USE OF SOCIAL MEDIA TO MONITOR AND PREDICT OUTBREAKS AND PUBLIC OPINION ON HEALTH TOPICS

by

Alessio Signorini

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science in the
Graduate College of The
University of Iowa

December 2014

Thesis Supervisor: Professor Alberto Maria Segre

Graduate College
The University of Iowa
Iowa City, Iowa

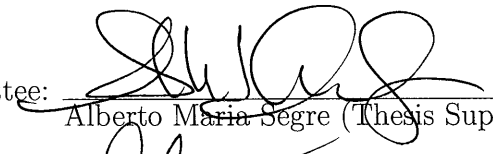
CERTIFICATE OF APPROVAL

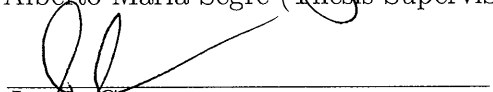
PH.D. THESIS

This is to certify that the Ph.D. thesis of

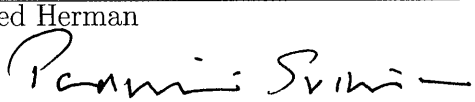
Alessio Signorini

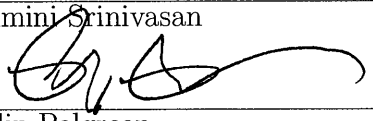
has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Computer Science at the December 2014
graduation.

Thesis committee: 
Alberto Maria Segre (Thesis Supervisor)


James Cremer


Ted Herman


Padmini Srinivasan


Philip Polgreen

ACKNOWLEDGMENTS

First and foremost I want to thank my advisor Alberto Maria Segre for helping me with my studies and my life from the moment I set foot in the United States. I really appreciate all his contributions of times, ideas, funding and support to make my P.h.D. experience productive and stimulating. He has been the best advisor I could hope for and this thesis has been completed largely thanks to his patience, his nudges, and the freedom he allowed me to have in exploring my own research interest.

I would like to thank my father Gianluca Signorini and my grandfather Luciano Redini for having introduced me to electronics and computers. I would also like to thank my mother Antonella and my entire family for being patient with me while away for so long to pursue my dreams.

Special thanks also go to my mentors Bruno Codenotti, Antonio Gulli, Apostolos Gerasoulis, Kimbal Musk and David Tisch for having guided me along the way and supported my ambitions and crazy projects. I would have never accomplished all I have so far without their support.

Lastly, I would like to thank Philip Polgreen for having introduced me to the field of epidemiology and worked with me on this research, Sheryl Semler and Catherine Till for all the help in making sure I was always registered for class and had all my papers in order, and all the great friends I have around the world, who consciously or unconsciously had a major impact in my life.

ABSTRACT

The world in which we live has changed rapidly over the last few decades. Threats of bioterrorism, influenza pandemics, and emerging infectious diseases coupled with unprecedented population mobility led to the development of public health surveillance systems. These systems are useful in detecting and responding to infectious disease outbreaks but often operate with a considerable delay and fail to provide the necessary lead time for optimal public health response.

In contrast, syndromic surveillance systems rely on clinical features (e.g., activities prompted by the onset of symptoms) that are discernible prior to diagnosis to warn of changes in disease activity. Although less precise, these systems can offer considerable lead time. Patient information may be acquired from multiple existing sources established for other purposes, including, for example, emergency department primary complaints, ambulance dispatch data, and over-the-counter medication sales. Unfortunately, these data are often expensive, sometimes difficult to obtain and almost always hard to integrate.

Fortunately, the proliferation of online social networks makes much more information about our daily habits and lifestyles freely available and easily accessible on the web. Twitter, Facebook and FourSquare are only a few examples of the many websites where people voluntarily post updates on their daily behaviors, health status, and physical location.

In this thesis we develop and apply methods to collect, filter and analyze the content of social media postings in order to make predictions. As a proof of

concept we used Twitter data to predict public opinion in the form of the outcome of a popular television show. We then used the same methods to monitor and track public perception of influenza during the H1N1 epidemic, and even to predict disease burden in real time, which is a measurable advance over current public health practice. Finally, we used location specific social media data to model human travels and show how this data can improve our prediction of disease burden.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 DISEASE SURVEILLANCE 1

1.1 Importance of Surveillance 2
1.2 Types of Surveillance Systems 4
1.3 National Electronic Disease Surveillance System 7
1.4 Introduction to Social Media 11
1.4.1 Blogs 18
1.4.2 Wikipedia 20
1.4.3 Twitter 21
1.4.4 Facebook 27
1.4.5 Flickr 30
1.4.6 FourSquare 32
1.4.7 Other Sources of Data: Proxy & Search Logs 34
1.4.8 Privacy Concerns 36
1.5 New Technology and Disease Surveillance 38
1.5.1 Related Research 41
1.5.2 Social Media for Disease Surveillance 44

2 RESEARCH APPROACH AND METHODOLOGIES 45

2.1 Twitter 46
2.1.1 Anatomy of a Tweet 47
2.1.2 Twitter's API 53
2.1.3 Data Gathering and Normalization 55
2.1.4 Stemming 57
2.1.5 Language Classification 59
2.1.6 Spam and Unrelated Tweet Removal 62
2.1.7 General Applicability 65
2.2 Support Vector Regression 65

3 APPLICATIONS 68

3.1 Predicting the American Idol 2009 Winner 69
3.2 Monitoring the Swine Flu Pandemic 73
3.3 Using Twitter to Estimate H1N1 Influenza Activity 77
3.4 Inferring Travel from Social Media 80
3.5 Predicting Local Flu Trends using Geolocated Tweets 85
3.5.1 City Level Flu Data 86
3.5.2 Travel Data 87
3.5.3 Flu Correlation between Cities: distance vs. flow 91
3.5.4 Predicting Flu Trends across Cities 92

4 CONCLUSION 96

APPENDIX

LIST OF KEYWORDS 101
A.1 Keywords used in study of Section 2.1.7 102
A.2 List of suffixes removed in step 5 of Porter's Algorithm . 102

BIBLIOGRAPHY 103

LIST OF TABLES

Table

1.1 Percentage of Americans performing common activities online . . . 12
1.2 Types of Tweets posted by users . . . 26
1.3 Types of Link shared on Tweets by users . . . 27
1.4 Values of commonly awarded checkins on Foursquare . . . 33
2.1 Search Keywords used as filters in Twitter's API . . . 56
3.1 MMWR Cities removed due to lack of data . . . 88
3.2 MMWR Cities removed due to overlapping metro areas . . . 90
3.3 Population vs. Twitter Penetration - Top 10 cities . . . 91
3.4 Square Correlation Coefficients for each approach . . . 94
3.5 Square Correlation Coefficients for most predictable cities . . . 95
3.6 Square Correlation Coefficients for most difficult cities . . . 95

LIST OF FIGURES

Figure

1.1 Number of Visits to "2009 Swine Flu Outbreaks" page on Wikipedia 21
1.2 Top categories of Foursquare checkins 32
1.3 Examples of Foursquare Badges 34
2.1 Example of Tweets 47
2.2 Use of Hashtags in a Tweet 48
2.3 Example of a Retweet 51
2.4 Example of Favorites for a Tweet 52
2.5 Example of Direct Reply on Twitter 53
2.6 Area used as geographical filter for Twitter 56
2.7 Percentage of Non-English Tweets given Twitter Profile Language . 60
2.8 Percentage of English vs. non-English Tweets given a US Timezone 61
2.9 Percentage of English Tweets by Hour of Day 61
2.10 Percentage of Spam/Non-Spam Tweets with certain features 64
3.1 Tweet Volume associated to each American Idol 8 contestant 70
3.2 Google Search Volume for American Idol 7 Contestants 71
3.3 Google Search Volume for American Idol 8 Contestants 72
3.4 Relative Number of Tweets during the Finale of American Idol 8 . . 72
3.5 Screenshot of H1N1 Realtime Monitor Interface 75
3.6 Tweet Volume for each Category by Date 76
3.7 Predicted vs. Reported ILI% in the U.S. for the 2009 Flu Season . . 79
3.8 Predicted vs. Reported ILI% in Region 2 for the 2009 Flu Season . 80
3.9 Statistics on Geographical Distance between Foursquare Checkins . 81
3.10 Statistics on Time between Foursquare Checkins 82
3.11 Travel plot with state level resolution 83
3.12 Density of Foursquare checkins in New York City 83
3.13 Density of Foursquare checkins in Manhattan by time of day 84
3.14 User paths across New York City inferred through Foursquare checkins 85
3.15 Flu & Pneumonia Deaths in New York City, NY for 2012 88
3.16 Distance vs. Correlation for Atlanta, GA 92
3.17 Flow vs. Correlation for Atlanta, GA 93

CHAPTER 1
DISEASE SURVEILLANCE

The world in which we live has changed rapidly over the last few decades. Threats of bioterrorism, influenza pandemics, and emerging infectious diseases coupled with unprecedented population mobility led to the development of surveillance systems for public health. According to Thacker and Berkelman [84] these systems perform an "ongoing systematic collection, analysis, and interpretation of data, closely integrated with the timely dissemination of these data to those responsible for preventing and controlling disease and injury" and are generally put in place by governmental organizations (e.g., ministries of health or finance) to assess in real time the health status and behavior of certain populations to allow decision makers to lead and manage resources more effectively. Since these monitoring systems can directly measure what is happening in a population they can be used both to assess the need for an intervention and directly verify its effects.

The key objective of public health surveillance is to guide interventions. The monitoring systems put in place generally aim to gather scientific and factual data essential to make informed decision and plan appropriate public health responses, and their design and implementation is often influenced by their objectives. Different public health objectives and the actions necessary to reach them may require different information systems. The type of action to be taken, when and how often it needs to be performed, what information is needed to take or monitor the action and how frequently the information is needed determines the type of surveillance or health information system to be used. For example, if the goal is to prevent the spread of acute infectious diseases (e.g., SARS) the surveillance system needs to be

effective in providing early warning signs so that managers can intervene quickly and stop potential epidemics. In contrast, the surveillance of chronic diseases (e.g., tuberculosis) or health-related behaviors (e.g., tobacco smoking) that have a relatively slow change rate, can be simply performed through demographic and health surveys done once a year.

1.1 Importance of Surveillance

The World Health Organization (WHO) and the World Bank consider [109] surveillance to be an essential function of a public health system, improving the efficiency and effectiveness of the services performed thanks to targeted interventions and documentation of its effects on the population. Since 1975, the Center for Diseases Control and Prevention (CDC) and the WHO have collaborated with more than 30 countries to strengthen health systems and address training needs for disease detection and response in a country-specific, flexible, and sustainable manner. State members of WHO need to comply with the guidelines set by the International Health Regulations and have key persons and core capacities in surveillance.

In 1993 the WHO developed (in Africa) the Integrated Disease Surveillance and Response (IDSR) strategy [110] which linked epidemiological and laboratory data at all levels of the health system, putting an emphasis on integrating surveillance with response. The approach was very comprehensive and included detection, registration and confirmation of case-patient, reporting, analysis and use of data, outbreak investigations and contact tracing.

In the late 1980s, while monitoring a population of 60 million people the Philippine Department of Health's (PDOH) integrated management information system [29] detected less than one outbreak per year. Nine years later the PDOH introduced the National Epidemic Sentinel Surveillance System, a hospital-based

sentinel surveillance system which provided rules for both the flow of data and the personnel requirements. The pilot study was a success and the system was integrated into the public health system and expanded to include HIV serological and behavioral risk surveillance. In 1995 alone, the system detected and investigated about 80 outbreaks.

In 2005, China launched its first Field Epidemiology Training Program (FETP) to rapidly expand its surveillance and response capacity, while Brazil and Argentina chose to use World Bank funds to improve their own systems. At the same time, with more and more data available through various channels, the U.S. Agency for International Development (USAID) redesigned its surveillance strategy to focus on the use of data to improve public health interventions [98]. These new systems were adapted to their local reality by many countries: Guatemala's marriage of its FETP (part of a larger, Central American FETP) with the Data for Decision Making program [64] is one example, and India, with its decentralized system, complex cultural and population dynamics, and wide variance in the sophistication of public health institutions, provides another model for strengthening national surveillance.

As of today more than half of world's population lives in a country where public health surveillance is carried out by staff members and trainees of FETPs or allied programs. Programs like the Epidemic Intelligence Service in the United States, the European Program for Intervention Epidemiology Training, and Public Health Schools without Walls provide most of the surveillance and response to emerging infections in these countries in addition to train the majority of the public health workers in the sector.

1.2 Types of Surveillance Systems

In their 1976 article [38] on the International Journal of Epidemiology, Foege and others stated

> the reason for collecting, analyzing, and disseminating information on a disease is to control that disease. Collection and analysis should not be allowed to consume resources if action does not follow

Public health surveillance systems should be implemented in such a way to provide valid and timely information to decision makers at the lowest possible cost. The utility of the data collected can be viewed as immediate, annual or archival, on the basis of the actions that can be taken. Similarly, spatial resolution of the data collected (e.g., macro vs. micro areas) may be sacrificed with the aim to improve timeliness and save resources. For these reasons, it is not always possible nor effective to deploy complex surveillance systems. In developing countries, for example, a critical challenge of the health sector is to ensure quality and effectiveness of the surveillance in decentralized environments. National-level programs and surveillance system managers may lose control of the quality and timeliness of the data collected and donors, perceiving weakness in the national system, may create parallel nongovernmental surveillance systems to gather directly the data they need. These systems generally work in the short-term, but in the long run, weaken even more the public health surveillance programs already in place.

Many types of surveillance systems exist [4] and are effectively deployed everywhere around the world. Among the most commonly used ones are:

Vital Statistics Keeping records of the number of births and deaths has been long used as indicator of overall population health. Infant mortality rate (the

number of deaths among infants per 1,000 births) is also used as risk factor for a variety of adverse health outcomes. In the United States (US), vital statistics are available from the National Center for Health Statistics and from state vital records offices. The CDC also operates an online system (called CDC WONDER) containing data on births, deaths, and many diseases.

Registries Registries are a simple type of surveillance system used for particular conditions (e.g., cancer or birth defects). They are often established at a state level to collect information about the number of people diagnosed with a certain conditions and are generally used to improve prevention programs.

Population Surveys Routine surveys are surveillance tools are generally repeated on a regular basis [73] and can be very useful in monitoring chronic diseases and health-related behaviors. While theoretically simple to implement, surveys require a clear definition of the target population to which the results can be generalized. In addition, to avoid bias, the sample size needs to be adequate to the health condition under surveillance (i.e., rare conditions require substantial samples). Two well known national surveys conducted in the U.S. are the Youth Risk Behavior Survey (YRBS) and the Behavior Risk Factor Surveillance System (BRFSS). In these surveys high school students and adults are asked about health-related behaviors such as substance use, nutrition, sexual behavior, and physical activity. The results are used to monitor trends in health behavior (e.g, YRBS showed decline in youth smoking from 36% in 1997 to 20% in 2007), plan public health programs, and evaluate public health policies at national and state levels.

Disease Reporting The International Health Regulations introduced by the WHO require timely reporting to public health officials for certain diseases. In

addition, countries are also required to report any public health emergency of international concern. In the United States, disease reporting is mandated by state law and the list of reportable diseases vary by state. States report nationally notifiable diseases to the CDC on a voluntary basis.

Adverse Event Surveillance The purpose of these systems is to gather information about negative effects experienced by people who have taken prescribed drugs and other therapeutic agents. Reports may come from health care providers (e.g., physicians, pharmacists, and nurses) as well as members of the general public, such as patients or lawyers, and manufacturers. Some examples of adverse events surveillance focused on patient safety are the FDA Adverse Events Reporting System FAERS [37] and the Vaccine Adverse Events Reporting System[1] (VAERS). The former is operated by the Food and Drug Administration (FDA) while the latter is mostly operated by the CDC. Due to their passive nature, AERS and VAERS may suffer from underreporting or biased reporting, and while they cannot be used to determine whether a drug or vaccine caused a specific adverse health event, they are fairly useful as early warning signals.

Sentinel Surveillance In a sentinel surveillance system, a predefined sample of reporting sources agrees to report all cases of defined conditions [73]. When properly implemented, sentinel-based systems offer an effective method of flexible monitoring with limited resources. While these systems are very effective in detecting large health problems, they may be insensitive to rare events (e.g., emergence of a new disease). One of the most well known sentinel surveillance systems used in the United States is for influenza, where selected health care providers report the number of cases of influenza-like illness to their state health department on a

[1] vaers.hhs.gov

weekly basis, allowing monitoring of macro trends using a relatively small amount of information.

Zoonotic Disease Surveillance Zoonotic surveillance system involve systems for detecting animals infected with diseases that can be transmitted to humans. Efforts of this type were very effective [6] in 2001 during an epidemic of West Nile Virus (WNV) in Florida, and led to public health control measures, such as advising the public to protect against mosquito bites and intensifying mosquito abatement efforts.

Laboratory Data Public health laboratories that routinely conduct tests for viruses, bacteria, and other pathogens can be another useful source of surveillance data. Laboratory serotyping provides information about cases that are likely to be linked to a common source and is useful for detecting local, state, or national outbreaks.

Syndromic Surveillance This method of surveillance has been introduced only recently and uses clinical information about disease signs and symptoms as opposed to diagnosis data. It can be active or passive and is based entirely on clinical features (e.g., collecting cases of diarrhea) without any clinical or laboratory diagnosis (e.g., cases of cholera). One important source of data are hospital emergency rooms, which can provide the health department with early notification of new outbreaks.

1.3 National Electronic Disease Surveillance System

According to the CDC, the majority of the cases of diseases and other conditions of interest are generally identified within the health care system. Once identified, these

are typically reported to a local health department, which aggregates them (either digitally or on paper-based forms) before sending them to the state health department where they are manually entered into the state's electronic system. Some of these data may then be aggregated at federal level. These reporting processes are generally the same, regardless of the disease or condition that is being reported, and the data transfer often occurs long after disease incidences are first reported.

Many problems can arise during the reporting process, and these, in turn, often place a large burden on the medical care staff who have responsibility for the reporting. For example, it is often left up to the health provider staff (which is frequently already overworked) to determine if a case meets public health surveillance case definitions and to figure out how to fill out the wide variety of forms produced by CDC and health departments. In some cases, the staff may also need to spend significant time tracking down all the records that need to be attached to the report. The result is that many diseases are underreported, inadequately documented or inaccurately recorded.

According to the CDC in the late 1990s, more than 100 different systems were used to transmit reports to the federal agency. These systems were isolated from one another due to differing data standards, legacy systems, patient privacy concerns and a lack of tools for information exchange. To reduce the burden imposed on medical care staff, minimize human error, and facilitate the transmission of these important medical data, the CDC designed and introduced the National Electronic Disease Surveillance System (NEDSS). The system was created to integrate and replace a number of existing CDC surveillance systems, including the National Electronic Telecommunications System for Surveillance (NETSS), HIV/AIDS reporting systems, vaccination programs, and tracking systems for tuberculosis and

other infectious diseases.

NEDSS is a secure online framework that allows healthcare professionals and government agencies to communicate about disease patterns and coordinate national response to outbreaks. The framework includes a set of specifications for software, hardware, databases and data format standards. Its base system is a platform that state agencies and health care providers can use to integrate surveillance systems data processing in a secure environment. It is composed by five main components:

- a Web-based module that allows easy online entry and management of data sets, including demographic and disease data;
- a Web application server called Silverstream that supports these Web-based modules;
- an integrated database management system;
- messaging software (i.e., HL7 Standard[2]) that allows electronic data interchange between state agencies and the CDC or state laboratories; and
- intranet-based authentication and authorization for security that is fully compliant with HIPAA regulations.

Once NEDSS is fully implemented across the United States, public health professionals and government agencies will receive timely alerts of disease outbreaks and bioterrorism attacks. The Centers for Disease Control and Prevention is in charge of maintaining and expanding NEDSS at the core of the Public Health Information Network (PHIN). The CDC requires that hospitals, clinics and state health agencies all adopt NEDSS standards so that the speed, accuracy, standardization and

[2]www.hl7.org

viability of data about diseases is improved.

The introduction of standards assures consistent data collection practices across the country. The public health data model and common data standards recommend, among other things, a minimum set of demographic data that should be collected as part of routine surveillance. In addition, the guidelines provide a uniform method for coding data (e.g., LOINC [72] as the standard for transmitting laboratory test names and SNOMED [51] as the standard for transmitting test results) on the data collection forms and defines its content (e.g., disease diagnosis, risk factor information, lab confirmation results, and patient demographics).

NEDSS also includes recommendations for standards that can be used for the automatic electronic reporting of surveillance data. Specifically, it provides guidelines for a standard data architecture and electronic data interchange format (i.e., HL7 Standard) to allow computer systems to automatically generate digital case reports ready to be sent to local or state health departments. These types of standards ease the burden on large organizations that already have computerized data systems (such as regional laboratories, hospitals, managed care organizations) and ensure that all cases that are in the providers data systems are reported to public health officers.

Standardized data collection forms ease the burden on physicians and their staff providing a single web-based data entry portal for all reportable conditions. Similarly, larger organizations can use automated electronic data exchanges that impose minimal burden on health-care reporters.

As of today, 46 states, New York City, and Washington, D.C., send case notifications to National Notifiable Disease Surveillance System (NNDSS) through a NEDSS-compatible system. According to the CDC [10], to be considered NEDSS

compatible, states must have information systems meeting these requirements:

- disease data entry directly through an Internet browser-based system;
- Electronic Laboratory Reporting (ELR);
- integration of multiple health information databases into a single repository; and
- electronic messaging. capabilities

The combination of these features allows states to create a single repository containing all the health information which is directly accessible by health investigators, and a secure channel to efficiently share data with the CDC and other health agencies.

1.4 Introduction to Social Media

According to a recent study [99] of the U.S. Department of Commerce, the number of households with a computer increased from 36% to 76% between 1997 and 2011, with 72% of these using the computers to connect to the Internet. The same study reveals that 27% of people are able to access the Internet both inside and outside the home from multiple devices. Similar statistics can be found in the updated reports [52] of Internet World Stats, which show 85% Internet penetration among the population of the US. While the world average is much lower (39%), Oceania, Australia and Europe closely follow North Americans with 67% penetration among their respective populations. It is interesting to note how in Europe the northern states (e.g., Iceland, Norway, Finland and Netherlands) lead the chart with an average Internet penetration of nearly 95%.

A Nielsen report [86] on US Internet Usage shows daily usage by the average American of about 60 hours per month, with the majority of the accesses performed through their smartphones. Clearly, the increased popularity of computers with high speed connections has changed the lives and behavior of millions of people. According to a Mediamark Research Survey done in fall 2008 [55], many tasks that were once done manually are now typically completed online (see 1.1).

Activity	**% of Americans**
Read News Online	46.00%
Paid Bills Online	39.60%
Personal Shopping	37.20%
Shared Photos	25.40%
Searched for Recipes	24.80%
Arranged a Trip	20.50%
Obtained Medical Advice	19.90%
Looked for Movie Showtimes	19.70%
Searched for Employment	15.30%
Traded Stocks	13.20%
Listened to the Radio	13.10%

Table 1.1: Percentage of Americans performing common activities online

The growth of the Internet has also increased the amount of information accessible to the general public on any topic. The web is seen by many as a giant (free) library where anything can be found. In fact, many schools and teachers have

had to introduce strict no-Internet-references policies in their classes, forcing the students to find "real" sources as references for their assignments. Electronic encyclopedias, epitomized for many years by the Encyclopedia Britannica and Microsoft Encarta, have now been replaced by Internet-based crowd-sourced publications like Wikipedia. Other paper-only publications have suffered a similar fate: many scientific journals and conference proceedings are often no longer offered on paper but rather distributed on some sort of electronic medium, such as on DVDs or memory cards. All this content is also made available on a website, where it is easily found and indexed by the major search engines.

For any type of content, accessibility and searchability are very important properties in today's connected world, where geographical locations and political borders are less of an impediment. Thanks to the Internet, researchers in Italy can easily share their data with groups based in Tokyo, or compare their results with the early outcomes of similar studies done in Canada, all practically in real time. Experiments of this kind are already a reality: in November of 2000, a monkey at Duke University in North Carolina was connected through the Internet to a robotic arm in the Massachusetts Institute of Technology (MIT) Touch Lab, more than 600 miles away [27]. Maps of planets and stars, weather forecasts and histories, geological charts and photos that were once available only to a restricted circle of scientists and graduate students, can now be found online by anyone in just a few minutes.

Among the most popular sites are health-care related websites, and a recent PEW Internet survey [66] reported that more than 72% of Internet users have looked up health information online in the last year. Questions that were once answered by consulting the Medical Encyclopedia are now answered online. Even

small laboratory studies, which a few years ago were at best published in low-circulation venues, receive a lot of attention thanks to references from passionate bloggers who elevate interest in these studies, making them generally available.

Other websites cover a wide range of possible medically-related necessities. Fitness and weight loss are among the most popular, with sites like Self, Men's Health and Weight Watcher leading the category with the highest number of visitors [2]. Another popular category are the disease-centered websites, where any user can try to auto-diagnose by selecting the symptoms experienced and letting the site suggest possible causes. Among the most popular sites in this category [3] we can find WebMD, Mayo Clinic and Yahoo!Health. Finally, support group websites (for addictions, substance abuse, or rare diseases) are also among the most visited. These are generally non-profit sites that aim to connect people in similar situations, to exchange information, help users deal with their shared problems and speed recovery. A recent PEW Internet Survey [41] reported that 8% of Internet users living with a chronic disease participate in an online discussion or forum. One of the most well-known site of this category is PatientsLikeMe[3], which as more than 250,000 members at the time of writing.

Example: 2005 Hurricane Katrina In times of crisis and emergency, the members of the public and affected communities are often the first to react, respond, and mobilize in order to help others in need. With the advent of web 2.0 and social media technologies, both bystanders and victims can and have been using these tools to communicate, document (i.e., citizen journalism), and rally aid in innovative ways. For example, when the Hurricane Katrina hit the US Gulf Coast,

[3] www.patientslikeme.com

> Louisiana and Mississippi took the brunt of the damage. Hundreds of thou- sands were displaced and at least 1,800 people lost their lives. The storm severely damaged the communication infrastructure and caused widespread power failures. The resulting devastation left many relief and federal organizations overwhelmed. A large number of grassroots efforts such as katrinahousing.net by the University of Michigan and "Craigslist Katrina Relief" emerged to provide aid, housing, necessities, and employment to those affected. To deal with the dearth of timely and accurate information, locals turned elsewhere for information, and actively worked to generate and disseminate accurate information. The Times-Picayune, a local newspaper, created discussion boards that were neighborhood specific and provided maps and satellite images. This was the first disaster where more formalized posting procedures were attempted by group administrators, but these were met with limited success. For example, on one group photo tagging instructions such as #KatrinaMissing, #KatrinaFound, #KatrinaOkay were issued to try and create a database of survivors, victims, and missing persons, but these instructions were not often observed by members.

Blogs, have become quite popular, and a recent study [50] by IgniteSpot reports that more than 77% of Internet users read blogs. While some users use their blogs to collect and share their favorite recipes, thoughts and projects, some blogs reached great popularity and become fully developed online magazines (e.g., Mashable[4]).

Blogs are not the only way people share and collaborate on content on the

[4]www.mashable.com

Internet. The most well-known collaborative effort is Wikipedia[5], a free encyclopedia created, edited and updated by users around the world. Since its creation in 2001, Wikipedia has attracted more than 75,000 editors [107] who have created more than 15 million documents. Although the English version is the largest (4.6M documents), many of the articles are available in 260 languages and is used by more than 400M unique users every month[106].

Together with Wikipedia, other social networking sites found fertile ground in the Internet during recent years. A recent report [67] by PEW Internet shows that more than 74% Americans have an account on at least one social network, and according to data [17] from Cisco, a shocking 90% of 18-30 years old check their account shortly after waking up. Projects like Facebook[6] and Twitter[7] have became hugely popular, gathering billions of active users of all ages. Users connect (or re-connect) with friends, partners and colleagues, sharing photos, videos, and other personal information.

In addition, thanks to relatively recent but now wide-spread embedding of GPS hardware in mobile devices (80% at the end of 2011 [71]), a new wave of applications regularly exploit this hardware in order to incorporate geographic information into social network traffic. This development permits people to focus on "hyper-local conversations", and applications like FourSquare[8] encourage people to declare their location (i.e., "checkin"), to help them connect with nearby friends and receive discounts for their loyalty to a brand or a particular shop.

[5] www.wikipedia.org
[6] www.facebook.com
[7] www.twitter.com
[8] www.foursquare.com

Example: 2008 China Sichuan Earthquake When the 2008 Sichuan earthquake occurred the famous Bay Area tech blogger Robert Scoble posted the event [9] on social media before either the mainstream media or the US Geological Survey could issue news releases. The official reports and news came about one hour later. Due to a combination of heavy damage to the telecommunication infrastructure and overwhelming call volume, both landline and cell phone services in the area failed. Many turned to the Internet for help and information.

There are two widely publicized success stories where Tianya[10] members provided critical information that proved useful to the authorities. In the first case [76], the military was attempting send relief to a remote area but were unable to find suitable landing strip and had to delay their efforts. Upon hearing this, a young woman who had grown up in the area but was currently away at school posted on Tianya the location of a suitable helicopter landing spot. The post was forwarded thousands of times to all of the major online communities until it eventually reached the military. Upon contacting the student, the military was able to land where she had described and deploy troops and equipment to those in need.

In the second case [112], the forum provided valuable feedback to government officials. A message that raised much concern from members provided details about the possible embezzlement of relief supplies by officials. This post attracted the attention of the government, who quickly

[9] https://twitter.com/Scobleizer/statuses/809121152
[10] http://www.tianya.cn/

investigated the situation and punished the offending individuals.

1.4.1 Blogs

Blogs were originally conceived of as replacements for old-fashion diaries; private sites that could be easily edited from anywhere and could also be enriched with all sort of media (e.g., photos, videos, music). According to recent estimates [108] [102], about 2.8M blog posts are published every day and globally more than 650 million users read blogs. In addition to a few, very important, commercial instances (e.g., Mashable, TechCrunch, DailyBeast), blogs are widely used by the tech community to share snippets of code, technical advice, and ideas. In contrast, the non-tech community generally uses blogs to share their thoughts, record recipes, give fashion advice, or to collect and document important moments in their lives (e.g., weddings, vacations).

In their blog posts, people express personal feelings and opinions about life, products, recent news or events. Since many users treat their blogs as a personal diary, the language adopted and the entities cited can often allow the identification of many personal details. For example, it is not uncommon to find posts entitled "my 30th birthday" [28], which allow analysts to determine the age of the writer with high precision. Some posts may describe an evening out, mentioning identifiable landmarks (e.g., "we got a cab to lower Manhattan"), places (e.g., "Time's Square was packed") or venues (e.g., "we had dinner at the Four Seasons"). Other posts offer clues about the gender of the writer, for example, comments about a new pair of shoes, relationship problems or a new dress might suggest a female writer, while opinions on the current situation of the stock market or the weekend's sport results, increase the probability of facing a male blogger.

While such details might help to identify the location, gender and age of the writer, the complexity of the language used in the posts makes it difficult to automatically identify the mood and attitude of the writer (e.g., happy, confused, frustrated) as well as the category of the post (e.g., sports, politics, history). Although difficult to achieve, automatic categorization of blog posts could be very useful in many occasions, as for example while trying to summarize the opinion of the public about certain products or topics.

There have already been many attempts to classify blog posts. In 2005, Gilad Mishne published a paper describing the early outcomes of his experiments leading to the development of MoodViews [62]. In his work, Gilad obtained about 850,000 mood-annotated blog posts from LiveJournal[11] and tried to identify discriminative features (and their weights) in the post's text for each different mood. Unfortunately, the precision achieved by the method tested is barely above (67%) the baseline (50%, or random guessing) and more work is clearly necessary to make it usable.

Similar work has been been published by Tyrrel et al. in 2006 [14]. In their work the authors simplified the approach taken by Mishne and tried to classify the posts into just 3 main classes: objective, positive or negative. The classification method was based mainly on the identification of the polarity of adjectives and verbs which they obtained from Wikitionary[12] and the weight of each term was computed using Support Vector Machine (SVM) classification. The final accuracy of the method was close to 90%.

Automatic classification of blog posts could be really useful in identifying the perceptions of the general public regarding some products or topics. In the health

[11] www.livejournal.com
[12] www.wiktionary.org

context, it could be useful to identify moods and opinions about certain diseases or vaccines which might permit public health officials to better address problems and concerns.

1.4.2 Wikipedia

Wikipedia is a collaboratively edited, multilingual, free Internet encyclopedia that is supported by the non-profit Wikimedia Foundation. It was launched on January 2001 by Jimmy Wales and Larry Sanger and has more than 18 billion page views. Today it contains more than 15 million articles written in 267 languages [106].

Wikipedia is one of the principal sources of information on the Internet and its pages often appear among the top 3 URLs in search engines results. This free encyclopedia is maintained by thousands [107] of passionate volunteers all over the world who constantly create, update and perfect its articles. In the past few years, Wikipedia has reacted to new trends and topics with great speed. Deaths of celebrities and major political events are often captured on its pages only few minutes after the corresponding event.

For example, during the recent swine flu outbreak, information about recent events was published in the "Swine Influenza" article on April 24th, just minutes after the first CDC public announcement, and a dedicated article was created on the following day. The "2009 Swine Flu Outbreak" article on Wikipedia received 1.5M visits during its first 5 days, with a peak of 417,200 on April 29th. Figure 1.1 shows the distribution of page views for the page "H1N1" for the month of April 2009.

This suggests that monitoring pages visits, creations and updates could offer an accurate picture of the most interesting current topics as perceived by the general

public [36].

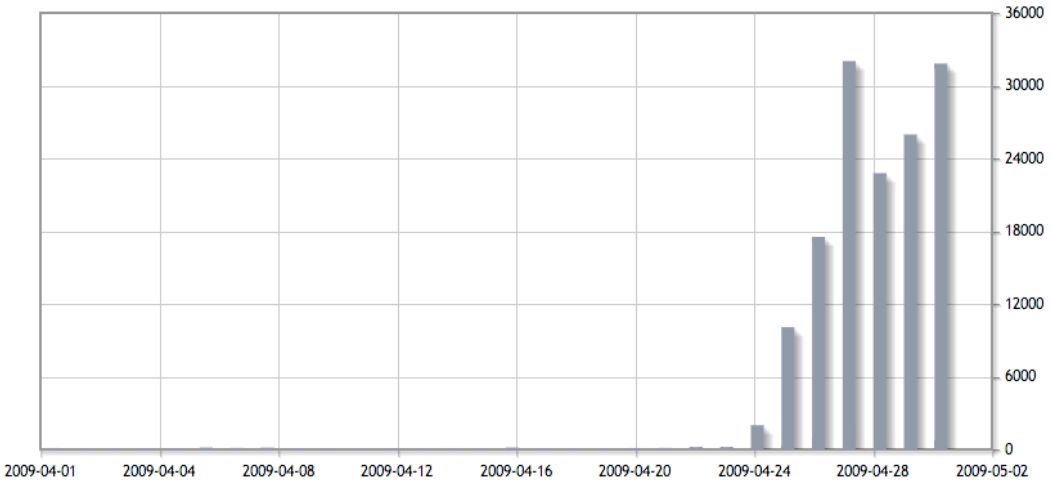


Figure 1.1: Number of Visits to "2009 Swine Flu Outbreaks" page on Wikipedia

1.4.3 Twitter

Twitter is an online social-network and microblogging platform which enables user to share short text messages. It was launched in July 2006 by Jack Dorsey, Evan Williams, Biz Stone and Noah Glass. After an initial slow adoption period, the service rapidly gained worldwide popularity, reaching 645 million registered users in 2014, who posted more than 500 million tweets per day [93].

Messages on the platform are called "tweets" and can be sent through the web, a mobile app or as text message. They are limited to 140 characters and can be shared privately or publicly. Users may "follow" other users to receive notifications of all their tweets. Subscribers are known as "followers" and this information is public in the sense that any user can see who is following whom.

Due to their nature, tweets can be quite noisy and contain many misspellings,

abbreviations and slang terms. In addition, Twitter users have developed or adopted special keywords and conventions in their messages:

- hashtags are words or phrases prefixed by a "#" and are used to represent the topic or type of a post, in such a way to make it easier to group them together;
- the "@" sign followed by an username is used for mentioning or replying to the other user; and
- tweets that start with "RT" are known as "retweets". They represent a form of endorsement of the original tweet by the author of the retweet, which decided to repost it to its own follower.

Hashtags that are used at increasingly greater rates than other hashtags are said to be "trending". Trending topics become popular either through a concerted effort by users, or because of an event that prompts people to talk about one specific topic. These topics help Twitter and its users to understand what is happening in the world.

Although the general public is increasingly careful about their privacy rights at home or over the phone, surprisingly many happily embrace online social networks and use them as an integral part of their daily lives, routinely sharing information about themselves, their mood, their activities, and so on. It is not uncommon to see lengthy conversations about a variety of current affairs on Twitter or Facebook. Whenever new gossip emerges or some public announcement is made, millions of messages are exchanged on social networks. For example, according to a report from Topsy Labs [87], in the first hour after the news of Whitney Houston's death surfaced, about 2.5 million related tweets were exchanged. Similarly, at the end of

the Super Bowl 2011, Twitter reported [92] that more than 4,000 tweets were being sent every second.

Example: 2007 California Wildfires In the fall of 2007, over 20 wildfires raged in California from Santa Barbara to San Diego, burning 500,000 hectares (approximately 1.25 million acres) and forcing large-scale evacuations. According to a survey of those affected, locals were not satisfied with the quality and quantity of information available from traditional media providers or authorities. Citizen reviews of the local news were better, but they complained that these providers were unable to keep up-to-date with rapid changes and were not accessible via TV or radio after evacuating the local area. Worse yet, the county emergency website was not able to handle the increased traffic and frequently crashed. Instead, several community websites emerged or changed their focus to aid residents. Rimoftheworld.net, a longstanding community website for residents in San Bernardino, allowed residents to submit news stories, discuss evacuation routes and fire prevention strategies on discussion boards, and post maps of the surrounding area. The administrators of the website worked with local firefighters and emergency services to circulate both official and eyewitness information. This emergency was one of the first to occur after the popularization of Twitter: in a survey of affected residents, 10% reported using the service for information, with most of these using the service for the first time. In particular, two San Diego residents dedicated themselves to gathering

information from all possible sources (e.g., friends, news, their own observations) and then posted all of their findings on Twitter. They provided unique and specific details by venturing around the city, taking photos of their friends' houses and listing inventory of local supermarkets, thus telling others where they could buy supplies. The importance of Twitter hashtag #sandiegofire came into focus during this event to aid those looking for information. Although many users began to adopt the convention, there was no clear consensus, and a number of different keywords emerged.

While gossip, news and general chit-chat represent a large share of the tweets sent every day, many people also use these social networks to share mood, feelings and personal concerns with their peers. During the H1N1 outbreak of 2009, for example, it was common to see people venting their theories or conjectures relating to public health announcements (e.g., "CDC Data Shows H1N1 Vaccine Perfect for Population Control" [13], their fear of having contracted the disease (e.g., "Im feeling sick... Hope its not H1N1"), and the panic-induced actions they were taking (e.g., "Got 2 facial masks but nobody around wearing it, gosh I don't want to attract the virus").

Example: 2009 Red River Flooding Several researchers took the opportunity to study Red River-related Twitter activity during the 2009 flood season. A detailed analysis [78] of over 7,000 Red River tweets for content and activity by Kate Starbird & Alexis Arbeit found that individuals made up the largest proportion of users (37%), but in terms of tweet volume, dedicated flood information accounts produced the most

[13] https://twitter.com/JohnMalcolmMe/status/13589474634104832

tweets (44%). Twitter activity was also affected by the public's risk perception, with tweet activity spiking when the threat was growing, and peaking when the risk was highest. The authors reported that while first hand information was the least common type of content (10%), derivative information, produced through a user-driven circle of information shaping and sharing using retweets, was the most common (over 75%). Synthetized information was tweeted most often by the media and was the second most popular type of tweeted content. Tech-savvy locals created a flood-service account that would automatically publish tweets whenever there was an update on the US Geological Survey website. The authors also reported that two main catagories of retweets emerged: general information with broad appeal, generally shared by those not directly affected by the flooding, and information that had local utility, always circulated by locals.

Example: 2008 Hurricane Gustav & Ike Hurricane Gustav and Ike occurred within one week of each other in the southern USA (August 25 and September 1, respectively). While in actuality neither hurricane rated on the same scale of destruction as Hurricane Katrina, residents and government agencies alike were concerned and the events highly publicized. A study [49] of hurricane-related tweets by Hughes & Palen found that activity spiked when the hurricanes represented the most danger (i.e., when the hurricanes made landfall). The author reported that a minority of users generated a large number of tweets, and that this percentage was constant across all events, suggesting that a few select users act as information hubs to disseminate information while

the majority are bystanders. In addition, the number of tweeted URLs was higher (in fact, almost double) during emergency events than at other times.

San Antonio-based market-research firm Pear Analytics analyzed [56] 2,000 tweets (originating from the US and in English) over a two-week period in August 2009 from 11:00 am to 5:00 pm (CST) and separated them into six categories as shown in table 1.2:

Type	**% of Tweets**
Pointless babble	40
Conversational	38
Pass-along value	9
Self-promotion	6
Spam	4
News	4

Table 1.2: Types of Tweets posted by users

Over the last few years, more and more tweets started embedding multi-media content, such as links to images, videos or news articles. A white-paper [61] by LTU Technologies (summarized in table 1.3) reports that 36% of tweets contain an image, 16% link to an article and 9% to a video.

With the increased diffusion of GPS equipped mobile devices more and more tweets started containing geolocation information in them. The location is often approximative (e.g., city level) but in some cases it contains the exact GPS coordinates of the author of the tweet and can be a great source of real-time of geolocated

information.

Link Type	**% of Tweets**
Images	36
Article	16
Video	9
Product	8
Front Page	7

Table 1.3: Types of Link shared on Tweets by users

1.4.4 Facebook

Facebook is perhaps the most well-known and widely used social network. It was founded in February 2004 by Mark Zuckerberg, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz and Chris Hughes. The service was initially limited to Harvard students but was then gradually extended to other groups [33]. By May 2005 it added support for more than 800 universities; in early 2006 also high schools students obtained access to Facebook. By September 2006 anyone with an email address or a phone number could join the social network. According to their official report [33], the service now has 1.31 billion monthly active users.

To share and access most of the content on Facebook users are required to register and create a personal profile. Various types of content can be added and shared through the profile, such as videos, pictures, songs and links. While not as popular as Foursquare's, Facebook also has checkin capabilities and users are free to augment their posts with the exact location.

On Facebook, users can enter mutual agreements of friendship. In general, friends of a user are allowed to see more content on a user's profile than other Facebook members. Most of the content shared on Facebook offers very detailed privacy options so users can decide if to share it publicly or only with some selected individual (e.g., their friends, a group, or a few manually selected users). Users may also join common-interest user groups, organized by workplace, school or college, or other characteristics, and map their friends onto (multiple) lists such as "work" or "acquaintances".

Users can endorse content on Facebook and around the web through the "like" feature, represented by the omnipresent blue thumbs-up button. Likes are used by Facebook to rank content internally and to prevent spam. The service also allows users to send messages directly to other users either as a live chat message or asynchronously, using an email-like mechanism. The Facebook authentication system can also be used to authenticate and login into other sites [34].

According to recent statistics [35] more than 4.75 billion content items are shared on Facebook every day, 350 million of these are photos, 70 million contain links and 200 million are messages sent to each other [79].

Example: 2007 Virginia Tech Shootings On April 16, 2007, a Virginia Tech student murdered two students then proceeded through the campus, shooting dozens of fellow students and professors, ending the crisis by killing himself. Before noon that day, 33 people were dead [19] and the community was both grieving their loss and frustrated with the University's lack of communication and inability to provide students with timely warnings during the crisis. Within a half hour of the last shooting, students began to post messages on Facebook asking if their

friends were okay. Within 90 minutes, the first Wikipedia page on the tragedy was published and the Facebook group "A Tribute to those who passed at the Virginia Tech Shooting" was created. Shortly thereafter the "I'm Ok at VT" Facebook group started, encouraging students to check in and let others know they were safe. All three became central sources of information for the next 24 hours as students worked together to determine the names of the victims. Students shared what they found while other members would ask for verification and attempt to cross-reference with other sources. As a result, the communities were self-correcting and established reporting norms (e.g., students had to explain their relation to the deceased or information source). Vieweg et al. studied [101] the public's use of the Facebook and web 2.0 technologies to deal with information dearth, generate and disseminate information, and conduct collective problem solving. The study found that the online community was able to accurately compile the names of all 32 victims before Virginia Tech officially released their list.

Example: 2010 Haiti Earthquake Similar to other examples of crowdsourced photojournalism in disaster areas, moments after a catastrophic magnitude 7.0 earthquake struck Haiti, affected citizens were using their mobile phones to take photos of their plight and distribute them via Twitter. For some Haitians who lost their phone landlines, Facebook was the only way to communicate their status to loved ones and learn about the fate of others. According to a study [81] from Sysomos, over 2.3 million tweets were sent between January 12 to 14 and over 1,500 Facebook status posts per minute contained the word

> "Haiti." Amatorial relief websites by both Haitians and tech-savvy volunteers quickly sprung up to offer aid. Mobile giving had been piloted in other disasters, but the 2010 earthquake was the most successful to-date. In response to the tragedy, Twitter updated their official blog to announce a Red Cross texting campaign [90] where users could text "HAITI" to donate $10 to the relief effort. Within 48 hours, over $3 million dollars had been raised, thanks in large part to viral dissemination via Twitter.

Facebook's posts could be a great source of real-time information. Public posts could be used to identify trends and monitor public opinion about certain topics, likes and shared links could indicate interesting topics or hidden communities, and the geolocation information attached to the posts could help build models of how users travel.

1.4.5 Flickr

Flickr[14] is an image and video hosting website launched in 2004 by Stewart Butterfield and Caterina Fake, and acquired in March 2005 by Yahoo. As of 2013 the service had more than 85 Million members [1] and more than 6 Billion photos uploaded.

The service allows users to upload photos and videos into various "photostreams", that can then be organized in albums. Each photo can belong to multiple albums and may have a title, a description, some tags and EXIF data attached to it. These metadata will be indexed by Flickr's internal search engine if their owner consents. Pictures on the service can be kept private or made public, specifying the

[14] www.flickr.com

type of licensing.

On Flickr users can follow each other's photostream and can join groups. Groups generally have an associated pool of photos, to which each member can contribute.

Example: 2004 Indian Ocean Earthquake & Tsunami Mobile phone technology, blogging, and photo-sharing were at the forefront of the Indian Ocean disaster. Cell phones with cameras, then a novel technology, were widely used to capture images of the devastation and citizens shared them with the world before the mainstream media could respond. The tsunami was the first instance of disaster-related activity on Flickr and photo groups were created to share news, strengthen the community, document history, educate distant observers, and rally for aid. Mobile phones were also heavily used for texting for help and locating survivors as phone landlines were down and voice calls were often dropped due to high bandwidth use. Public blogs also played an unprecedented role. The "Southeast Asia Earthquake & Tsunami" blog [15] was launched by 3 individuals to provide aid, news, and information about family members to affected people. The blog also allowed visitors to post their needs or what help they could offer. A list of confirmed deaths, image galleries, and links to aid agencies were also constantly updated. The blog was so successful that it reached 1.1 million hits within 10 days of its launch; a worldwide blogging landmark.

[15] tsunamihelp.blogspot.com

1.4.6 FourSquare

Foursquare is a location-based social networking website for mobile devices. Users "check in" (or *checkin*) at the venues they visit to earn points, badges and discover friends who are already there. The service was created in 2009 by Dennis Crowley and Naveen Selvadurai as the second iteration of a similar idea (named Dodgeball) that they built a few years before and sold to Google. As of January 2014, FourSquare has 45 million users and receives more than 5 billion checkins every day [40].

While the user base of FourSquare is equally distributed among genders, a 2011 study [103] conducted by the Wall Street Journal reported that only 38% of Foursquare checkins are made by women. In the study, the top checkin categories (figure 1.2) are Corporate Offices and Homes, followed by Coffee Shops, Bars, Gyms, Grocery Stores, Parks, Restaurants and Transportation.

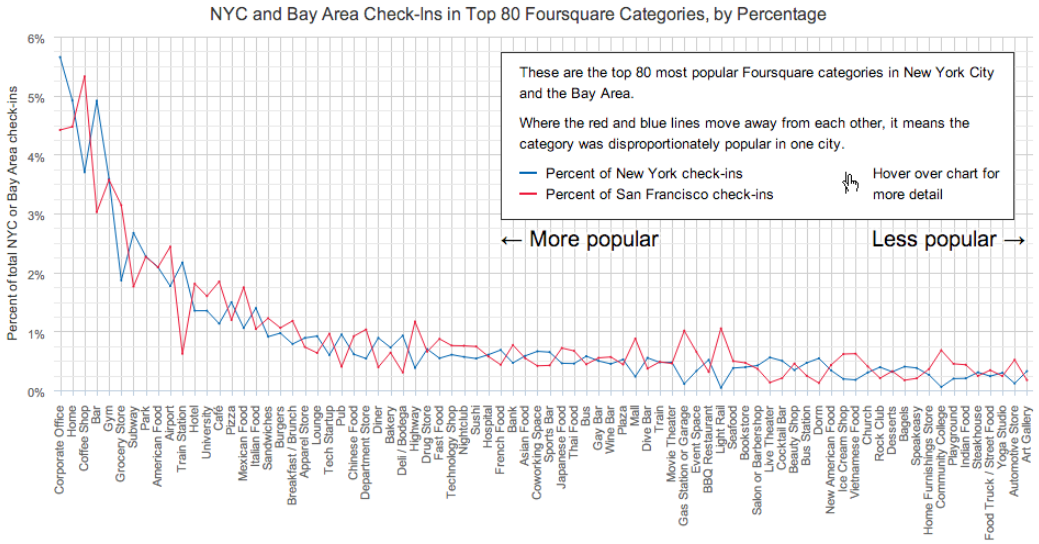


Figure 1.2: Top categories of Foursquare checkins

To checkin into venues, users need to have GPS enabled devices and must be in close proximity to the intended checkin location. Each time the user checks into a place, they receive points and can check their standing against friends on a leaderboard. There are more than 100 reasons Foursquare awards points. Some of the most commonly awarded ones are shown in 1.4.

Activity	**Value**
Checking into a new venue	3 points
Becoming the Mayor of a venue	5 points
First of your friends to checkin at a venue	3 points
Checking into a new category	4 points
Checking into a known venue	1 points

Table 1.4: Values of commonly awarded checkins on Foursquare

Foursquare users are encouraged to be hyper-local and hyper-specific with their checkins, including the area of a building (e.g., the terminal in an airport) or the specific activity performed while at a venue (e.g., listening to a concert). Users can create a "to do" list for their private use and add "tips" to venues that other users can read, which serve as suggestions for things to do, see or eat at the location.

Together with points, users can earn badges (e.g., the ones shown in figure 1.3) for their checkins. Badges are awarded at checkin for locations with certain tags, for checkin frequency, or for other patterns such as the time of checkin. There are a handful of introductory badges that are earned as milestones in usage. Some badges can only be earned in a specific city or event, and others can only be earned if the venue has certain tags. A notable one is the NASA Explorer badge, unlocked

on October 22, 2010 by astronaut Douglas H. Wheelock checking into foursquare from the International Space Station.

Figure 1.3: Examples of Foursquare Badges

Another form of recognition and gamification introduced by FourSquare is the concept of "mayorship" of a venue: the user who has checked in to a venue on more days than anyone else in the past 60 days, and the checkins are valid under Foursquare's time and distance protocols, will be crowned mayor. Being the "mayor" of a venue is mostly a vanity thing, but in some occasions it allows to unlock discounts for services at the venue (e.g., the "barista badge" of Starbucks provides discounts at the venue [18].

Checkins can be a great source of real-time data on user habits and their movements inside and across cities. According to an infographic [39] published by Foursquare, a big percentage of user checkins are in travel-related venues (e.g., train stations, airports, subways). The same report shows how the top states for gym checkins are California, Illinois, Minnesota, New York and Washington. Chapter 4 shows some possible uses of these data to create a travel model in cities and across the United States.

1.4.7 Other Sources of Data: Proxy & Search Logs

While according to a recent study [26] the majority of Internet users are registered and frequently make use of some social media application or site (e.g., Facebook,

Twitter, Blog), a small percentage of the population still does not. For example, only 32% of the population aged 65 and over use social media sites.

Fortunately, even these less engaged Internet users use the Internet to find answers to their health questions and looking at their traffic patterns could provide interesting insights. For example, health-related websites like WebMD and MayoClinic will receive a higher-than-usual amount of traffic in time of crisis or when a pandemic warning is in effect (e.g., traffic on the Wikipedia H1N1 page spiked during the 2009 pandemic). People who fear they may have contracted the illness will probably visit such sites to compare their symptoms with the ones reported there. In addition, official health outlets (e.g., CDC.gov) will also be frequently checked for updates by concerned users.

Closed communities (e.g., companies or university dorms) and other large private networks (e.g., Verizon's mobile Internet network) use various level of proxy servers and firewalls to regulate traffic. All traffic generated by the users of the network goes through these servers and a log of who accesses what may be easily kept by each server. Having access to these logs could provide interesting data to monitor the response of the population to certain news. The analysis of these logs might also support the study of correlation between user's behavior and medical data on a variety of topics.

In addition to the page visited, knowing what people look for on search engines for can provide interesting insights on their health interests and fears. Today people rely more and more on the results provided by search engines to accomplish many tasks, even not strictly related to the web. For example, almost all the current search engines allow users to discover the current time in various cities of the world (e.g., search for "time in Rome, Italy" on Ask.com) as well as movie theater listings (e.g.,

search for "80302 movies" on Google) or the correct spelling of a word (e.g., search for "analyzing" on Yahoo!). As the reach of the Internet grew, people also started using web search engines as substitutes for their medical encyclopedias and updated information on health questions. A recent PEW Research [66] reported that 77% of participants initiated their health-related investigation on a search engine.

Similarly to the traffic that goes through proxy servers and firewalls, all the queries submitted to a search engine are aggregated and saved for later analysis in databases which are commonly referred to as "query logs". Over the past few years, query log analysis generated many interesting studies in a broad range of fields.

1.4.8 Privacy Concerns

While the broad availability of customer data and the recent improvements in data mining techniques please marketers and companies, they raise many privacy concerns among users and customers. The idea that so much data has been collected about one's activities and that all these data sources could potentially be linked together to produce an accurate and complete picture of each user is definitively raising increasing concerns.

In her 1998 report [9] Ann Cavoukian, Commissioner for the Ontario Information and Privacy Committee, claimed that data mining "may be the most fundamental challenge that privacy advocates will face in the next decade". In her report, she recommends that, at the moment of purchase, customers be given a choice among 3 levels of opt-out policies:

1. Do not allow any data mining of user's data;

2. Allow data mining only for internal use only; and

3. Allow data mining for both internal and external uses.

In addition, if we start relying on social media analysis for disease surveillance, sophisticated solutions may need to be built to filter out bad and misleading data produced intentionally to confuse the system. Hackers and scammers have been creating fake accounts on social media sites for years to post links and fake news in the hope to increase the visibility of their product or scam. Similarly, even a non-malicious user, could be just trying to having fun with the system. For example, creating a new account on Twitter is easy, resulting in many attempts to spoof accounts created for important individuals and well known businesses. To overcome this problem, in 2008 Twitter launched a verification program to allow celebrities and companies to certify that the accounts bearing their names were real, that is, actually under that individual's control.

Privacy concerns are even more pressing when dealing with medical data, since a data leak or massive data aggregation could influence an individual's insurance status. In today's healthcare, lab results, scans, diagnosis, etc, are all stored in digital format in a variety of data centers. Assembling a complete picture of an individual's condition requires that these data be aggregated.

The Health Insurance Portability and Accountability Act (HIPAA) sets standards [24] for how these data should be stored, transmitted and accessed. It separates Personal Identifiable Information (PII) and Personal Health Information (PHI), allowing the transmission of the latter for research purpose as long as users are not identifiable. It also identifies who is subject to these regulations and in which capacity, and requires the provisioning of emergency access plans in case of breach or natural disaster.

While this is great for the patients, it can also raise lots of security concerns.

Hundreds of papers and books have been published in recent years just on this topic, with the aim of exposing the flaws of the system and increasing the confidence in data mining techniques with solutions that allow anonymous aggregation of the data while preserving its important properties. Most of the solutions proposed, as for example the one published by Segre & al., take advantage of cryptographic algorithms to scramble identifying fields while still allowing statically useful data analysis [75].

1.5 New Technology and Disease Surveillance

Although the introduction of NEDSS improved the effectiveness of US health care surveillance systems, it still relies on a small number of humans (e.g., doctors or nurses) to manually report the cases of diseases and conditions they encounter. Whether submitted using paper forms or electronically, the system relies heavily on each medical office's effort to promptly transcribe and report their cases. Doctor's offices are notoriously understaffed, especially when economic conditions are poor. The use of computers and digital communications helped relieve some of this pain, but we are still far from what it could be achieved even just fully exploiting existing technologies.

The collection of disease-related data and their correct reporting typically relies on the use of surveys, with all of the problems generally associated with this instrument (poor return rates, slow turnaround, etc.). For influenza like illnesses (ILI), for example, it is ultimately up to the health care provider to determine if a specific case meets public health surveillance case definition, to fill out the appropriate forms from the CDC and the provider's respective state health department, and then to actually submit the report.

Nevertheless, it is important to realize that many people never actually consult a doctor for what they perceive to be common or minor health problems. In fact, a Consumer Health-Care Product Association survey [20] reported that nearly 80% of Americans relied on over-the-counter medications to treat a personal condition and that 73% would rather treat themselves at home than see a doctor. For these reasons, it is very likely that many potentially interesting diseases and conditions will remain unreported and thus potentially undetected.

Similarly, surveys done to gather information on the effectiveness of a health campaign are often conducted over the phone by calling an appropriately sized random sample of home phone numbers. This is a difficult and lengthy manual process, subject to all of the sampling and response problems generally associated with political polling which may in turn lead to unreliable results. For example, thanks to the diffusion of Caller ID technology, many families screen their calls, while others simply disconnect as soon as it is clear they are being asked to participate in a survey. In addition, because many people work outside the home, it is difficult to time calls correctly.

For the next step to happen it is necessary that all the parties involved develop a consensus on critical surveillance content and commit to transform surveillance from laboriously collected archives of after-the-fact statistics to meaningful real-time monitoring of the status of local health, capable of providing real-time early warnings for potentially devastating outbreaks.

Information technology and informatics can help on this front. If data standards will be fully developed and adopted, so that various systems can integrate, data collection for surveillance could be highly automated by the same electronic systems already used to support clinical care. Whenever a health event occurs (e.g.,

death, disease, or injury) a message could be sent to the responsible public health department including all the necessary information on the provider, the patient's details (e.g., name and home address) and its health records (e.g., immunizations, treatments, risk factors). New algorithms can then be developed and applied to the data collected to determine wherever an alert should be sent, what is its priority, and when it should be escalated to national (or even international) level.

Even before this ideal capacity becomes widespread, technology such as cell-phone based systems could accelerate the collection and transmission of important health data (e.g., fever outbreaks). In the last years the adoption of cell-phones in developing countries has been extraordinary and wireless networks may soon be able to replace the need for expensive landline-based systems. Wireless Internet access can facilitate the communication between local and district health posts and allow to expand their knowledge and capacities, for example, through telemedicine. Since almost all recent wireless devices are capable of accurately pinpoint their location through the Global Positioning System (GPS), up-to-date maps can be generated in real time describing local health status.

Example: 2003 SARS Epidemic The 2003 SARS epidemic in China occurred just prior the widespread use of many existing social media platforms. Instead, cell phones played a major role in public communication during this health emergency. Due to China's strict censorship policy little information regarding an increasing number of "atypical pneumonia" cases was released, but the people of Guangdong province were aware of SARS and the potential problem before the mainstream media as the number of text messages sent in the area sky-rocketed [114] in the days leading up to the Chinese governments official report.

> In February, a media and Internet blackout on SARS was enforced across China and news providers did not report on or acknowledge the existence of the disease. Without any means of acquiring or verifying information, the public began to circulate texts regarding SARS outbreaks, folk remedies (most of which were inaccurate, e.g., drinking teas and vinegars), and rumours. Cell phone applications were also built to help the public battle SARS. Sunday Communications, a cell phone service provider, allowed subscribers [111] to receive alerts by text if they were within one kilometre of an infected building in Hong Kong. In other cases, dissatisfied computer-savvy Chinese citizens created independent websites (e.g., sosick.org) listing areas of suspected or confirmed SARS cases.

1.5.1 Related Research

Traditional surveillance systems like those deployed by the CDC mostly rely on physician visit data (i.e., billing information) to estimate the prevalence of Influenza-Like Illnesses (ILI) and other diseases. Billing data is generally considered to be accurate, largely because real dollars are at stake. These data are collected and then aggregated to give a national picture within a 2-3 weeks reporting time lag.

In an attempt to reduce this time lag , in 2003 Espino, Hogan and Wagner suggested [31] using received call volume on telephone triage advice lines as a proxy for surveillance data. Their study demonstrated good cross correlation between calls made to emergency rooms and doctor's after-hours lines and the ILI percentage data published by the CDC 1-5 weeks later. In a similar fashion, Magruger reported [59] how over-the-counter drug sales volume can be used as an early estimator for physician visits. In his research, he demonstrated a correlation between flu remedy

sales data and doctor visits due to influenza, and between chest-rub sales data and cases of bronchitis. Unfortunately, the average lead time for sales data was only 3 days, still not enough to make these results useful for purposes other than optimizing drug distribution and supply.

With the wide-spread adoption of cellphones and the Internet, many people started using search engines to find information about specific diseases or medical problems. According to a PEW report [66] more than 85 million Americans look for health information online every year.

In 2004, Johnson & Al. demonstrated [54] how health website access logs data correlated to official ILI reports (but, unfortunately, still lacked sufficient timeliness). In 2006, Eysenbach introduced [32] a novel approach to flu surveillance using related query volume on search engines. Since query logs were not available, the researchers bought Google Ads keywords for "flu" and "flu symptoms" obtaining detailed statistics on the weekly volume of queries performed and correlated these with officially reported Canadian ILI percentage.

In 2008 a study [68] conducted by Polgreen & al. using Yahoo! search queries confirmed these results. In their study, the authors studied the correlation between the percentage of ILI-related queries and official CDC data, developing a linear model which allows to predict influenza outbreaks 1-3 weeks in advance. A similar model was also developed to predict an increase in mortality attributable to pneumonia and influenza up to 5 weeks in advance. In both experiments, the queries used were identified by the presence of a few specific influenza-related terms.

Google Flu is the best-known query log analysis effort. In their paper [42] the authors analyzed hundreds of billions of queries contained in 5 years of Google query logs. The query logs were anonymized, but information about the location of

the users (obtained through geo-location of the source IP address) was retained to provide localized statistics. Flu-related queries were automatically identified by an automated classifier when performed on their system on each particular location. The results obtained during their experiments were validated against official CDC data on Influenza-Like Illness doctor visits. During their experiments, the authors identified 45 search queries which are significantly more useful in predicting the number and location of ILI-visits as depicted by CDC data. These queries were then used to to develop a linear model using weekly ILI percentages between 2003 and 2007. The model was able to obtain a good fit with CDC-reported ILI percentages with a mean correlation of 0.90. The model was also validated against a previously untested data from 2007 through 2008 and showed a mean correlation of 0.97. Data from the state of Utah allowed the authors to test the model on a more local scale, obtaining a mean correlation of 0.90.

During the H1N1 outbreak of 2010, Chew & Al. performed a detailed analysis [15] of more than 2 million tweets containing the keywords "H1N1", "swine flu" or "swineflu". Their study demonstrated how Twitter content could be used by health authorities as a complementary source of information to monitor public opinion and respond to public concerns in a timely fashion.

Similarly, in 2011, Chumara & Al. demonstrated [16] how informal media (e.g., news, blogs and tweets) could be used to monitor and make predictions during the 2012 Haitian cholera outbreak. Since data from those sources is typically available 2-3 weeks in advance of official reports, they represent an alternative to official data and allow for faster and more effective resource deployment.

1.5.2 Social Media for Disease Surveillance

The main goal of this work is to demonstrate how the analysis of social media data can effectively be used to make predictions on health related topics. In this thesis we show how the analysis of the content of Twitter posts allowed us to monitor and track public perception of the 2009 H1N1 flu epidemic. We also demonstrate how the same methodologies could be used to monitor and accurately predict flu trends at national and regional levels, a significant improvement over current health practice. Finally, we show how geolocated social media posts (e.g., Foursquare checkins) can be used as an effective but inexpensive source of data to create accurate travel models, that can help us to more accurately predict flu trends at city level.

CHAPTER 2
RESEARCH APPROACH AND METHODOLOGIES

The advances in computing and networking technology of the last years make it possible to collect large quantities of data describing complex social systems which, until only recently, were simply too large to store and too complicated to analyze. Inexpensive computational power makes it possible to create and calibrate detailed system models from these data. Such models generally describe systems using mathematical constructs. They are commonly used in engineering disciplines (e.g., computer science), the natural sciences (e.g., physics) and the social sciences (e.g., economics) to estimate and predict how a system under study will behave under certain conditions. If properly validated, such models can also be used to test hypothesis and perform simulations.

Many medical fields have also greatly benefited from mathematical modeling: researchers almost completely rely on models and simulation to perform their work. Epidemiologist, for example, use parameterized disease models to predict how quickly and broadly outbreaks may spread and help prepare the appropriate medical response. These models can be of great benefit to clinicians by providing the foundation for planning and cost-effectiveness analysis of health measures. With the widespread adoption of text-messaging, instant-messaging, and social networks like Twitter and Facebook, mathematical modeling can also be applied in social contexts. For example, health organizations may want to create models to devise the most effective strategy to implement awareness campaigns and predict their impact. Upon deployment, models may help to devise the optimal sampling strategy necessary to accurately estimate the effectiveness of the campaign while minimizing

cost of sampling.

Disease models are parameterized mathematical representations of clinical conditions, intended to summarize what is known about the disease epidemiology, prevention and treatment. Parameter values are generally fit from historical disease data when such data are available. Unfortunately, such data are not always available for new situations or emerging diseases. In these cases, models are often generated using generic parameters taken from similar diseases. As the scope of a model increases to include, for example, epidemiological factors, modelers also need to gather current data about, e.g., a population, its mobility and the surrounding environment so as to provide appropriate model parameters. Interesting data may include, for example, the size of the population, the average distance traveled, the health status of each individual and the weather conditions. These data are used to fit the parameters of the model, as well as to validate the model's accuracy.

The spread of the reach of the Internet and the increase of social web activity could represent a good supplement to official data. On a daily basis, millions of social network status updates, blog posts and search queries travel through the network. In these messages, people express their feelings, look for solutions to their problems, or seek suggestions from peers. Monitoring and analyzing these data could provide hints on the perception and mood of the public with respect to certain health subjects, as well as clues to new and potentially unreported outbreaks.

2.1 Twitter

The primary data source for this work is Twitter, the real-time micro-blogging site. With more than 500 million [93] posts per day, the analysis and classification of this real-time stream of information could be very useful to monitor and detect early signs of diseases outbreaks as well as to measure public perception of disease-related

topics.

2.1.1 Anatomy of a Tweet

Tweets are short text messages exchanged on the social network. The length of each tweet is limited to 140 characters due to the original limitation of GSM SMS text messages. They can be posted publicly or kept private and thus visible only to the followers of the author. At minimum, each tweet contains the *username* of the author, the *timestamp* in which it was sent and the *text* of the post. Figure 2.1 shows some examples from notable accounts:

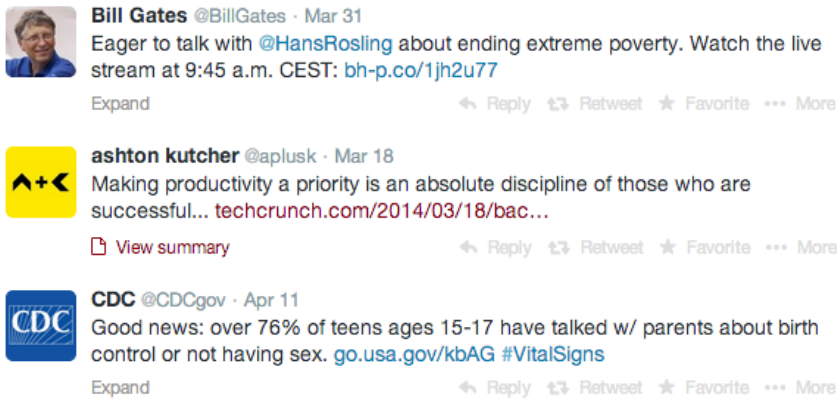


Figure 2.1: Example of Tweets

It is common to find special keywords in tweets. The most popular ones are user mentions and *hashtags*. The former are generally usernames prefixed by a commercial at (e.g., @a_signorini) and their purpose is to get the attention of a particular member. *Hashtags* are words or unspaced phrased preceded by a pound sign (e.g., #awesome). These keywords were initially used as references, so that all posts around the same topic (e.g., #SXSW2014) could be quickly found through search,

but in the last years they are used more liberally often even in place of actual sentences (e.g., #missingyou instead of "I miss you"). See figure 2.2 for an example.

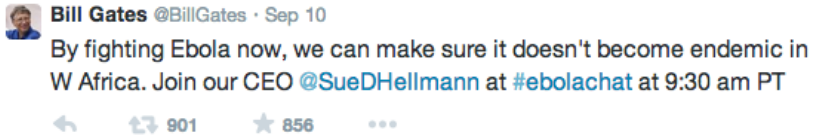


Figure 2.2: Use of Hashtags in a Tweet

What is actually shown by the interface on Twitter website is just a small sample of the information embedded in each tweet. A much better way to consume the stream of tweets is use one of the many APIs available. For example, each tweet retrieved through the interface includes the complete profile of its creator, completed of name, location, language, date of creation, timezone, number of followers, background color, etc.


```
"user": {
  "profile_sidebar_fill_color": "DDEEF6",
  "profile_sidebar_border_color": "C0DEED",
  "profile_background_tile": false,
  "name": "Twitter API",
  "profile_image_url": "http://a0.twimg.com/...74872/7normal.png",
  "created_at": "Wed May 23 06:01:13 +0000 2007",
  "location": "San Francisco, CA",
  "follow_request_sent": false,
  "profile_link_color": "0084B4",
  "is_translator": false,
  "id_str": "6253282",
  "entities": {
    "url": {
      "urls": [
        {
          "expanded_url": null,
          "url": "http://dev.twitter.com",
          "indices": [0, 22]
        }
      ]
    },
```

```
    "description": {
      "urls": []
    }
  },
  "default_profile": true,
  "contributors_enabled": true,
  "favourites_count": 24,
  "url": "http://dev.twitter.com",
  "profile_image_url_https": "https://si0.twimg.com/...74872/normal.png",
  "utc_offset": -28800,
  "id": 6253282,
  "profile_use_background_image": true,
  "listed_count": 10774,
  "profile_text_color": "333333",
  "lang": "en",
  "followers_count": 1212963,
  "protected": false,
  "notifications": null,
  "profile_background_image_url_https": "https://si0.twimg.com/.../bg.png",
  "profile_background_color": "C0DEED",
  "verified": true,
  "geo_enabled": true,
  "time_zone": "Pacific Time (US & Canada)",
  "description": "The Real Twitter API. I tweet ... service issues ",
  "default_profile_image": false,
  "profile_background_image_url": "http://a0.twimg.com/.../theme1/bg.png",
  "statuses_count": 3333,
  "friends_count": 31,
  "following": true,
  "show_all_inline_media": false,
  "screen_name": "twitterapi"
}
```


The Twitter API also detects *entities* (e.g., places, names, companies) in the text of the tweet as well as user mentions and *links* (i.e., URLs), and breaks them out in a separate section.

Given the restriction on the number of characters, links in tweets are generally shortened through some service before being used in the text with the aim to save precious characters. These services generally have very short domain names (e.g., bit.ly) and create a short hash (e.g., abcde) of each URL which redirects (e.g., `http://bit.ly/abcde`) to the full form (e.g., `http://www.cnn.com/politics/`

`20140911/test.html`) when visited. Many URL shortening services (e.g., Bit.ly[1], TinyURL[2], Bit.do[3]) have been launched in the past years, and big companies like Google (goo.gl) and Twitter (t.co) launched their own. While providing a useful service to the users, these services collect important statistics on the most visited and clicked links, which are often used while ranking the importance of a page (e.g., for Google) or a post/tweet (e.g., for Facebook or Twitter). Twitter API results presents URLs (e.g., `http://bit.ly/abcde`) both in their shortened and expanded form.


```
"entities": {
  "urls": [
    {
      "expanded_url": "https://dev.twitter.com/terms/display-guidelines",
      "url": "https://t.co/Ed4omjYs",
      "indices": [76, 97],
      "display_url": "dev.twitter.com/terms/display-\u2026"
    }
  ],
  "hashtags": [
    {
      "text": "Twitterbird",
      "indices": [19, 31]
    }
  ],
  "user_mentions": []
}
```


Other special fields in the tweet data indicate if the tweet has been *retweeted* or *favorited* by the user or if it is a *reply* to some other tweet.


```
{
  "coordinates": null,
  "favorited": false,
  "truncated": false,
  "created_at": "Wed Jun 06 20:07:10 +0000 2012",
  "id_str": "210462857140252672",
  "in_reply_to_user_id_str": null,
```


[1] bit.ly
[2] tinyurl.com
[3] bit.do

```
  "in_reply_to_screen_name": null,
  "source": "web",
  "in_reply_to_status_id": null
}
```


Retweets are form of public endorsement of the content of a tweet, where a user finds the content of a post so valuable to be willing to share it with its own followers, while attributing the credits to its original author. Originally, *retweets* where a user created movement and their content looked just like any other tweet to Twitter's API. For this reason, users generally prefixed *retweets* with the keyword "RT". In the recent years Twitter's interface changed and a tag clearly marks retweets, making the practice unnecessary (see Figure 2.3). In addition, Twitter realized that *retweets* are an important metric to assess the quality and importance of a tweet and created additional attributes (e.g., `retweet_count`) to capture these data.

Figure 2.3: Example of a Retweet

Another form of endorsement for the content of a tweet is to *favorite* it. This form of endorsement is private and only the author of the tweet and the user who favorited it know about it (see Figure 2.4). User may *favorite* a tweet to privately demonstrate support to its author without having to write a full reply or retweet (e.g., similarly to what a Like does on Facebook) or to simply flag so it can be easily found later. Twitter's interface and API provide special endpoints [95] to retrieve all the tweets favorited by the authenticated user. Similarly to retweets, the number

of *favorites* can be used as an important metrics for the quality and impact of a tweet, and is captured in the attribute `favorite_count` in the latest version [94] of Twitter's API.

Figure 2.4: Example of Favorites for a Tweet

Finally, the current Twitter interface (also the one of many apps) make it very easy for a user to *reply* to a user or in response to a specific tweet. *Replies* are simply public direct messages, which *text* generally starts with the handler or the recipient (see Figure 2.5). Whenever a *reply* is initiated through an interface, the information about who is replying to what is captured in special fields (i.e., `in_reply_to_screen_name` and `in_reply_to_status_id`). This data help capture relationships between tweets and users, allowing Twitter to discover communities of users and suggest other users to follow.

More recently, many Twitter clients started exploiting the GPS capabilities of modern mobile devices by including geographic *coordinates* in tweets, effectively pinpointing the exact location from where the message originated (as opposed to simply where the user typically resides).

```
"coordinates": {
    "coordinates": [-75.14310264, 40.05701649],
    "type":"Point"
```

Figure 2.5: Example of Direct Reply on Twitter

}

The user has control over the accuracy of the location embedded in a tweet. By default the location is provided at the neighborhood level, but it can also be abstracted to city, state or country. In some cases, for example when tweeting from a specific location (e.g., a restaurant), exact coordinates can be embedded in the tweet. This is especially true for all the services that use Twitter to publicize the activities of the user, such as Yelp[4], FourSquare, RunKeeper[5], etc.

2.1.2 Twitter's API

When the data collection process for this project started, the only interface available to programmatically gather tweets was the REST search interface [97]. Using this interface to query Twitter's search engine at regular intervals allowed us to retrieve the last 20 tweets containing at least one of the keywords mentioned in the query.

[4]yelp.com
[5]runkeeper.com

The search interface had some limitations on the length of the query and the allowed frequency of the requests [96]. Generic keywords returned too many results for each search and, given the necessity for a delay between each successive query, this would cause the system to miss important tweets. To improve our coverage we optimized the query interval and the keyword distribution in an attempt to capture all the matching tweets. Unfortunately, this search-based approach combined with the (then unknown to us) caching mechanisms of Twitter's server could not guarantee a complete sample.

In October 2009, Twitter released the first version of its Streaming API [89]. This interface supports opening a single connection to Twitter's servers and receiving a continuous stream of all the tweets matching certain conditions, or *filters*. These filters can be written so that the connection will return all the tweets matching certain keywords or authored by certain users. This API is both efficient and effective in capturing all the posts relevant to a certain domain (it also offers a "sample" endpoint that returns a random fraction of all the tweets posted, independent of domain). About one year later, in September 2010, Twitter introduced a location-based (i.e., geographic) filter [91] in the Streaming API. Thanks to this new filter it is now possible to retrieve all the tweets authored in a particular area of the world, described by a rectangular bounding box defined by the coordinates of its corners.

As with the search interface, "rate limiting" also applies to the Streaming API. If the filters applied are too broad, rate limiting messages will appear in the stream of tweets to inform the developer that their feed is incomplete. Clever use of filters can probably allow most research-related use of the data stream with just a free account, but more comprehensive data gathering will require special business

agreements with Twitter. In particular, the company offers a "firehose" access level which allows to receive all the tweets exchanged on the network.

2.1.3 Data Gathering and Normalization

We started collecting data from Twitter in March, 29th 2009, right at the beginning of the H1N1 epidemic. Given our relatively narrow interests, we used a predefined set of flu-related queries, intending to gather all the tweets related to our particular domain. Our keyword set contained generic terms such as "flu" and "influenza", but also more H1N1-specific keywords such as "swine flu", "h1n1" and "pneumonia." Since we were also interested in monitoring the public's response to the media announcements we also started tracking correlated keywords. Terms like "purell", "hand sanitizer", "hand washing" and "surgical masks" allowed us to track messages connected to the fear of contracting the disease, while terms like "relenza", "zanamivir", "tamiflu", "amantadine", "rimantadine" and "oseltamivir" allowed us to track posts related to possible cures.

On October 1st 2009, we switched to the Streaming API and added a few more generic keywords such as "infection", "sick", "headache", "hospital", "shortage", and a couple very specific one ("guillain" and "barre"). In addition, on October 20th, 2009 we also started retrieving a random sample of all published tweets (independent from our keywords) through the "sample" endpoints of the Streaming API.

Finally, when the location based filter became publicly available we started retrieving a third stream of tweets sampled from all tweets sent from coordinates within a bounding box (north-west corner -124.80:22.02, south-east corner -66.18:49.72) loosely defined to enclose the United States (see Figure 2.6).

Since tweets are generally generated by users, their content is often quite messy.

Date	Search Keywords
March 29th, 2009	flu, influenza, "swine flu", h1n1, pneumonia
April 10th, 2009	purell, "hand sanitizer", "hand washing", "surgical masks"
April 17th, 2009	relenza, zanamivir, tamiflu, amantadine, rimantadine, oseltamivir
October 1st, 2009	"infection", "sick", "headache", "hospital", "shortage", "guillan", "barre

Table 2.1: Search Keywords used as filters in Twitter's API

Figure 2.6: Area used as geographical filter for Twitter

For example, some tweets contain non-ASCII characters, and others contain tweet-specific jargon keywords (e.g., "RT" for re-tweet), hashtags (e.g., "#awesome"), usernames (e.g., "@a_signorini") or links. To make sure to have a clean and processable dataset, we applied the following cleaning steps to each tweet:

1. as long as tweet contains HTML entities, decode them;
2. if twitter contains non-ascii characters or is shorter than 5 characters, discard the tweet;
3. lowercase the tweet;
4. then remove:
   (a) all numbers and terms starting with numbers;
   (b) all terms starting with a protocolo (e.g., "http://"); and
   (c) all terms starting with @ and # ;
5. ignore the tweet if less than 5 characters.

Once the text was clean, we obtained a set of terms splitting at white spaces. Because tweets are often sent from mobile devices, there is an increased probability of introducing typos and misspellings. During our experiments, we removed all terms which occurred very rarely (less than 5 times) or were shorter than 3 characters.

2.1.4 Stemming

To further clean our dataset, we applied stemming to the remaining terms. Algorithms for stemming have been studied in Computer Science since 1960s and are

now quite commonly used in information retrieval. Their main objective is to reduce each word to its root (e.g., "sickest" to "sick"). The most basic algorithms are simply lookup tables pre-populated with popular words and exceptions. These are fast and easy to understand. Another class of algorithms performs suffix-stripping and rely on a small set of rules (or steps) that are applied on the term to find its root form. The most famous algorithm of this type was invented by Porter in 1980 [70].

Lemmatization algorithms perform part-of-speech analysis to reduce the subset of rules that can be applied to the term. This class of algorithms is generally an improvement over suffix-stripping, but any mistake in the identification of the correct category for the term limits the added benefits of lemmatization. An example of this class of algorithm is the Y-Stemmer [113].

Finally, stochastic algorithms involve the use of probability to determine which is the most probable root form of a word. These algorithms are generally trained on some input data from the language of interest to form the probabilistic model, and then produce a probable root for each word in the test set.

While many other sub-classes of stemming algorithms exist (as well as some hybrid approaches, especially for Arabic text [25], many architectures simply use the well-known Porter Algorithm. Many implementations [65] of this algorithm exists (some even written by Porter himself) but the generic English version is almost always based on performing 6 basic steps [77] on each term:

1. strip plurals and -ed or -ing suffixes;

2. replace terminal y with i when there is another vowel in the stem;

3. reduce double suffixes to single ones (e.g., fool-ish-ly, care-less-ly);

4. remove known morphological suffixes (-icate, -ative, -alize, -iciti, -ical, -ful, -ness);

5. remove some more known suffixes[6] (-able, -ment, etc.); and

6. remove final -e.

In our experiment we used the well-known Lingua::Stem::En library available on CPAN. This algorithm is based on Frakes and Cox's 1986 implementation of the Porter's Algorithm for stemming.

2.1.5 Language Classification

Language classification of text is a problem tackled in many disciplines and is particularly difficult when the amount of text to analyze is very limited or otherwise constrained. For texts longer than 400 characters, Cavnar and Trenkle demonstrated [8] that it is possible to achieve 99.8% accuracy using n-gram classification, but when dealing with short and ill-written texts such as tweets, Tromp and Pechenizky reported [88] that the accuracy of such a method often drops below 90%, which could translate in more than 50 million misclassified tweets per day.

Fortunately, the keywords chosen in our filters were mostly English medical terms, thus the majority of the posts retrieved through the Streaming API were already in English and we did not need to apply any further language classification. That said, during our experiments, we explored the possibility of using additional information included in the tweet stream that had not previously been exploited to further improve the accuracy of language identification.

For example, Twitter allows its users to select a language for the web interface.

[6] Complete list in Appendix A.2

People who sign up for the service using a mobile device generally do not bother changing it (the default setting is English) to their native language because they tend to consume the data through a (generally localized) mobile application. In addition, many international users seem to prefer to have the interface in English. That said, there seem to be a modest percentage (11%) of users who did go into the trouble of changing the language of the interface, and for these, a high percentage (96%) of the tweets they publish are generally written in their native language (see Figure 2.7). Using the language selected for the web interface while trying to recognize the language of the tweet could boost the accuracy of the classification.

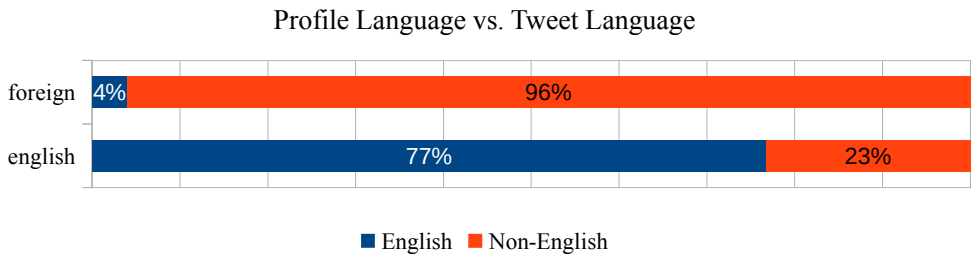


Figure 2.7: Percentage of Non-English Tweets given Twitter Profile Language

Similarly, users can specify their time zone in the profile. As for the language of the interface, many users (35%) do not bother to change it from the default setting, but for those who do (see Figure 2.8), it can be used as another parameter in the language identification process. For example, tweets produced by profiles in US time zones (e.g., "Mountain Time") are more likely to be written in English than those produced in European time zones (e.g., "GMT+3"). Using the time zone in the language classification process could thus improve the accuracy of the process.

Even where the time zone has not been selected or is inaccurate (e.g., the user

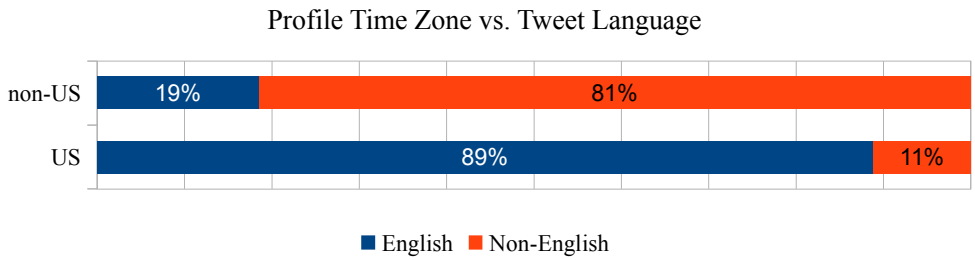


Figure 2.8: Percentage of English vs. non-English Tweets given a US Timezone

is traveling), knowing that people generally post their tweets during the day (e.g., 7am to 11pm) implies that the (absolute) time of the tweet may be an additional indicator of its geographic origin (e.g., tweets sent during US night time are less likely to be in English). During our experiments we did not use these findings, but including the time of the tweet in the language classification process could produced an additional boost in the classification accuracy.

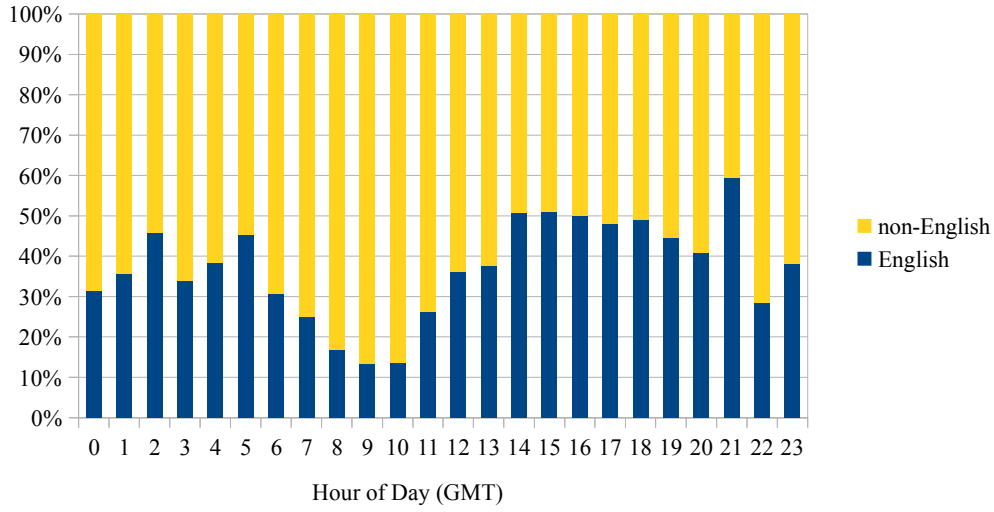


Figure 2.9: Percentage of English Tweets by Hour of Day

In a similar fashion, Twitter recommends that users specify their home location in their profile. While many people (33%) seem to enjoy providing fictional locations (e.g., "heaven"), others specify their city (e.g., "Boulder, CO") or at least their state (e.g., "California"). In addition, more recently, many tweets written on mobile handsets contain exact GPS coordinates allowing to more easily identify their origin. As with time zone, this attribute may not be accurate (e.g., due to travel or spoofing) but tweets sent from profiles who claim to be based in the US are more likely to be written in English. Including this attribute in the language classification process could further boost its accuracy.

Other profile attributes such as the full name of the user (when specified) or the location/language of the people they follow could have a positive correlation with the favorite language of the user and thus the language of their tweets.

2.1.6 Spam and Unrelated Tweet Removal

Over the last few years, email, social networks and forum, became one of the main communication methods of Internet users. As their popularity increased, marketers took the opportunity to use these new channels to advertise their products and services to thousands of users with little effort and low costs. The practice of using electronic messaging systems to send unsolicited messages indiscriminately is generally referred as "*spamming*". According to the Internet Society [83] the term "*spam*" derives from a sketch in a 1970's British TV series, and the first documented spam was a message advertising the availability of a new model of Digital Equipment Corporation computers sent by Gary Thuerk to 393 recipients on ARPANET in 1978. Instead of sending a separate message to each person, he had his assistant write a single mass e-mail and send it simultaneously to every user.

The practice was not appreciated by the community but seem to have generated some sales.

As the Twitter platform became more popular, people all over the world started using it as a marketing mechanism, thus introducing spam in the system. It is not uncommon to find automated tweets (about 2%) from companies selling drugs online (e.g., "*Boiron Oscillococcinum Natural Flu Relief 6 pack with 6 Dose Bonus Pack, 12 total! http://t.co/hGISpYz2*"). Moreover, while the keywords chosen retrieve mostly (about 75%) health-related tweets, it is possible to find irrelevant tweets (e.g., "*I'm sick of my job*"). Since both spam and unrelated tweets may negatively influence our results, we introduced additional filters in the preparation step to accurately filter out spam (e.g., drug advertisements) and eliminate unrelated tweets (e.g., non health tweets) in order to improve the quality of our input and consequently the accuracy of the final model.

We used Naive Bayesian classification [60] to identify health-related tweets. In our experiments we used the CPAN module `Algorithm::NaiveBayes` and trained our model on a set of 2000 tweets produced by known health authorities (e.g., @who, @CDCFlu, @nytimeshealth, @goodhealth, etc.). Although the initial model was simple, a test performed over 250 manually labeled tweets showed that the accuracy of the classification was quite high (92%). Errors in the classification were generally due to very short tweets (2 or 3 words) or very ambiguous content (e.g., "my job makes me sick!").

Unfortunately, spammers have refined their craft over the last few years, and their fake (and often illegal) drug advertisement posts often appear as perfectly valid health tweets. While a large portion (98%) of these tweets contain a URL that point back to the sale site, user comments on public media announcements

also often contain a URL that points to the source. Fortunately, tweets meant to sell drugs generally (85%) included sale-terms such as "cheap" or "discount", which made identification easier. In addition, while professional spammers are getting better at simulating "real" users, their profiles are often quite generic (e.g., no full name), contain default settings (e.g., English language and GMT time zone), unpronounceable usernames (e.g., @gawuwoguhadi), have a URL in their bio, and generally have a comparable number of followers and users who they are following. We added these features to our Naive Bayesian classifier which was trained and tested again on the labeled dataset obtaining an accuracy of 97%. The remaining errors were mostly due to very clever spammers who created very realistic profiles and used well crafted health-related text in their tweets but linked to their own sales-related site.

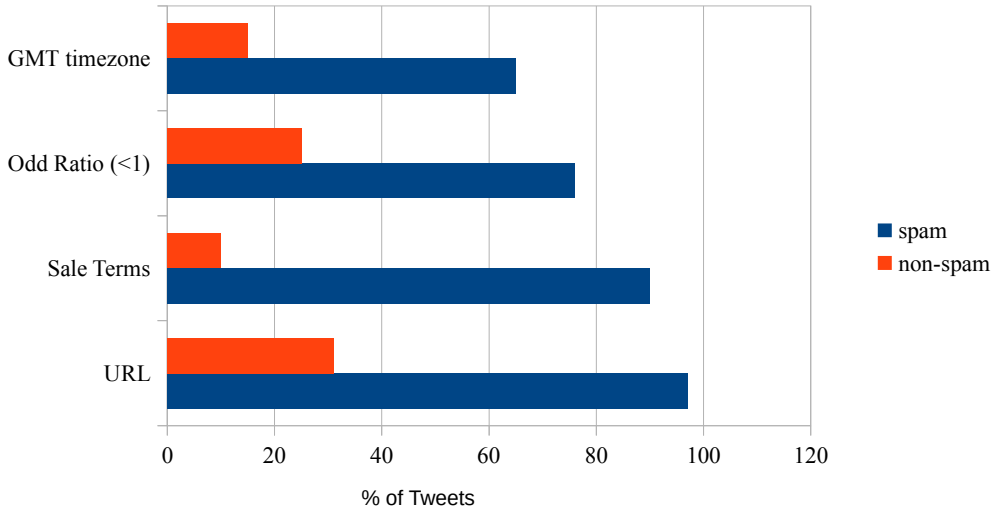


Figure 2.10: Percentage of Spam/Non-Spam Tweets with certain features

Geolocated tweets require a similar level of preprocessing. Some sites use Foursquare

checkins to measure the popularity of venues, making them a perfect target for spam. In addition, automated bots (e.g., job postings) often add geographic coordinates to their tweets, making it appear as if the poster were traveling all over the place. In our work, we eliminated users who appeared to travel excessively quickly (>1800km/h) or checked in too frequently (>1 checkin every 5 seconds). We also eliminated casual users (<5 checkins) since they simply did not provide sufficient data to form a reliable model of the user's movement.

2.1.7 General Applicability

The methodologies explained have general applicability and be used as guidelines in studies from any domain. As an example, they have been used in a recent infodemiology study [116] from Zhange & Al, to gather insights about the characteristics of fitness-related messages exchanged on Twitter. In this study, we used a set of 113 keywords (see Appendix A.1) as filter for Twitter's Streaming API between January 1st, 2011 and March 31st, 2011 and collected of about 1 million fitness-related tweets. Spam and foreign tweets were excluded with the methodologies described in the previous sections. The study manually analyzed about 5,000 tweets from the sample and reported that about 60% of them implied a user's intention to participate in some physical activity, but that only 10% of them contained any hint of social support. While this particular study was not relevant to the topic of this thesis, it does lend credibility to the generalizability of the methods developed here.

2.2 Support Vector Regression

Support Vector Regression [5] is an instance of the more general class of Support Vector Machines (SVM), a supervised learning method generally applied to solve

classification problems [105]. A classification system categorizes examples as instances of some class or concept. For example, one might build a classification system to discriminate between low and high risk for hospital readmission on the basis of information provided in a patient record. A learning method attempts to automatically construct a classification system from a collection, or training set, of input examples. Elements of the training set are usually represented as a collection of values for pre-specified features or attributes; for this example, these features could be such measurable properties as age, recent hospitalizations, recent clinic visits, etc. Training set elements are marked *a priori* with their outcome, or class membership (e.g.,"high risk"). Once generated, the classification system can then be used to predict the outcome of future examples on the basis of their respective feature values. Commonly-used learning methods include neural networks, Bayesian classifiers, nearest-neighbor methods, and so on; here, we use SVMs.

SVMs use quadratic programming, a numerical optimization technique, to calculate a maximum-margin separator, that is, the hyperplane that maximally separates data points belonging to different classes in the multidimensional feature space, while tolerating only a pre-specified error rate. Since the data are often not linearly separable (e.g., there is no simple linear expression, or hyperplane, that separates high risk from low risk of hospital readmission), a kernel function is used to project the data into a higher-dimensional space. If the new space has a sufficiently large number of dimensions, it ensures that a maximum-margin separating hyperplane exists and will be found efficiently, even if the original data are not linearly separable. Commonly-used kernels include the radial basis function, hyperbolic tangent function, and the polynomial kernel function (used in this application).

When used for regression, SVMs produce a nonlinear model that minimizes

a preselected linear-error-cost function where features serve as regression variables. Each input data point is described as a collection of values for a known set of variables or features. In our experiments, each data point represented a tweet and the feature set was defined as the collection of terms in the dictionary appearing more than 10 times per week. For each time interval (i.e., a week of the flu season), the value of a feature was given by its usage statistic for the corresponding term. Thus each tweet had been encoded as a feature vector of length equal to the number of dictionary terms occurring more than 10 times per week, where the value assigned is the fraction of total tweets in that time interval that contain the corresponding dictionary term after stemming.

In the works reported here, we relied on the widely adopted open-source `libSVM` library [13] for our computations.

CHAPTER 3
APPLICATIONS

Social media adoption has been increasing at steady pace for the last years across all ages and genders. As these mediums became more and more popular, marketers all over the world have started to encourage customers to use them to express their opinions and look for help. It is not uncommon in these days to see a hashtag in the corner of a TV show or below a manifest, or being invited to share a picture to receive a discount in a store. Even companies that were once difficult to reach (e.g., airline companies) created social media teams dedicated to scout the web for brand-related chatter and to address every complaint publicly, creating a bidirectional channel with the customers and increasing brand loyalty. Similarly, athletes and celebrities of all types started using these media to reach their fans, and app developers started adding functionality that allows their software to automatically post the user's result (e.g., the length and calories consumed in a run) on popular services.

The result of all these interactions is a very rich dataset that captures data about and the opinion of many diverse people across many different topics. On Twitter it is easy to find tweets expressing, in real time, people's feelings about the current episode of their favorite show, predictions or orders made on the stock market, fitness activities (e.g., runs) just performed, weather information, job postings, reactions to public news and blog posts, complaints about service interruptions (e.g., Gmail), and many others.

Over the course of the last few years we have been collecting, monitoring and analyzing the Twitter stream and have proven its utility in many situations. In this

chapter we describe a few successful applications of this type of analysis.

3.1 Predicting the American Idol 2009 Winner

As a first proof of concept exercise, we aimed to predict the outcome of a popular user-driven entertainment contest using Twitter. American Idol (AI) is a reality-show competition to find new solo musical talent. It debuted on June 11, 2002 and has since become one of the most popular shows on American television. The program is a spinoff from Pop Idol, a reality program created by British entertainment executive Simon Fuller and first aired in 2001 in the United Kingdom.

The program conducts a series of nation-wide auditions looking for the best performers. The American public decide the outcome of the final stages through telephone voting. The judges (usually record producers, singers or music executives) on the show give critiques to the contestants after each performance. On American television, the show is usually aired on two consecutive weekdays: on the first evening each contestant performs one or more songs, and on the following night the outcome of the phone voting is announced and one of the contestants is sent home.

Given the popularity of the show and the fact that its target audience demographic is largely compatible with the power users of Twitter, we decided to analyze the AI-related tweet stream. We used the Twitter search API to retrieve tweets which contained the full name of each of the last 5 contestants (Adam Lambert, Danny Gokey, Matt Giraud, Allison Iraheta and Kris Allen) of American Idol Season 8 between April 28, 2009 and May 20, 2009. Since the number of fans of each contestant might vary, we decided to normalize the daily counts of each contestant dividing by the daily average number of tweets (for the same contestant) received over the month collected.

Interestingly, we observed that even on such a small time scale the sequence

of peaks of tweets with each contestant's name closely tracks the order in which the contestant sang during the show. Figure 3.1 depicts the temporal distribution of the tweets for each contestant during the night of May 5th, 2009. Analyzing the sequence of the peaks it is possible to reconstruct the order in which the contestants sang: Kris Allen, Adam Lambert, Danny Gokey and Allison Iraheta.

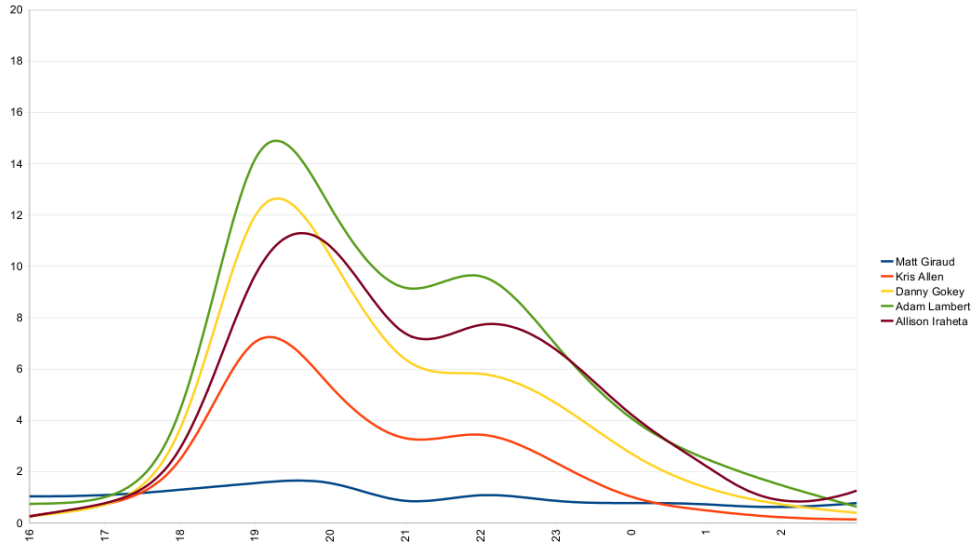


Figure 3.1: Tweet Volume associated to each American Idol 8 contestant

On May 20, 2009 (the morning before the finale), Lara Hejtmane published an article [47] on Mashable[1] where she applied Google Flu's prediction model to guess the outcome of the popular TV show. In her study, Hejtmanek observed how in the previous years the distribution of queries based on finalists' names closely matched their final order in the show for each season. Figure 3.2 shows search query trends for American Idol 7. Analyzing the most recent query trends (Figure 3.3), the author announced that Adam Lambert would win season 8 of American Idol. This curious use of search query trends allowed the article to gain substantial popularity

[1] www.mashable.com

on the web and the prediction made by Hejtmanek was endorsed by many other bloggers.

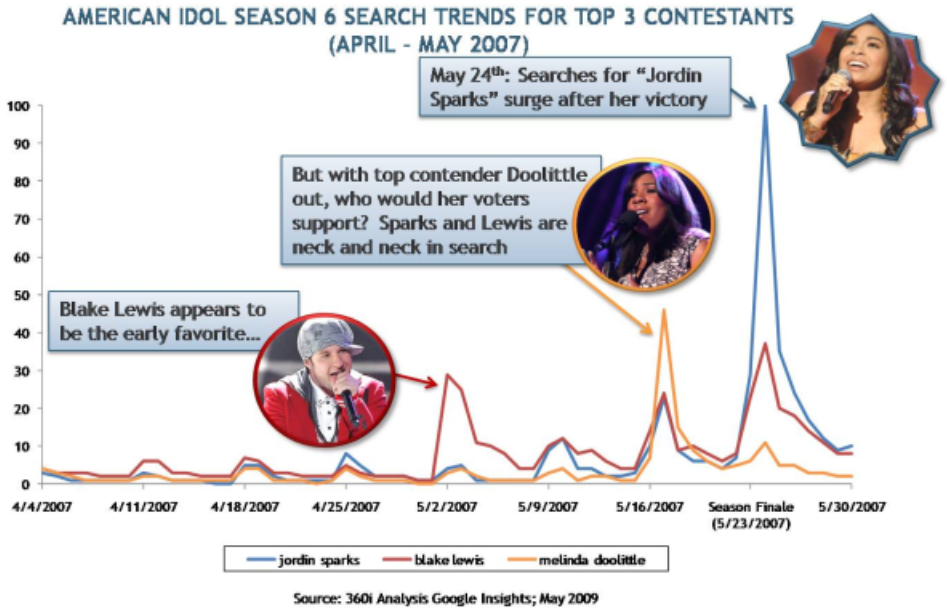


Figure 3.2: Google Search Volume for American Idol 7 Contestants

Inspired by this article, we studied the number of tweets published for each finalist during the 2-night finale. Our analysis discovered that the total number of tweets published for each contestant did not offer any particular clue on who might be the public's favorite. For this reason, we decided to limit our analysis only to positive tweets (e.g., containing words like "love", "best" and "win").

Figure 3.4 above shows the relative number of positive tweets (over the average for the previous week) obtained by each contestant during their final performance. Observing this graph, Kris Allen's performance seemed to have achieved a higher level of public appreciation with respect to that of his rival Adam Lambert. For this reason, despite what Google's search terms may indicate, we were confident

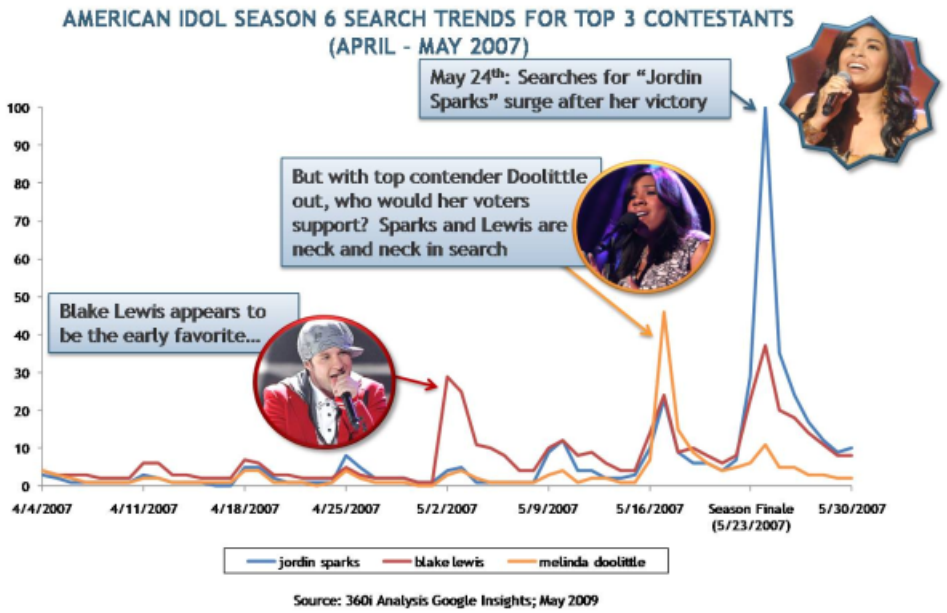


Figure 3.3: Google Search Volume for American Idol 8 Contestants

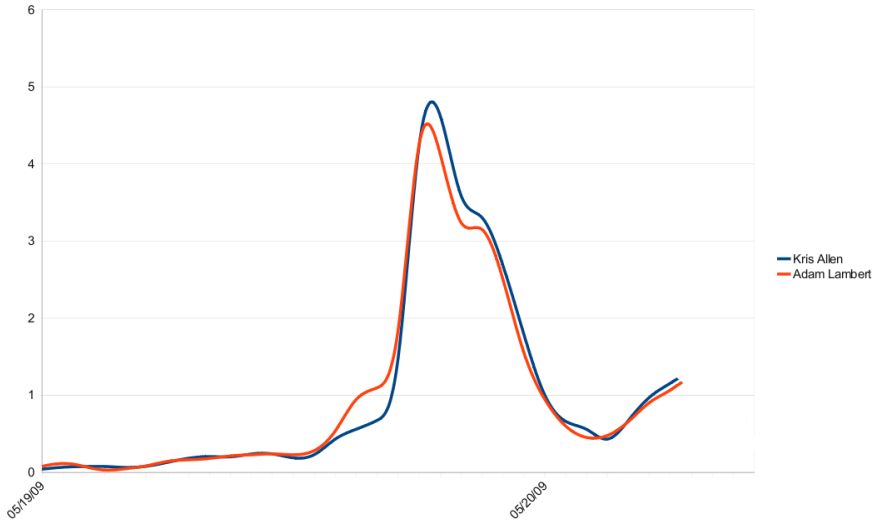


Figure 3.4: Relative Number of Tweets during the Finale of American Idol 8

that Kris Allen would win American Idol 2009: our prediction was, in fact, correct [104]. This small (and perhaps frivolous) study seemed to indicate that social web activities are in fact highly correlated with public perceptions of certain topics, and pushed us into thinking that the aggregation of information implicitly released in public conversations (e.g., a tweet or a blog post) could be very effective in a public health context.

3.2 Monitoring the Swine Flu Pandemic

Predicting the outcome of a reality show does not necessarily constitutes a contribution to the public good. However, the same techniques can be used to predict public perception of other more important societal concerns, such as the reach of public health messaging.

Novel influenza A (H1N1) is a relatively new flu virus of swine origin that was first detected at the beginning of April 2009 in certain regions of Mexico. This mutation of the virus, capable of infecting humans, spread from person-to-person sparking outbreaks of influenza all over the United States as well as internationally. The Centers for Disease Control and Prevention (CDC) in Atlanta, GA issued the first outbreak report on April 23, 2009, after which human cases of H1N1 infection were identified in San Diego County and Imperial County, California as well as in San Antonio, Texas. Media outlets all over the world depicted this pandemic as disastrous, forecasting thousands of deaths and hospitalizations.

On April 26 of the same year, the CDC published some general prevention guidelines (e.g., cover your nose and mouth when sneezing, wash your hands often, etc.) while announcing that face masks had been distributed in community settings where the spread of influenza had been detected. In the same update, the CDC

announced that the virus seemed to be susceptible to common antivirals such as Tamiflu and Relenza. While the number of new cases identified increased only by a few dozen per day, the increase in number of articles and news reports published was hundreds of times higher, making the spread of influenza a common topic of discussion. Fearing a pandemic, many prepared for the worst and stockpiled food, water and medical supplies. Travel to and from Mexico (but also anywhere within the U.S.) was curtailed, and in many airports passengers and workers started wearing surgical masks at all times.

Fueled by the desire to monitor and estimate the response to the situation, we started collecting influenza related tweets on April 28, 2009. Using the search API and the keywords reported in Section 2.2, we retrieved all the H1N1-related tweets published between April 28 and May 15, 2009. Each entry was timestamped and contained additional information about the user (e.g., geographic location). Swine-related entries were identified by searching through Twitter's public stream for tweets matching specific keywords:

- swine AND (flu OR influenza);
- H1N1;
- (face OR surgical) AND (mask OR masks);
- relenza OR zanamivir;
- tamiflu OR oseltamivir; and
- (hand AND (wash OR washing)) OR handwashing.

At the same time we created a client-side JavaScript application to monitor H1N1-related tweets published in the United States in real time. This interface (Figure

3.5) continuously updates a Google map with the last 500 most recent matching tweets, yielding a real-time view of flu-related public sentiment. Users could read any tweet by placing the cursor over its corresponding colored dot on the map.

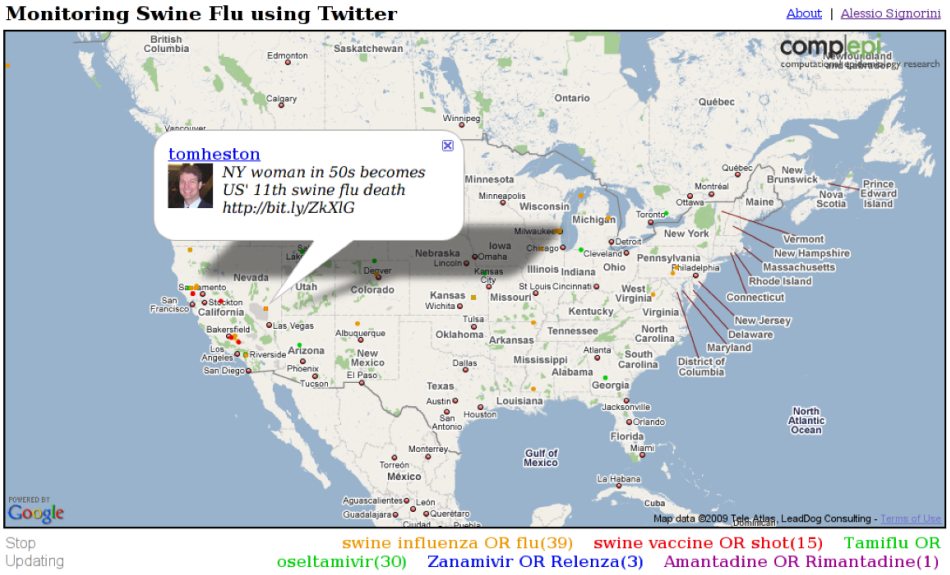


Figure 3.5: Screenshot of H1N1 Realtime Monitor Interface

During the period in question, we collected a total of 592,543 H1N1-related tweets. After removing irrelevant entries from the data, we aggregated the tweets into categories (e.g., antiviral, handwashing, masks) and compared their temporal distributions with the increase of cases identified, and the public announcements of the CDC and other health organizations.

Looking at Figure 3.6 it is interesting to notice how the majority of the tweets were published before May 7, 2009, when the number of cases detected was still well under one thousand. The volume of conversations on H1N1-related topics does not seem to grow proportionally with the number of cases detected, which might suggest a high correlation between the tweet stream and the evolving perception

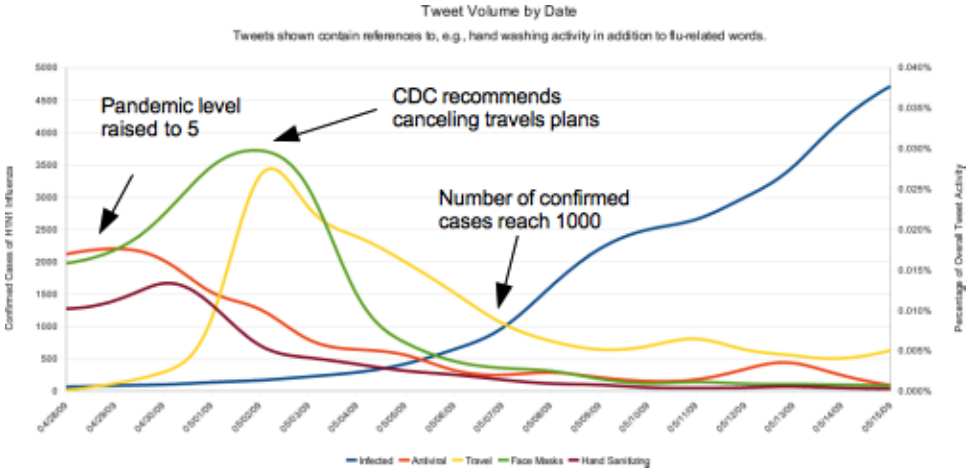


Figure 3.6: Tweet Volume for each Category by Date

of the outbreak on the part of the public rather than the actual number of cases detected. Moreover, the peaks of the various categories appear to track public announcements by various health organizations.

For example, tweets containing references to antiviral drugs peaked on April 29, the same day the World Health Organization (WHO) raised the pandemic warning level to 5, but fell as soon as official reports indicated that most U.S. cases were relatively mild and did not require hospitalization. Nevertheless, as a reaction to the warning, numerous media agencies republished the safety guidelines issued by the CDC just a few days earlier, which possibly accounted for the peak in tweets with hand-sanitizing references on the following day. In addition, many health organizations recommended canceling all unnecessary travel and wearing surgical masks as a precaution while in crowded public spaces, such as planes or airports.

It is interesting to note how the number of tweets corresponding to these topics peaked on the day following the announcement [100]. Thus the evidence suggests that it might be possible to use tweet analysis as an inexpensive way to determine

not only the public's level of anxiety and concern, but also to gauge response to news and official public health messaging.

3.3 Using Twitter to Estimate H1N1 Influenza Activity

The following influenza season, inspired by the Yahoo! [68] and Google's [42] studies, we wanted to demonstrate how the number of occurrences of certain keywords in tweets can be used to estimate the level of influenza activity at both the national and regional levels, with a high degree of accuracy, and several weeks in advance with respect to official reporting methodologies.

Because CDC-reported data are generally only available one to two weeks after the fact, real-time estimates constitute an important tool for public health practitioners. Although influenza is not a nationally notifiable disease in the U.S., an influenza surveillance program does exist [12]. One component of this surveillance program is tracking reported influenza-like illness (ILI) during influenza season (usually October through May), since earlier detection can improve both clinical and public health responses. As explained in Section 3.2, members of the US Outpatient Influenza-like Illness Surveillance Network (ILINet) report [11] the total number of patients seen along with the number with ILIs (i.e., body temperature of 37.8'C or greater, cough and/or sore throat without any other apparent cause). Because ILI data are not uniformly reported by each state, the data are aggregated within the 10 CDC Influenza Reporting Regions and subsequently weighted by regional population [11].

To track the 2009-2010 flu-season, we collected roughly 8 million flu-related tweets between October 1, 2009 and May 20, 2010 following the procedure described

in Section 2.1. Unfortunately, many of the tweets had to be discarded because did not meet the criteria on location (i.e., non-US), language (i.e., non-English), spam or content described in Section 2.1.3, reducing our dataset to 4,199,166 tweets.

Our first experiment was to estimate weekly ILI% values at the national level. Each weekly model was trained using support-vector regression. For each week in our dataset, the training vector was composed by the term-frequency statistics for the term set just described and the objective was the official CDC-reported national ILI% value for that week.

To verify the accuracy of our models, we used a standard leaving-one-out cross-validation methodology [74], training 32 times, once on each 31 week subset of the training data and then testing on the remaining week. Figure 3.7 compares the 32 estimated (red line) ILI values obtained with target ILI values reported by the CDC (blue line). These estimates are point estimates, which do not reflect temporal aspects of the data. Even so, the estimates of national ILI values produced by the system are fairly accurate, with an average error of 0.28% (min = 0.04%, max = 0.93%) and a standard deviation of 0.23%.

Encouraged by the results obtained at national scale we computed real-time estimates of ILI% activity in a single CDC region. This experiment required that we to be able to unequivocally associate each tweet to a US state. To that extent, we only considered tweets that embedded precise geolocation data, or for which the user declared a well-formed US location that matched either:

- <full_state_name >;
- <state >, <country >;
- <city >, <state >; or

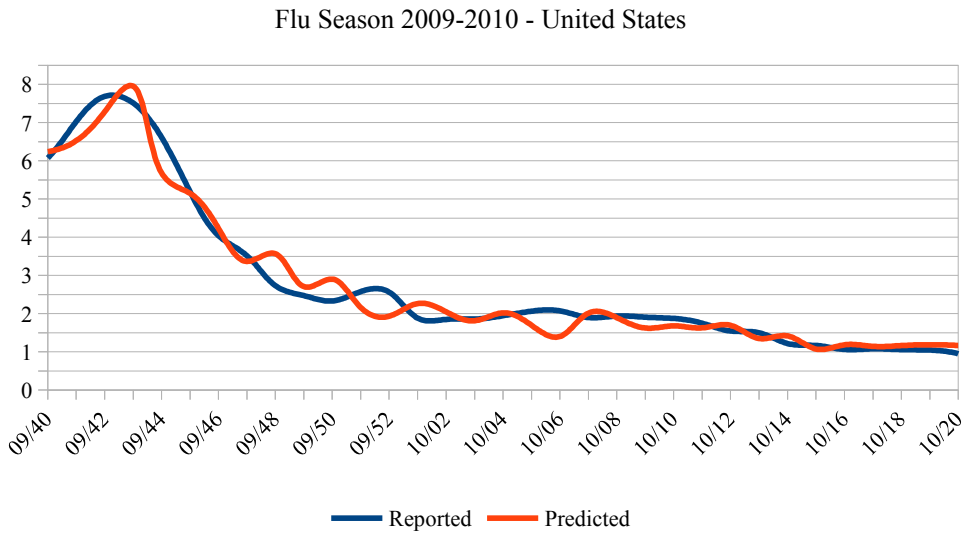


Figure 3.7: Predicted vs. Reported ILI% in the U.S. for the 2009 Flu Season

- <city >, <state >, <country >

where <state> could either be the full name or the abbreviation, and <country> could either be "US", "USA" or "United States". This requirement reduced our dataset to 905,497 tweets.

Using the same methodology as for the national level experiment, we fit geolocated tweets to CDC region ILI% readings from 9 of the ten CDC regions to construct a model trained on all their data. We then used the model to estimate ILI% values for the remaining CDC region (Region 2, New Jersey and New York). The regional model (Figure 3.8) still approximates the epidemic curve as reported by ILI data, although the prediction (based on significantly fewer tweets) is somewhat less precise with an average error of 0.37% (min=0.01%, max=1.25%) and a standard deviation of 0.26%.

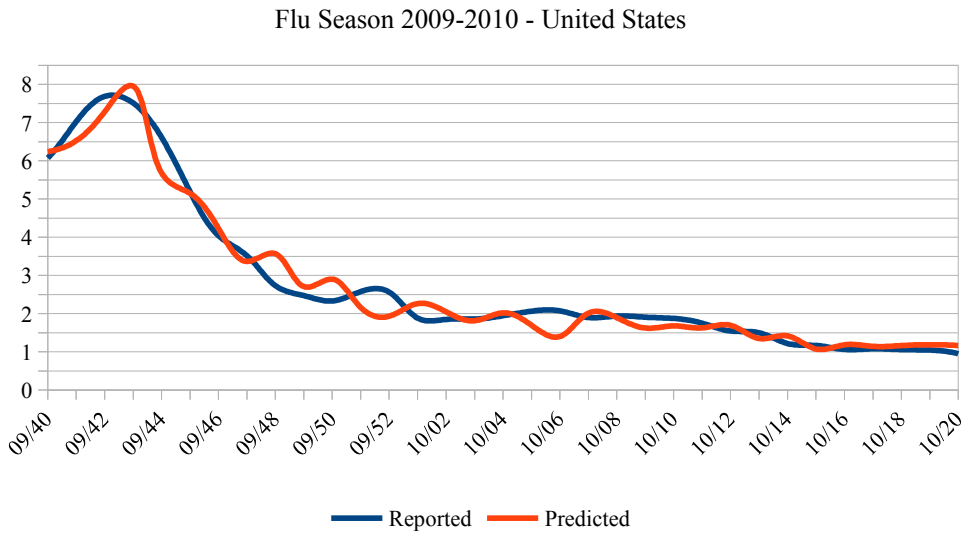


Figure 3.8: Predicted vs. Reported ILI% in Region 2 for the 2009 Flu Season

3.4 Inferring Travel from Social Media

The spread of infectious diseases is facilitated by human travel. Disease is often introduced by travelers and then spread among susceptible individuals. Likewise, uninfected susceptible travelers can move into (infected) populations thereby sustaining the spread of an infectious disease. Several disease-modeling efforts [45] have incorporated travel and census data in an effort to better understand the spread of disease. Unfortunately, most travel data are not fine grained enough to capture individual movements over long periods and large spaces. Alternative methods, such as tracking currency movements [7] or cell phone calls [43], have been suggested to measure how people move with higher resolution but these are often sparse, expensive and not readily available to researchers.

Over the last few years, more and more mobile devices have been equipped with GPS receivers, which application developers have hurried to exploit. Location

aware applications have been growing in popularity and many of the tweets sent today embed exact geo-location information. In our study, we took advantage of the geographic coordinates included in tweets to build national and city-level travel models aggregating the movements of users. Using the methodology described in chapter 2.2, we collected 68 million geocoded entries (regular tweets and checkins embedded in tweets) from 3.2 million users using the Twitter streaming API for the period from September 11, 2010 through January 28, 2011. Prior to analysis, we eliminated the checkins not originating in the US as well as any suspicious ones following the rules explained in Section 2.1.6.

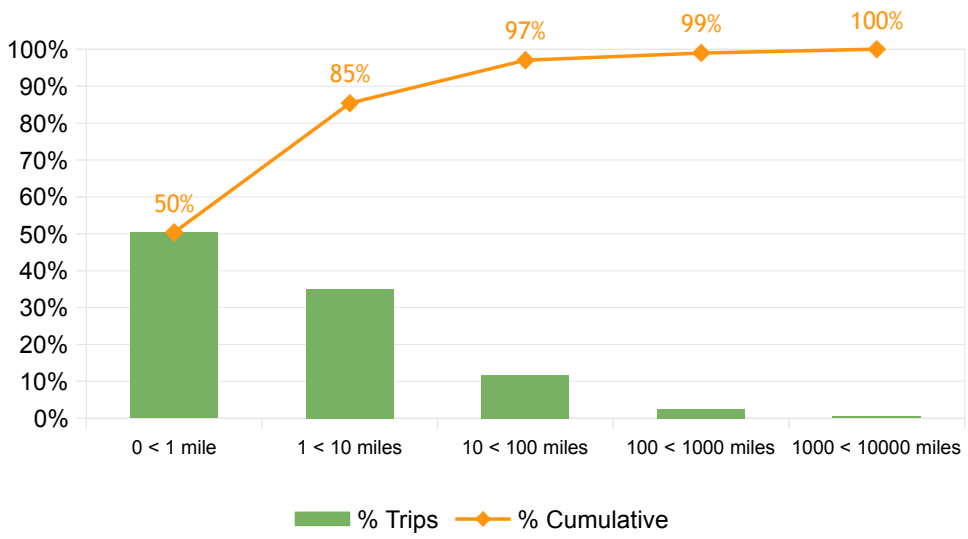


Figure 3.9: Statistics on Geographical Distance between Foursquare Checkins

By linking each individual users' consecutive location records together, we computed the statistical distribution of time intervals (Figure 3.10) and distance traveled (Figure 3.9) between records. These data showed that power users of those applications use them very frequently, with half of the checkins being less than 6

hours and not more than 1 km apart from each other.

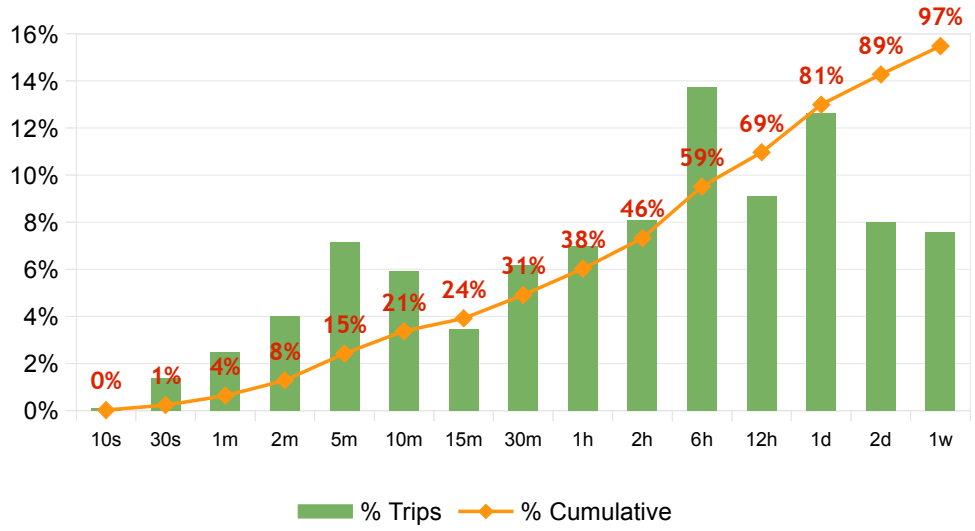


Figure 3.10: Statistics on Time between Foursquare Checkins

In one of our experiments we analyzed a random subset of about 3 million record intervals from about 165,000 users, spanning between September 11, 2010 and October 26, 2010. For the national model we aggregated these data by state using the GPS coordinates and a precision of about a mile over state borders. Figure 3.11 represents a plot of these data over a US map and shows the density of the movements between states as lines between their capitals.

Similarly, for a few big cities the amount of data available was sufficient to create interesting street level models. One of such place was New York City, for which we were able to create a local check-in density map (Figure 3.12) by aggregating users' check-ins with 500-meter resolution. A larger bubble represents a denser set of records in that geographic area.

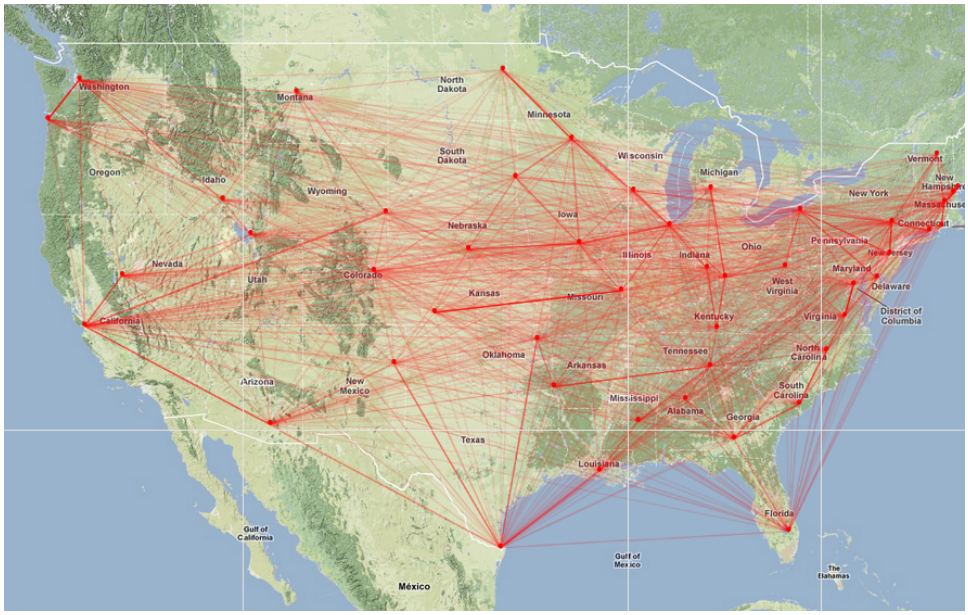


Figure 3.11: Travel plot with state level resolution

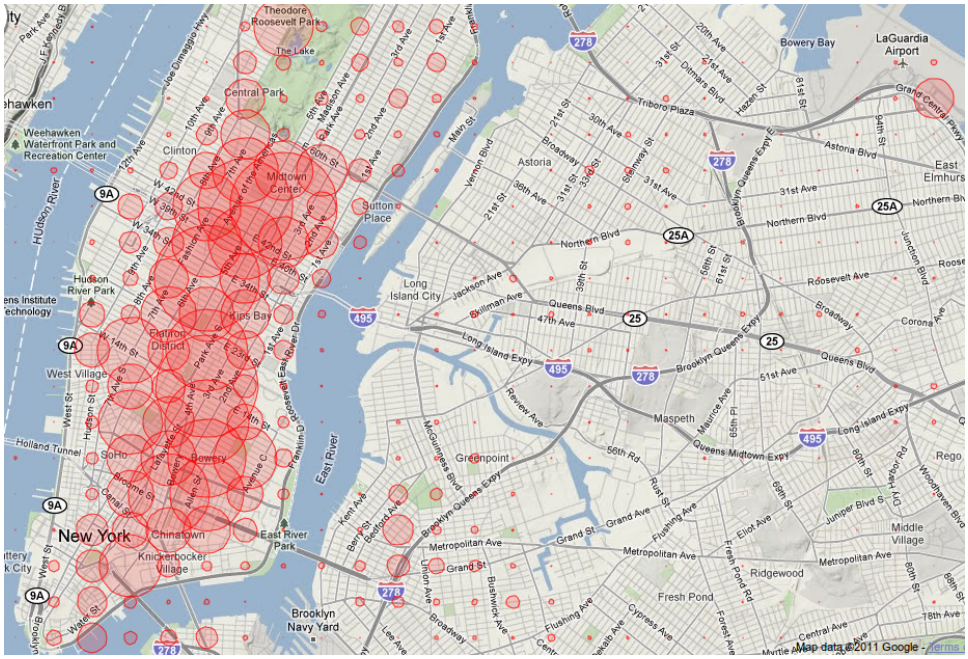


Figure 3.12: Density of Foursquare checkins in New York City

Exploring the density of the checkins at different times of the day shows macro-trends in movements as depicted in Figure 3.13. In the case of New York City, for example, we notice rush towards the Wall Street area (biggest bubble at the bottom of the 6am and 2pm picture) during the middle of the day, and a retreat towards areas with more entertainment venues (e.g., Time Square) during the evening hours.

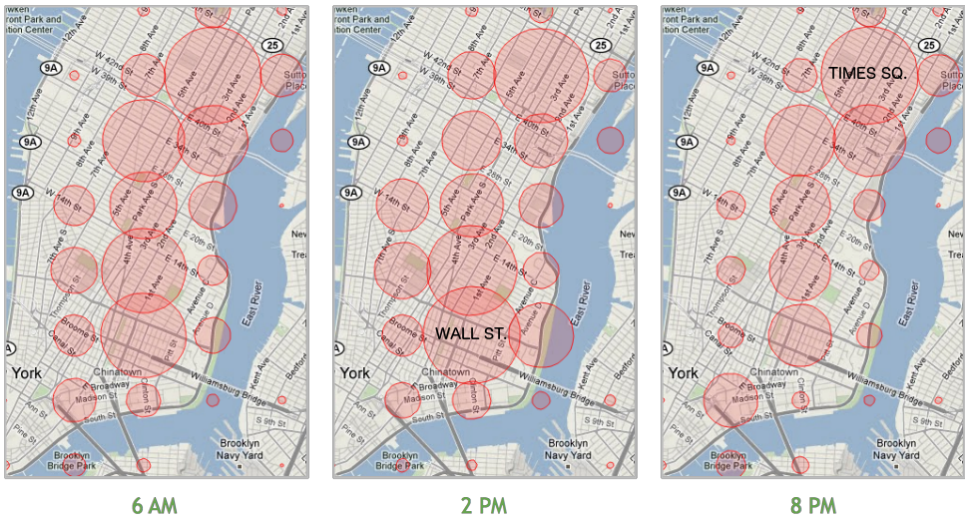


Figure 3.13: Density of Foursquare checkins in Manhattan by time of day

The movements of individuals can be easily tracked across the city by linking (together) their consecutive checkins. Figure 3.14 gives some examples for New York City.

The results of our work suggests that social media location data can be used as multiscale proxy for travel at the national, state and urban level. These data are inexpensive and easily obtained. Furthermore, they can be used not only to understand historical travel but also to monitor in real-time changes in travel behavior to help inform disease surveillance.

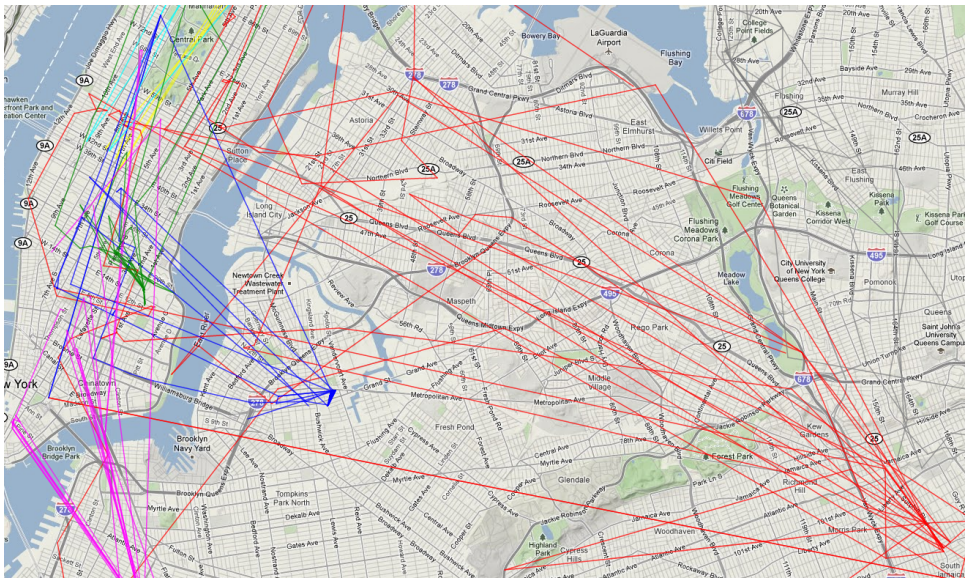


Figure 3.14: User paths across New York City inferred through Foursquare checkins

3.5 Predicting Local Flu Trends using Geolocated Tweets

Although global connectivity reduced distances in many ways, traveling, both for work and pleasure, is still frequent and necessary among almost every population. The recent surge in low-cost airlines and general flight availability steadily increased the number of passenger traveling both domestically and internationally year after year. According to the U.S. Department of Transportation [63], more than 645 million passengers were carried in the US in 2013.

The spread of infectious diseases is facilitated by human travel. Disease is often introduced by travelers and then spread among susceptible individuals. Likewise, uninfected susceptible travelers can move into populations sustaining the spread of an infectious disease. Airports and airplanes dramatically increase the probability and frequency of contact between individuals, even across populations that would not ordinarily meet. Many studies [82] [30] suggested imposing restrictions on airline

travels to limit the spread of epidemics, but restrictions were rarely put in place.

While intuitive, the correlation between movements of travelers and flu has been notoriously hard to demonstrate [82], due in part to the difficulty of obtaining data to create a reliable travel model. Several disease-modeling efforts [46] have incorporated travel and census data in an effort to better understand the spread of disease. Unfortunately, most travel data are not fine grained enough to capture individual movements over long periods and large spaces. Alternative methods, such as tracking currency movements [7] or cell phone calls [43], have been suggested to measure how people move with higher resolution but these are often sparse, expensive and not readily available to researchers.

Flight ticketing data would represent a great source of information on how people move around the US. Unfortunately, especially after the events of September 11, 2001 it has been difficult to obtain access to this data for research purposes, and purchasing it from the International Air Transport Association (IATA) can be quite expensive. On the other hand, the widespread adoption of geolocation-enabled application may provide an alternative solution to this problem.

In here we demonstrate how geolocated tweets can be used to infer the movements of its author, and how the aggregation of these movements allows to create travel models which can then be used to improve the accuracy in the prediction of local flu patterns across many US cities.

3.5.1 City Level Flu Data

The Center for Diseases Control and Prevention (CDC) publishes ILI weekly data at regional level and a few states (e.g., Colorado) make their weekly statistics available for download through their health department's website. Unfortunately, given our

focus on travels across cities and its correlation to flu trends, these statistics were not fine grained enough to be used in our experiments.

To approximate city level data we decided to use the CDC Morbidity and Mortality Weekly Report (MMWR) as proxy for local flu activity. This report is composed by the number of deaths voluntarily reported by 122 cities (most with a population of 100,000 or more) divided by age range, and due to pneumonia and influenza. Although often incomplete and noisy, these numbers could be use to approximate (albeit with some delay) flu activity levels in each city.

To deal with late reporting of deaths and noise in the data, we performed our experiments on a smoothed version created by averaging the value of each week with the ones of the previous and following 2 weeks. For each city/year, the data has also been normalized dividing by the maximum value observed, so that the peak is always 1 and different cities can easily be compared. In the rest of this section we will refer to this measure as "flu trend". Figure 3.15 shows an example of the raw mortality raw data and its smoothed flu trend for the city of New York, NY in 2012.

The MMWR dataset is very noisy and required a good amount of cleanup. Among the 122 cities in the reports, 3 always report only partial data (Pittsburg, Scranton, and Erie in Pennsylvania), and 20 reported data for less than 45 weeks of the year (table 3.1). Those cities were eliminated from the dataset.

3.5.2 Travel Data

Using the methodology described in Section 3.4, in 2012 we collected about 240 million geolocated tweets (authored by about 4 million users) that were posted inside the United States according to the coordinates embedded in the metadata.

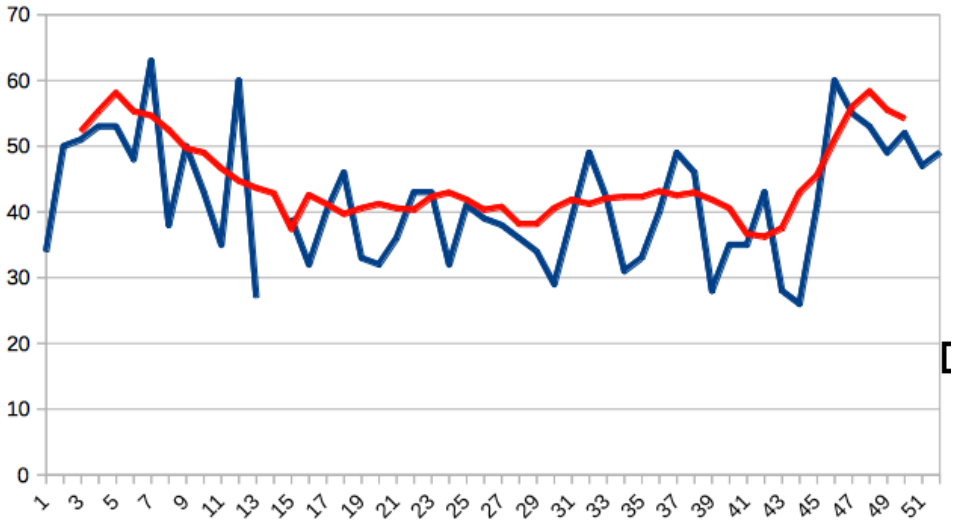


Figure 3.15: Flu & Pneumonia Deaths in New York City, NY for 2012

New Orleans (LA), Baton Rouge (LA), Trenton (NJ), Gary (IN)
Jacksonville (FL), New Bedford (MA), Shreveport (LA), Utica (NY)
Wilmington (DE), Pueblo (CO), Allentown (PA), Lincoln (NE)
Miami (FL), Cambridge (MA), Tampa (FL), Norfolk (VA)
Lexington (KY), Denver (CO), Reading (PA), Providence (RI)

Table 3.1: MMWR Cities removed due to lack of data

The statistical characteristics of this dataset were similar to the ones described in Section 3.4.

For each city in the MMWR dataset, we obtained the geographical coordinates using the Google Geolocation API [44] and population information using the web interface available at City Data[2]. Since each tweet has to be univocally mapped to one of these cities we had to remove 20 cities with overlapping metro areas (see table 3.2). Honolulu (HI) was also removed from the dataset because no tweets were collected for this location. Each tweet in a 20 miles radius from the coordinates of one of the cities was assumed to belong to that city. This process reduced the dataset to 124 million tweets posted by 2.2 million users.

For each user, we sorted the tweets according to the timestamp and considered "home" the city in which the majority of the tweets were produced. With that information in hand, we assumed that the user took a "trip" whenever a tweet in the home city was followed by one in another city. We then aggregated the number of trips between each two cities to obtain the final counts.

To account for the different penetration of Twitter and geolocation technologies in different metro areas we used the population of the originating city to proportionally scale the counts. For each city, we computed the ratio between the number of people for which the city was considered "home" and its population (see Table 3.3 for a list of the 10 largest cities and their tweet volumes), and used it to scale the volume of the flow of people traveling to other cities.

[2]www.city-data.com

Biggest City	**Nearby Cities**
New York City (NY)	Elizabeth (NJ), Yonkers (NJ), Jersey City (NJ), Newark (NJ), Paterson (NJ)
Los Angeles (CA)	Long Beach (CA), Pasadena (CA), Glendale (CA)
Cambridge (MA)	Boston (MA), Somerville (MA), Lynn (MA)
San Francisco (CA)	Berkeley (CA)
Albany (NY)	Schenectady (NY)
Minneapolis (MN)	St. Paul (MN)
Tampa (FL)	St. Petersburg (FL)
Providence (RI)	Fall River (MA)
Waterbury (CT)	New Haven (CT)
Kansas City (MO)	Kansas City (KS)
Philadelphia (PA)	Camden (NJ)

Table 3.2: MMWR Cities removed due to overlapping metro areas

City	**Population**	**Home for**	**%**
New York City, NY	8406000	297486	3.54%
Los Angeles, CA	3857799	206378	5.35%
Chicago, IL	2714856	94952	3.50%
Houston, TX	2160821	71780	3.32%
Philadelphia, PA	1547607	77626	5.02%
Phoenix, AZ	1488750	42206	2.83%
San Antonio, TX	1382951	32312	2.34%
San Diego, CA	1338348	42659	3.19%
Dallas, TX	1241162	77844	6.27%
San Jose, CA	982765	27737	2.82%

Table 3.3: Population vs. Twitter Penetration - Top 10 cities

3.5.3 Flu Correlation between Cities: distance vs. flow

The first goal of this work was to demonstrate that a flow-based metric may better represented the similarity between the flu trends of two cities than distance alone. The idea that two nearby cities influence each other seems intuitively correct due to terrestrial travel but does not take into account travel over large distance which is presumably more likely to be air travels. Conversely, a distance matrix composed only using airline ticketing information neglects short-range movements.

Using geolocated tweets as source of movement, we not only capture intra-state travels done by car but also long-distance travel, presumably by air. In particular, our data represent the volume of individuals traveling between two cities, irrespective of the transportation method chosen.

A scatter plot of distance versus flu trends similarity for 2011 (Figure 3.16) shows very little correlation between the two metrics. While there seem to be a faded correlation, close cities do not necessarily show similar flu trends.

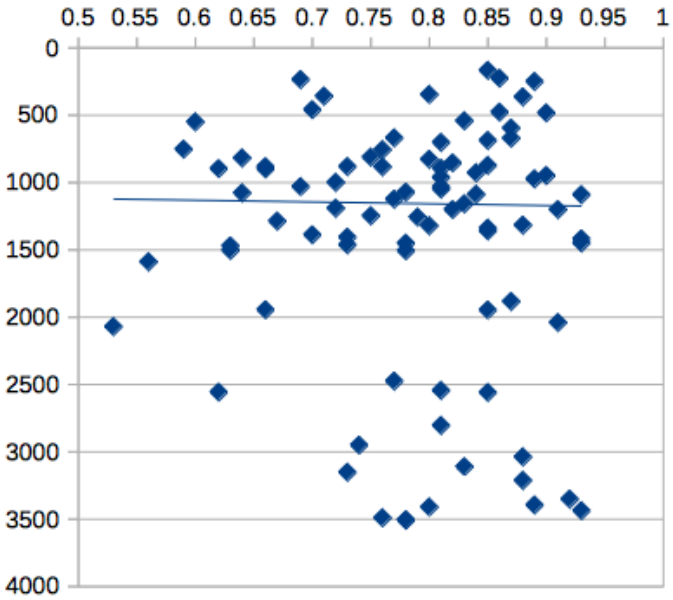


Figure 3.16: Distance vs. Correlation for Atlanta, GA

On the other hand, a scatter plot of the number of people traveling into the city versus the flu trends similarity (Figure 3.17) definitively shows a more visible trend.

3.5.4 Predicting Flu Trends across Cities

The second goal of this research was to determine if a travel model obtained through social media could in fact improve the prediction of flu trends. To verify this hypothesis, we obtained flu trends data (generated using the methods of Section 3.5.1) for 2011 and 2012 for each of the 78 cities left in the combined dataset.

For each city, we trained an SVM using 2011 data and we tested on 2012 data. The models were created through Support Vector Regression with a polynomial

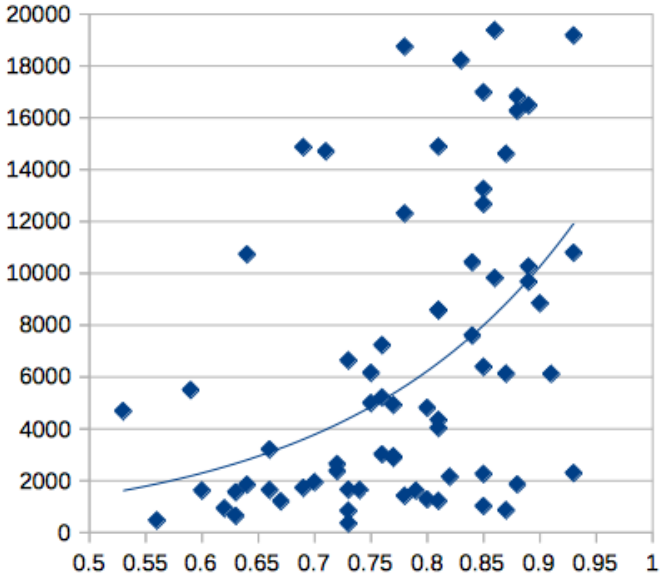


Figure 3.17: Flow vs. Correlation for Atlanta, GA

kernel using the well known LibSVM library. For each week, the target variable was the value of the flu trend for that week, and each feature was represented by the value of the flu trend in one of the other correlate cities two weeks before.

To compare the effectiveness of different approaches in selecting which cities should be used as features for each city, we decided to select the top-N (here, N=20) correlated cities according to the following conditions:

- **distance**, closest 20 cities
- **similarity**, cities with 2011 flu trend similarity greater than 0.8
- **flow**, top 20 cities with highest volume of travels towards the target city
- **combined**, top 10 cities from correlation and top 10 cities from flow

Averaging the results of all the cities, we obtain the following squared correlation coefficient for each approach:

	Distance	**Similarity**	**Flow**	**Combined**
Average	0.23086	0.26971	0.28080	0.28543
Minimum	0.00009	0.00004	0.00027	0.00001
Maximum	0.7780	0.73145	0.83247	0.79387

Table 3.4: Square Correlation Coefficients for each approach

While the averages are low, a flow based approach for selecting the most correlated cities yields a 21% improvement in the prediction with respect of pure distance. Combining 2011 similarity and flow yields a 23% improvement respect to distance alone. Table 3.5 shows the most predictable cities while Table 3.6 shows the most least predictable ones.

The poor performance attained by all methodologies when dealing with these cities seem to suggest unpredictability of their flu trends. Some possible explanations for that:

- **Port-of-entry**, Washington (DC), Philadelphia (PA), Dallas (TX) are all big port-of-entries for many international flights which makes their flu season highly independent from other places in the US.
- **Noisy data**, Waterbury (CT) has only a 43 deaths reported in the period considered by this study. Relatively speaking, the difference between 1 and 2 deaths is huge and highly unpredictable, especially when using these to infer flu trends.
- **Not enough data**, Fort Wayne (IN) has been selected as home town for only 6481 users, in comparison, Las Vegas (NV) is considered home town for almost 300,000 users.

	Distance	**Similarity**	**Flow**	**Combined**
Las Vegas, NV	0.757804	0.562152	0.83247	0.735552
San Jose, CA	0.602464	0.689917	0.801097	0.793866
Albuquerque, NM	0.556896	0.626601	0.711688	0.494305
Tucson, AZ	0.740066	0.5472	0.694016	0.661322
San Antonio, TX	0.467228	0.636067	0.686346	0.612461

Table 3.5: Square Correlation Coefficients for most predictable cities

	Distance	**Similarity**	**Flow**	**Combined**
Washington, DC	0.043178	0.066362	0.013697	0.001167
Philadelphia, PA	0.000334	0.057870	0.007827	0.002519
Dallas, TX	0.000088	0.024712	0.005816	0.001576
Waterbury, CT	0.061919	0.004790	0.002763	0.000006
Fort Wayne, IN	0.005689	0.000038	0.000272	0.000518

Table 3.6: Square Correlation Coefficients for most difficult cities

Despite these issues, the experiments performed in this work seem to indicate that (1) travel models generated through social media realistically represent both short and long range movements of people and that (2) the volume of people traveling between two cities more accurately correlates with the similarity between their flu trends than distance alone. Furthermore, our experiments seem to confirm that considering the flow of people moving between cities can help improve the accuracy of flu trends forecasting.

CHAPTER 4
CONCLUSION

Our results demonstrate that social media data can be used to track users's interest and concerns related to public topics (e.g., H1N1 influenza), and also to accurately estimate disease prevalence. Although from a descriptive perspective, since no comparable data (e.g., survey results) are available, it is not possible to really validate some of the results, the trends observed are reasonable and quite consistent with expectations. For example, Twitter users' initial interest in antiviral drugs such as oseltamivir dropped at about the same time as official disease reports indicated most cases were relatively mild in nature, despite the fact that overall the number of cases was still increasing. Also, interest in hand hygiene and face masks seemed to be timed with public health messages from the CDC about the outbreak in early May. Interestingly, in October of 2009, concern regarding shortages did not appear nor did interest in rare side effects, perhaps because they did not occur in any widespread fashion. Here, absence of a sustained detectable signal may indicate an apathetic public, or may simply indicate a lack of information in the media. In either case, our work proposes a mechanism to capture these concerns in real time, pending future studies to confirm our results using appropriate techniques for analyzing autocorrelated data.

Our research also demonstrates that it is possible to estimate people and disease activity (e.g., travel and flu movement) in real time using social data. While influenza is well known and reoccurs each season with regular cycles, its geographic location, timing, and size varies complicating efforts to produce reliable and timely

activity estimates using traditional time series models [80]. The literature provides several examples of "syndromic approaches" to anticipating or forecasting ILI, including analyses of telephone triage calls [31], purchases of over-the-counter medications for respiratory diseases [48] [23], and school absenteeism [58]. Although in theory it is possible to gather diagnosis-level data in near-real time from emergency department visits [53] [115], doing so at a national level would require fusing, at considerable expense, data sources from different geographic areas and multiple firms (in the case of pharmacy data or billing data). In addition, while these efforts can yield information about future influenza activity days to weeks in advance of traditional sources (e.g., ILI surveillance), it is difficult to compare these approaches, because different geographic regions were studied and different statistical approaches were used [22].

In contrast, our estimation method is based on well-understood machine learning methods and uses a publicly available stream of tweets as input. The accuracy of the resulting real-time ILI estimates clearly demonstrates that the subset of tweets identified and used in our models contains information closely associated with disease activity. Our results show that we were able to establish a distinct relationship between Twitter data and the epidemic curve of the 2009 H1N1 outbreak, both at a national level and within geographic regions. Our approach, in contrast to others [69], does not attempt to forecast influenza activity, but aims to provide real-time estimates. Yet because our results are available "live" (i.e., as soon as the data are captured), estimates are available sooner than traditional public health reports, which tend to lag ILI activity by 1 or 2 weeks.

In addition, Twitter and search data are easily and efficiently collected, and can be processed automatically in real time. While search-term data related to

influenza is more available than in the past to investigators outside search engine companies, we think that our Twitter-based approach provides some unique advantages:

- First, the Twitter data provide more contextual information than a corpus of search queries (i.e., lists of key words), so that they can be used to investigate more than just disease activity. Contextual cues also enable the retrospective study of ancillary issues, such as treatment side effects or potential medication shortages. For example, while monitoring the swine flu pandemic (Section **??**), we also investigated perceptions regarding pregnancy and influenza in direct response to a specific question from a state epidemiologist who was concerned that women might avoid the new H1N1 vaccine because of pregnancy-related concerns. It is important for public health officials to know about such opinions, beliefs, and perceptions as they develop, so as to craft more effective communication strategies.

- Second, Cooper et al. [21] found that daily variations of search frequency in search query data regarding cancer were heavily influenced by news reports, making search query data a necessarily "noisy" marker for actual disease activity. Because the entire tweet is available, this is less of a problem for Twitter-based analysis using the support-vector regression method espoused here, since terms will emerge during model fitting to ensure noisy tweets are excluded. Similar data-mining approaches could also be applied to search data, but require access to more context and state information (e.g., search histories rather than unlinked individual queries) than is generally made available to outside investigators by search-engine firms. This is largely because

releasing fine-grained search data raises significant privacy issues, especially if it can be linked to individuals across multiple searches. In contrast, all of the Twitter data used here is placed in the public domain by the issuing user who chooses to broadcast his or her tweets to the world at large: indeed, Twitter and the Library of Congress have future plans to make every public tweet ever issued available to any interested party.

Despite these promising results, there are several limitations to our study:

- First, the use of Twitter is neither uniform across time nor geography. Mondays are usually the busiest for Twitter traffic, while the fewest tweets are issued on Sundays; also, people in California and New York produce far more tweets per person than those in the Midwestern states (or, for that matter, in Europe). When and where tweets are less frequent (or where only a subset of tweets contain geographic information), the performance of our model may suffer. The difference in accuracy at a national level and regional level observed in the Results could, in part, be explained by this lack of data. While the national model used, on average, 120,000 weekly tweets to make its weekly predictions, the regional one had only 3,000.

- A second limitation is that we only had one year of sampled data. More seasons, especially non-pandemic seasons, should help improve the accuracy of our ILI estimates, as would more complete access to the Twitter stream.

- Third, the demographic of Twitter users do not represent the general population, and in fact, the exact demographics of the Twitter population, especially the Twitter population that would tweet about health related concerns, is unknown and not easy to estimate. Finally, we need to determine how accurately

Twitter can estimate other population-based measures of influenza activity.

If future results are consistent with our findings, Twitter-based surveillance efforts like ours and similar efforts underway in two European research groups [85] [57] may provide an important and cost-effective supplement to traditional disease-surveillance systems, especially in areas of the United States where tweet density is high. We propose that Twitter data can also be used as a proxy measure of the effectiveness of pubic health messaging or public health campaigns. Our ability to detect trends and confirm observations from traditional surveillance approaches make this new form of surveillance a promising area of research at the interface between computer science, epidemiology, and medicine.

APPENDIX
LIST OF KEYWORDS

A.1 Keywords used in study of Section 2.1.7

aerobics, aikido, badminton, balanced, barbell, baseball, basketball, "baton twirling", bb, bball, b-ball, bicycle, bicycling, bike, biking, boat, bowling, canoe, canoeing, cardio, circuit, climb, climbed, climbing, cross county, dance, dancing, dodgeball, dumbbell, elliptical, exercise, fitness, football, golf, gym, gymnastics, hapkido, hike, hiked, hiking, hockey, hunt, hunting, interval, jazzercise, jog, jogged, jogging, judo, karate, kickball, lacrosse, martial, muscle, physical activity, pull-up, pump, push-up, racquetball, ran, rec, recreation, resistance, rock, row, rowing, run, salsa, sit-up, skate, skating, ski, skiing, soccer, softball, squash, squats, step, strengthen, stretch, stretching, swim, swimming, table tennis, tae kwon do, tennis, "track & field", treadmill, ultimate, upper class, vb, vball, v-ball, viffleball, volleyball, walk, walked, walking, water, weight, workout, wrestling, yoga, zumba

A.2 List of suffixes removed in step 5 of Porter's Algorithm

-al, -ance, -ence, -er, -ic, -able, -ible, -ant, -ement, -ment, -ent, -ion, -ou, -ism, -ate, -iti, -ous, -ive, -ize

REFERENCES

[1] The Verge Adrianna Jefferis. The man behind flickr on making the service 'awesome again'. http://www.theverge.com/2013/3/20/4121574/flickr/-chief-markus-spiering-talks-photos-and-marissa-mayer, March 2013.

[2] Alexa. Top health fitness sites. http://www.alexa.com/topsites/category/Top/Health/Fitness/News_and_Media/Magazines_and_E-zines, October 2014.

[3] Alexa. Top health sites. http://www.alexa.com/topsites/category/Top/Health/, October 2014.

[4] Dr. Carolina Arana. Types of surveillance systems. http://publichealth/observer.com/types-of-surveillance-systems/, October 2014.

[5] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10):203–224, 2007.

[6] Carina G M Blackmore, Lillian M Stark, William C Jeter, Robin L Oliveri, Robert G Brooks, Lisa A Conti, and Steven T Wiersma. Surveillance results from the first west nile virus transmission season in florida, 2001. *Am J Trop Med Hyg*, 69(2):141–150, Aug 2003.

[7] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[8] William B. Cavnar and John M. Trenkle. N-gram-based text categorization.

In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[9] Ann Cavoukian. Data mining: Staking a claim on your privacy. Technical report, Ontario Information and Privacy Commissioner, 1998.

[10] CDC. Nedss. `http://wwwn.cdc.gov/nndss/script/nedss.aspx`, October 2014.

[11] Center for Disease Control and Prevention. Overview of influenza surveillance in the united states. `http://www.cdc.gov/flu/weekly/overview.htm`.

[12] Center for Disease Control and Prevention. Weekly u.s. influenza surveillance report. `http://www.cdc.gov/flu/weekly/`.

[13] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machine. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`, 2011.

[14] Paula Chesley, Bruce Vincent, Li Xu, and Rohini K Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Training*, 580(263):233, 2006.

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

[16] Rumi Chunara, Jason R Andrews, and John S Brownstein. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*,

86(1):39–45, 2012.

[17] Cisco. Connected world technology report. http://www.cisco.com/c/en/us/solutions/enterprise/connected-world-technology-report/index.html, 2012.

[18] The New York Times Claire Cain Miller. Starbucks fans can become 'baristas' on foursquare. http://bits.blogs.nytimes.com/2010/03/11/starbucks/-fans-can-become-a-barista-on-foursquare/, March 2010.

[19] CNN. Virginia tech shooting - facts. http://www.cnn.com/2013/10/31/us/virginia-tech-shootings-fast-facts/, October 2013.

[20] Consumer Healthcare Product Association. Your health at hand: Perceptions of over-the-counter medicine in the u.s. http://www.yourhealth.athand.org/images/uploads/CHPA_YHH_Survey_062011.pdf, June 2011.

[21] Crystale Purvis Cooper, Kenneth P Mallon, Steven Leadbetter, Lori A Pollack, and Lucy A Peipins. Cancer internet search activity on a major search engine, united states 2001-2003. *Journal of medical Internet research*, 7(3), 2005.

[22] Lynne Dailey, Rochelle E Watkins, and Aileen J Plant. Timeliness of data sources used for influenza surveillance. *Journal of the American Medical Informatics Association*, 14(5):626–631, 2007.

[23] GR Davies and RG Finch. Sales of over-the-counter remedies as an early warning system for winter bed crises. *Clinical microbiology and infection*, 9(8):858–863, 2003.

[24] Department of Health and Human Services. Health information privacy. http://www.hhs.gov/ocr/privacy/.

[25] Tahar Dilekh and Ali Behloul. Implementation of a new hybrid method for stemming of arabic text. *analysis*, 3(4):5, 2012.

[26] Maeve Duggan and Joanna Brenner. Social networking site users. http://www.pewinternet.org/2013/02/14/social-networking-site-users/, February 2013.

[27] Duke Medicine News and Communication. Monkeys consciously control a robot arm using only brain signals; appear to "assimilate" arm as if it were their own, 2003.

[28] Megs Mahoney Dusil. 30 items i'd like for my 30th birthday. http://www.purseblog.com/general/30-items-id-like-for-my-30th-birthday, September 2013.

[29] White M. E. and S. M. McDonnell. Public health surveillance in low and middle income countries. In S. M. Teutsch and R. E. Churchill, editors, *Principles and Practices of Public Health Surveillance*, pages 287–315. Oxford University Press, 2000.

[30] Joshua M Epstein, D Michael Goedecke, Feng Yu, Robert J Morris, Diane K Wagener, and Georgiy V Bobashev. Controlling pandemic flu: the value of international air travel restrictions. *PloS one*, 2(5):e401, 2007.

[31] Jeremy U Espino, William R Hogan, and Michael M Wagner. Telephone triage: a timely data source for surveillance of influenza-like diseases. In

AMIA Annual Symposium Proceedings, volume 2003, page 215. American Medical Informatics Association, 2003.

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

[33] Facebook. Company info. `https://newsroom.fb.com/company-info/`, October 2014.

[34] Facebook. Facebook login overview. `https://developers.facebook.com/docs/facebook-login/overview/v2.1`, October 2014.

[35] Facebook, Ericsson and Qualcomm. A focus on efficiency. Technical report, Facebook, Ericsson and Qualcomm, September 2013.

[36] Geoffrey Fairchild. *Improving Disease Surveillance: Sentinel Surveillance Network Design and Novel Uses of Wikipedia*. PhD thesis, Department of Computer Science, University of Iowa, November 2014.

[37] FDA. Adverse event reporting system (faers). `http://www.fda.gov/Drugs/GuidanceComplianceRegulatory/Information/Surveillance/Adverse/Drug/Effects`, October 2014.

[38] W H Foege, R C Hogan, and L H Newton. Surveillance projects for selected diseases. *Int J Epidemiol*, 5(1):29–37, Mar 1976.

[39] Foursquare. 2010: Our year of 3400 `https://foursquare.com/infographics/2010infographic`, 2010.

[40] Foursquare. About. https://foursquare.com/about, October 2014.

[41] Susannah Fox and Kristen Purcell. Social media and health. Technical report, Pew Research, 2010.

[42] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

[43] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[44] Google. The google maps geolocation api. https://developers.google.com/maps/documentation/business/geolocation/, October 2014.

[45] B. T. Grenfell, O. N. Bjornstad, and J. Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 12 2001.

[46] BT Grenfell, ON Bjørnstad, and J Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[47] Lara Hejtmanek. American idol winner: Can google predict the results? *Mashable*, May 2009.

[48] William R Hogan, Fu-Chiang Tsui, Oleg Ivanov, Per H Gesteland, Shaun Grannis, J Marc Overhage, J Michael Robinson, and Michael M Wagner. Detection of pediatric respiratory and diarrheal outbreaks from sales of over-the-counter electrolyte products. *Journal of the American Medical Informatics*

Association, 10(6):555–562, 2003.

[49] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.

[50] IgniteSpot. Small business marketing idea. `http://www.ignitespot.com/small-business-marketing-idea/`, October 2014.

[51] International Health Terminology Standards Development Organization. Snomed clinical terms. `http://www.ihtsdo.org/snomed-ct`, October 2014.

[52] Internet World Stats. Internet world statistics. `http://www.internetworld/stats.com/stats.htm`, October 2014.

[53] Charlene Babcock Irvin, Patricia Petrella Nouhan, and Kimberly Rice. Syndromic analysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance. *Annals of emergency medicine*, 41(4):447–452, 2003.

[54] Heather A Johnson, Michael M Wagner, William R Hogan, Wendy Chapman, Robert T Olszewski, John Dowling, Gary Barnas, et al. Analysis of web access logs for surveillance of influenza. *Stud Health Technol Inform*, 107(Pt 2):1202–1206, 2004.

[55] Anne Marie Kelly. Downloading tv programs, watching videos, and making online phone calls represent the biggest one-year internet activity increase. Technical report, MediaMark Research and Intelligence, 2008.

[56] Ryan Kelly. Twitter study reveals interesting results about usage 40pointless babble. Technical report, Pear Analytics, 2009.

[57] Vasileios Lampos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

[58] Dennis D Lenaway and Audrey Ambler. Evaluation of a school-based influenza surveillance system. *Public health reports*, 110(3):333, 1995.

[59] S Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24(4):349–53, 2003.

[60] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[61] MDG Advertising. Images account for 36 percent of all twitter links shared. Technical report, MDG Advertising, 2013.

[62] Gilad Mishne and Maarten de Rijke Gilad. Moodviews: Tools for blog mood analysis. *ICWSM*, 2007.

[63] U.S. Department of Transportation's Bureau of Transportation Statistics. December 2013 u.s. airline systemwide passengers up 6.1 percent from december 2012. `http://www.rita.dot.gov/bts/press_releases/bts012_14`, March 2014.

[64] Marguerite Pappaioanou, Michael Malison, Karen Wilkins, Bradley Otto, Richard A Goodman, R Elliott Churchill, Mark White, and Stephen B Thacker. Strengthening capacity in developing countries for evidence-based public health: the data for decision-making project. *Soc Sci Med*, 57(10):1925–1937, Nov 2003.

[65] Gonzalo Parra, Steve Dyrdahl, and Brian Goetz. Java implementation of the original porter algorithm. http://tartarus.org/martin/PorterStemmer/java.txt, 2000.

[66] PEW Internet. Health fact sheet. `http://www.pewinternet.org/fact/-sheets/health-fact-sheet/`, 10 2014.

[67] PEW Internet. Social networking: Fact sheet. `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`, October 2014.

[68] Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[69] Philip M Polgreen, Forrest D Nelson, George R Neumann, and Robert A Weinstein. Use of prediction markets to forecast infectious disease activity. *Clinical Infectious Diseases*, 44(2):272–279, 2007.

[70] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[71] Jagdish Rebello. Four out of five cell phones to integrate gps by end of 2011. Technical report, iHS Technology, July 2010.

[72] Regenstrief. Logical observation identifier names and codes. http://loinc.org/, October 2014.

[73] Birkhead G. S. and C. M. Maylahn. State and local public health surveillance. *Principles and Practices of Public Health Surveillance*, page 270, 2000.

[74] Jeff Schneider. Cross validation. http://www.cs.cmu.edu/~schneide/tut5/node42.html, September 1998.

[75] Alberto Maria Segre, Andrew Wildenberg, Veronica Vieland, and Ying Zhang. Privacy-preserving data set union. In *Privacy in Statistical Databases*, pages 266–276. Springer, 2006.

[76] Sichuan Online. Qq posting girl help helicopter airborne wenchuan. http://news.xinhuanet.com/society/2008-05/18/content_8198824.htm, May 2008.

[77] Ilia Smirnov. Overview of stemming algorithms. *Mechanical Translation*, 2008.

[78] Kate Starbird, Leysia Palen, Amanda L Hughes, and Sarah Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2010.

[79] Statistics Brain. Facebook statistics. http://www.statisticbrain.com/facebook-statistics/, July 2013.

[80] Donna F Stroup, Stephen B Thacker, and Joy L Herndon. Application of

multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962–1983. *Statistics in Medicine*, 7(10):1045–1059, 1988.

[81] Sysomos. Haitian earthquake dominates twitter. `http://blog.sysomos.com/2010/01/15/haitian-earthquake-dominates-twitter/`, January 2010.

[82] Janice Tanne. Restricting air travel may slow spread of flu. *BMJ*, 333(7568):568, 2006.

[83] Brad Templeton. Origin of the term "spam" to mean net abuse. `http://www.templetons.com/brad/spamterm.html`.

[84] S B Thacker and R L Berkelman. Public health surveillance in the united states. *Epidemiol Rev*, 10:164–190, 1988. PIP: TJ: EPIDEMIOLOGIC REVIEWS.

[85] The International Council on Medical & Care Compunetics. The potential of twitter for early warning and outbreak detection, April 2010.

[86] The Nielsen Company. What's empowering the new digital consumer? Technical report, The Nielsen Company, 2014.

[87] Topsy. 2.5m tweets an hour as news of whitney houston's death spreads. `http://topsylabs.com/2012/02/12/2-5-million-tweets-an-hour`, February 2012.

[88] Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In *Proc. 20th Machine Learning conference of Belgium and The Netherlands*, pages 27–34, 2011.

[89] Twitter, Inc. Streaming api. https://dev.twitter.com/docs/api/streaming, October 2009.

[90] Twitter Inc. Healing haiti. https://blog.twitter.com/2010/healing/-haiti, 2010.

[91] Twitter, Inc. Streaming api: Location based filters. https://dev.twitter.com/docs/streaming-apis/parameters#locations, September 2010.

[92] Twitter, Inc. #superbowl. https://blog.twitter.com/2011/superbowl, February 2011.

[93] Twitter, Inc. About twitter, inc. https://about.twitter.com/company, October 2014.

[94] Twitter, Inc. Api overview: Tweets. https://dev.twitter.com/overview/api/tweets, October 2014.

[95] Twitter, Inc. Rest api: Get favorites list. https://dev.twitter.com/rest/reference/get/favorites/list, October 2014.

[96] Twitter, Inc. Rest api: Rate limiting. https://dev.twitter.com/docs/rate-limiting/1.1, October 2014.

[97] Twitter, Inc. Rest api: Search tweets. https://dev.twitter.com/docs/api/1.1/get/search/tweets, October 2014.

[98] U.S. Agency for International Development. Infectious disease and response strategy 2005. Washington, DC., 2005.

[99] U.S. Census Bureau. Computer and internet use in the united states, May 2013.

[100] U.S. Department of State. Mexico travel alert: H1n1 flu update. `https://blogs.state.gov/stories/2009/04/28/mexico-travel-alert-h1n1`, April 2009.

[101] Sarah Vieweg, Leysia Palen, Sophia B Liu, Amanda L Hughes, and Jeannette Sutton. Collective intelligence in disaster: An examination of the phenomenon in the aftermath of the 2007 virginia tech shootings. In *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*, 2008.

[102] W3 Techs. Usage statistics and market share of wordpress for websites. Technical report, W3 Techs, 2014.

[103] Wall Street Journal. Foursquare week. `http://graphicsweb.wsj.com/docs/FOURSQUAREWEEK1104/bygender.php`, 2011.

[104] Mark Walsh. Search wrong on "idol". *MediaPost*, May 2009.

[105] Sholom M. Weiss and Casimir A. Kulikowski. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1991.

[106] WikiMedia Foundation. Strategic plan 2011. `http://upload.wikimedia.org/wikipedia/foundation/c/c0/WMF_StrategicPlan2011_spreads.pdf`, October 2011.

[107] Wikipedia. Wikipedia - number of editors. http://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Number_of_editors, October 2014.

[108] Wordpress. A live look at activity across wordpress.com. http://en.wordpress.com/stats/, October 2014.

[109] World Bank. World development report 2000–2001: Attacking poverty. Technical report, World Bank, 2001.

[110] World Health Organization. Epidemiological surveillance of communicable disease at the district level. In *WHO Regional Committee for Africa, 43rd session*, 1993.

[111] Wired Xeni Jardin. Text messaging feeds sars rumors. http://archive.wired.com/medtech/health/news/2003/04/58506?currentPage=all, April 2003.

[112] Philip Fei Wu Yan Qu and Xiaoqing Wang. Online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake online community response to major disaster: A study of tianya forum in the 2008 sichuan earthquake. *HICSS-42*, 2009.

[113] VA Yatsko. Methods and algorithms for automatic text analysis. *Automatic Documentation and Mathematical Linguistics*, 45(5):224–231, 2011.

[114] Haiqing Yu. The power of thumbs: The politics of sms in urban china. In *Graduate Journal of Asia-Pacific Studies*, volume 2:2, pages 30–43, 2004.

[115] Christine M Yuan, S Love, and M Wilson. Syndromic surveillance at hospital emergency departments?southeastern virginia. *Morbidity and Mortality Weekly Report*, pages 56–58, 2004.

[116] Ni Zhang, Shelly Campo, Kathleen F Janz, Petya Eckler, Jingzhen Yang, Linda G Snetselaar, and Alessio Signorini. Electronic word of mouth on twitter about physical activity in the united states: exploratory infodemiology study. *Journal of medical Internet research*, 15(11), 2013.