & Pavlin, J. A. (2007). Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. *Emerging Infectious Diseases*, *13*(2), 207.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Artificial Intelligence*, 554–557. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2816/3234

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *ICWSM*.

Murthy, D., Gross, A., & Pensavalle, A. (2016). Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities. *Journal of Computer-Mediated Communication*, *21*(1), 33–49. https://doi.org/10.1111/jcc4.12144

Oyeyemi, S. O., Gabarron, E., & Wynn, R. (2014). Ebola, Twitter, and misinformation: a dangerous combination? *Bmj*, *349*, g6178.

Palen, L., Vieweg, S., Liu, S. B., & Hughes, A. L. (2009). Crisis in a networked world features of computer-mediated communication in the April 16, 2007, Virginia Tech Event. *Social Science Computer Review*, *27*(4), 467–480.

Paul, M. J., Dredze, M., & Broniatowski, D. (2014). Twitter improves influenza forecasting. *PLoS Currents*, *6*.

Peköz, E. A., Shwartz, M., Iezzoni, L. I., Ash, A. S., Posner, M. A., & Restuccia, J. D. (2003). Comparing the importance of disease rate versus practice style variations in explaining differences in small area hospitalization rates for two respiratory conditions. *Statistics in Medicine*, *22*(10), 1775–1786.

Pennacchiotti, M., & Popescu, A.-M. (2011). A Machine Learning Approach to Twitter User Classification. *Icwsm*, *11*(1), 281–288.

Pew Research Center. (2015). Social Media Usage: 2005-2015 | Pew Research Center. Retrieved January 16, 2017, from http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860).

Salathe, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., … others. (2012). Digital epidemiology. *PLoS Comput Biol*, *8*(7), e1002616.

Santos, J. C., & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, *11*(Suppl 1), S6.

Sarcevic, A., Palen, L., White, J., Starbird, K., Bagdouri, M., & Anderson, K. (2012). Beacons of hope in decentralized coordination: Learning from on-the-ground medical twitterers during the 2010 Haiti earthquake. In *Proceedings of the ACM 2012 conference on computer supported cooperative work* (pp. 47–56). ACM.

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.

Schull, M. J., Morrison, L. J., Vermeulen, M., & Redelmeier, D. A. (2003). Emergency department overcrowding and ambulance transport delays for patients with chest pain. *Canadian Medical Association Journal*, *168*(3), 277–283.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS One*, *6*(5), e19467.

Social Media Update. (2016). Retrieved January 12, 2017, from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference 2014 Proceedings*.

Starbird, K., & Palen, L. (2012). (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In *Proceedings of the acm 2012 conference on computer supported cooperative work* (pp. 7–16). ACM.

Sui, D., Elwood, S., & Goodchild, M. (2012). *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media.

Sutton, J. N., Palen, L., & Shklovski, I. (2008). *Backchannels on the front lines: Emergency uses of social media in the 2007 Southern California Wildfires*. University of Colorado.

Taubenberger, J. K., & Morens, D. M. (2008). The pathology of influenza virus infections. *Annu. Rev. Pathmechdis. Mech. Dis.*, *3*, 499–522.

Trzeciak, S., & Rivers, E. P. (2003). Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. *Emergency Medicine Journal*, *20*(5), 402–405.

Tulloch, D. (2014). Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice. *International Journal of Geographical Information Science*, *28*(4), 847–849.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178–185.

Twitter helps Chicago find sources of food poisoning | Reuters. (2014). Retrieved January 12, 2017, from http://www.reuters.com/article/us-chicago-twitter-food-poisoning-idUSKBN0GQ25820140826

Venkatraman, A., Mukhija, D., Kumar, N., & Nagpal, S. J. S. (2016). Zika virus misinformation on the internet. *Travel Medicine and Infectious Disease*, *14*(4), 421.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1079–1088). ACM.

Wennberg, J. E. (1979). Factors governing utilization of hospital services. *Hospital Practice*, *14*(9), 115–127.

Wilson, M. W., & Graham, M. (2013). Situating neogeography. *Environment and Planning A*, *45*(1), 3–9.

Wright, E., Khanfar, N. M., Harrington, C., & Kizer, L. E. (2010). The lasting effects of social media trends on advertising. *Journal of Business & Economics Research*, *8*(11), 73.

Wright, K. B. (2005). Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, *10*(3), 0.

Yamamoto, T., & Kitamura, R. (1999). An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days. *Transportation*, *26*(2), 231–250.

Yang, X., Wu, Z., & Li, Y. (2012). Using Internet reports for early estimates of the final death toll of earthquake-generated tsunami: the March 11, 2011, Tohoku, Japan, earthquake. *Annals of Geophysics*, *54*(6).

Yu, Y., & Wang, X. (2015). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. *Computers in Human Behavior*, *48*, 392–400. https://doi.org/10.1016/j.chb.2015.01.075