A Project Report

on

DISEASE PREDICTION USING SOCIAL MEDIA DATA

Submitted in partial fulfillment of the
requirement for the award of the degree of

B. Tech Computer Science And Engineering

Under The Supervision of
Dr. Munish Sabharwal
Designation- Dean (SCSE), Professor

Submitted By

Govinda Agarwal 18SCSE1010020
Nikhil Pandey 18SCSE1010216

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA
INDIA
2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE'S DECLARATION

We hereby certify that the work which is being presented in the thesis, entitled **"DISEASE PREDICTION USING SOCIAL MEDIA DATA"** in partial fulfillment of the requirements for the award of the B.Tech submitted in the School of Computing Science and Engineering of Galgotias University, Greater Noida, is an original work carried out during the period of August 2021 to December 2021, under the supervision of Dr. Munish Sabharwal, Dean & Professor in Department of Computer Science and Engineering/Computer Application and Information and Science, of School of Computing Science and Engineering, Galgotias University, Greater Noida

The matter presented in the thesis/project/dissertation has not been submitted by me/us for the award of any other degree of this or any other places.

Govinda Agarwal

Nikhil Pandey

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Munish Sabharwal

Dean (SCSE), Professor

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **18SCSE1010020 – GOVINDA AGARWAL, 18SCSE1010216 – NIKHIL PANDEY** has been held on _____and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place:

ACKNOWLEDGEMENT

It is our privilege and solemn duty to express our deepest sense of gratitude to Dr. Munish Sabharwal, under whose guidance we carried out this work. We are indebted to him for his invaluable supervision, heart full cooperation and timely aid and advice till the completion of the report in spite of his pressing engagements. We wish to record our sincere gratitude for his constant support and encouragement in preparation of this report.

We take this opportunity to express our hearty thanks to all those who helped us in the completion of our project work. We are incredibly grateful to the author of various research papers, for helping us become aware of the research currently ongoing in the field. We are very thankful to our parents for their constant support and love.

Last, but not least, we would like to thank our classmates for their valuable comments, suggestions and unconditional support.

ABSTRACT

Over the previous few decades, the world we live in has altered dramatically. The creation of public health monitoring systems was prompted by threats of bioterrorism, influenza pandemics, and developing infectious diseases, as well as unprecedented population movement. These systems are useful for detecting and responding to infectious disease outbreaks, but they frequently function with significant delays and do not give the necessary lead time for optimal public health response. To warn of changes in disease activity, syndromic monitoring systems rely on clinical traits (e.g., activities prompted by the development of symptoms) that are observable prior to diagnosis. These techniques, while less exact, can provide significant lead time. Patient data can be obtained from a variety of existing sources set up for various purposes, such as emergency department primary complaints, ambulance dispatch data, and over-the-counter medicine sales. Unfortunately, these data are frequently costly, difficult to get, and nearly impossible to combine.

Fortunately, with the rise of online social networks, considerably more information about our everyday routines and lives is now freely available and easily accessible on the internet. Twitter, Facebook, and Foursquare are just a few of the many websites where users freely share information about their everyday activities, health, and physical position. In order to make predictions, we create and implement methods for collecting, filtering, and analyzing the content of social media postings in this thesis. We modelled human trips using location-specific social media data and demonstrated how this data might help us better anticipate disease burden.

CONTENTS

CONTENTS

List of Tables

List of Figures

ABBREVIATIONS

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
BRFSS	Behavior Risk Factor Surveillance System
CDC	Centers for Disease Control and Prevention
COVID-19	Coronavirus Disease 2019
ELR	Electronic Laboratory Reporting
FAERS	FDA Adverse Events Reporting System
FDA	Food and Drug Administration
FETP	Field Epidemiology Training Program
GMT	Greenwich Mean Time
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HIPAA	Health Insurance Portability and Accountability Act
HIV	Human Immunodeficiency Virus
HL7	Health Level Seven
HTML	Hypertext Markup Language
IDSR	Integrated Disease Surveillance and Response
ILI	Influenza-like Illness
JPA	Java Persistence API
LOINC	Logical Observation Identifiers Names and Codes
NEDSS	National Electronic Disease Surveillance System
NETSS	National Electronic Telecommunications System for Surveillance
NLP	Natural language processing
OSN	Orbit Showtime Network
PDOH	Philippine Department of Health
PHIN	Public Health Information Network
SARS	Severe Acute Respiratory Syndrome
SMS	Short Message Service
SNOMED	Systematized Nomenclature of Medicine
TB	Tuberculosis

ABBREVIATIONS

URL	Uniform Resource Locator
USA	United States of America
USAID	United States Agency for International Development
VAERS	Vaccine Adverse Events Reporting System
WHO	World Health Organization
YRBS	Youth Risk Behavior Survey

CHAPTER 1

INTRODUCTION

The world during which we live has changed rapidly over the previous couple of 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 [1] these systems perform an ”ongoing systematic collection, analysis, and interpretation of knowledge , closely integrated with the timely dissemination of those 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.

Surveillance of public health is primarily used to inform interventions. The monitoring systems put in place are typically intended to collect scientific and factual data necessary for making informed decisions and planning appropriate public health actions, and their design and execution are often affected by their goals. Diverse public health goals and the activities required to achieve them may necessitate the use of different information systems. The sort of surveillance or health information system to be used is determined by the type of action to be taken, when and how often it must be performed, what information is required to conduct or monitor the action, and how frequently the information is required. If the aim is to avoid the spread of acute infectious illnesses (such as SARS), the monitoring system must be successful in detecting early warning signals so that management may respond immediately and avert epidemics. Surveillance of chronic illnesses (e.g. TB) or health-related behaviors (e.g. cigarette smoking) with a modest change rate, on the other hand, may be done easily by annual demographic and health surveys.

1.1 IMPORTANCE OFSURVEILLANCE

The World Health Organization (WHO) and the World Bank believe [2] surveillance to be an important role of a public health system, since it improves the efficiency and efficacy of services delivered through targeted interventions and recording of population impacts. Since 1975, the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) have worked with more than 30 nations to improve health systems and meet disease detection and response training requirements in a country-specific, flexible, and long-term way. WHO members must adhere to the International Health Regulations' standards and have essential personnel and core surveillance capabilities.

The Integrated Disease Surveillance and Response (IDSR) approach [110] was established by the WHO (in Africa) in 1993, and it connected epidemiological and laboratory data at all levels of the health system, with an emphasis on integrating surveillance and response. Detection, registration, and confirmation of case-patients, reporting, analysis, and utilization of data, epidemic investigations, and contact tracing were all part of the strategy.

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.

China established its first Field Epidemiology Training Program (FETP) in 2005 to swiftly increase its monitoring and response capability, whereas Brazil and Argentina decided to enhance their own systems using World Bank money. At the same time, the United States Agency for International Development (USAID) changed its surveillance strategy to focus on the use of data to enhance public health interventions as more data became accessible via multiple channels [98]. Many countries adapted these new systems to their local realities: one example is Guatemala's marriage of its FETP (part of a larger, Central American FETP) with the Data for Decision Making programme [64], as well as India, which has a decentralized system, complex cultural and population dynamics, and a wide

range of public health institution sophistication, presents a different strategy for bolstering national surveillance.

As of today more than half of the 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 training the majority of the public health workers in the sector.

1.2 TYPES OF SURVEILLANCESYSTEMS

Foege and colleagues said in a 1976 paper [38] published in the International Journal of Epidemiology that the purpose of collecting, evaluating, and distributing information about a disease is to control that illness. If no action is taken, collection and analysis should not be permitted to spend resources. Public health surveillance systems should be set up in such a manner that they give decision makers accurate and timely data at the lowest feasible cost. On the basis of the activities that may be performed, the utility of the data obtained can be classified as immediate, yearly, or archival. Similarly, in order to enhance timeliness and conserve costs, geographical resolution of the data gathered (e.g., macro vs. micro regions) may be compromised. Complex surveillance systems are not always practical or successful for these reasons. In poor nations, for example, ensuring the quality and efficacy of monitoring in decentralized contexts is a major problem. National-level programme and surveillance system managers may lose control over the quality and timeliness of data gathered, and funders may develop parallel nongovernmental surveillance systems to collect the data they want directly if they see a vulnerability in the national system. These methods typically function in the short term, but in the long run, they exacerbate the weaknesses of existing public health monitoring programmes.

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 an indicator of overall population health. Infant mortality rate(the5 number of deaths among infants per 1,000 births) is also used as a 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.

1.2.1 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 certain conditions and are generally used to improve prevention programs.

1.2.2 Population Surveys

Routine surveys are surveillance tools that 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 data is used to track changes in health behavior (for example, the YRBS revealed a decrease in teenage smoking from 36% in 1997 to 20% in 2007), develop public health initiatives, and assess national and state public health policy.

1.2.3 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 varies by state. States report nationally notifiable diseases to the CDC on a voluntary basis.

1.2.4 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 (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.

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

1.2.6 Zoonotic Disease Surveillance

Zoonotic surveillance systems involve systems for detecting animals infected with diseases that can be transmitted to humans. Operations like this were highly effective [6] in 2001 during a West Nile Virus

(WNV) epidemic in Florida, and led to public health actions including urging the public to protect themselves from mosquito bites and strengthening mosquito abatement efforts.

1.2.7 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.

1.2.8 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.

1.2.9 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, determining whether a case satisfies public health surveillance case criteria and figuring out how to fill out the broad range of

forms issued by the CDC and health departments is usually left to the health provider personnel (who are frequently already overworked). In certain circumstances, the staff may need to devote a substantial amount of time to locating all of the documents that must be attached to the report. As a result, many illnesses go unreported, are insufficiently documented, or are recorded incorrectly.

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 designed to replace and combine many current CDC surveillance systems, including the National Electronic Telecommunications System for Surveillance (NETSS), HIV/AIDS reporting systems, immunization programmes, and TB and other infectious disease monitoring systems.

The National Electronic Disease Surveillance System (NEDSS) is a secure online platform that allows healthcare professionals and government organizations to communicate about disease trends and coordinate national epidemic responses. A collection of specifications for software, hardware, databases, and data format standards are included in the framework. 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 made up of five major components:

- A Web-based module that provides for quick online entry and administration of data sets, such as demographic and illness information;
- Silverstream is a Web application server that supports various Web-based modules;
- An integrated database management system;
- Messaging software (i.e., HL7 Standard) that allows electronic data interchange between state agencies and the CDC or state laboratories; and
- Intranet-based authentication and authorization for complete compliance with HIPAA standards.

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 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 the 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:

- An Internet browser-based system is used to enter illness data;
- Electronic Laboratory Reporting (ELR);
- Bringing together numerous health-related 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.

1.3 INTRODUCTION TO SOCIALMEDIA

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 Media mark Research Survey done in fall 2008 [55], many tasks that were once done manually are now typically completed online.

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

As the Internet has grown in popularity, so has the amount of information available to the general public on any subject. Many people regard the internet as a massive (free) library where they can find anything. In fact, several schools and teachers have had to implement stringent no-Internet-references regulations in their classrooms, compelling students to discover “real” sources for their projects as references. 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. Researchers in Italy may readily exchange their data with groups in Tokyo or compare their findings with the early results of comparable experiments conducted in Canada, all in real time, thanks to the Internet. Experiments like these are already taking place: in November 2000, a chimpanzee at Duke University in North Carolina was connected over the Internet to a robotic arm at the Massachusetts

Institute of Technology (MIT) Touch Lab, which was located over 600 miles distant [27]. Planet and star maps, weather predictions and histories, geological charts and pictures that were formerly exclusively available to a select group of scientists and graduate students may now be obtained online in a matter of minutes by everyone.

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. Small laboratory research, which were just a few years ago only published in low-circulation venues, now attract a lot of attention owing to enthusiastic bloggers who raise interest in these findings and make them widely available.

Other websites cover a wide range of possible medically-related necessities. Fitness and weight lossare 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 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. Patientlike is one of the most well-known sites in this category, with over 250,000 users 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 Hurricane Katrina hit the US Gulf Coast, 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. A local newspaper, The Times-Picayune, developed neighborhood-specific discussion boards and supplied maps and satellite imagery. This was the first catastrophe in which group administrators attempted more organized posting processes, although they were greeted with minimal 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 Ignite Spot 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 reach great popularity and become fully developed online magazines (e.g. Mashable).

Blogs are not the only way people share and collaborate on content on the Internet. The most well-known collaborative effort is Wikipedia, a free encyclopedia created, edited and updated by users around the world. Since its creation in 2001, Wikipedia has attracted more than 75,000editors [107] who have created more than 15 million documents. Despite the fact that the English version has the most papers (4.6 million), many of the articles are available in 260 languages and are accessed by over 400 million unique visitors each month [106].

Other social networking sites, like Wikipedia, have found fruitful ground on the Internet in 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 such as Facebook and Twitter have exploded in popularity, attracting billions of users of all ages. Users make (or re-make) connections with friends, partners, and coworkers, exchanging photographs, videos, and other personal data.

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 exploits this hardware in order to incorporate geographic information into social network traffic. This advancement allows individuals to focus on "hyper-local discussions," and apps like FourSquare8 encourage people to announce their location (i.e., "checkin") in order to connect with nearby friends and get discounts for their devotion to a brand or a certain store.

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 have been two well-publicized success cases in which Tianya members have given authorities vital information. The military was attempting to send aid to a remote region in the first occurrence [76], but they were unable to identify a suitable landing strip and had to postpone their operations. 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 investigated the situation and punished the offending individuals.

1.3.1 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 sorts 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, Daily Beast), 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, 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 Mood Views [62].Gilad used LiveJournal11 to gather about 850,000 mood-annotated blog posts and attempted to discoverdiscriminativecharacteristics(andtheirweights)inthepost'scontentfor each mood.

Unfortunately, the precision attained by the approach evaluated is only somewhat better (67 percent) than the baseline (50 percent, or random guessing), indicating that additional effort is needed to make it practical.

Similar work has 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 Wiktionary 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 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.

1.3.2 Wikipedia

The non-profit Wikimedia Foundation supports Wikipedia, a collaboratively updated, multilingual, free Internet encyclopedia. Jimmy Wales and Larry Sanger founded it in January 2001, and it now has over 18 billion page views. Today it contains more than 15 million articles written in 267 languages [106].Wikipedia is one of the most important sources of knowledge on the Internet, and its pages frequently show in search engine results as one of the top three URLs. Thousands [107] of dedicated volunteers from all around the globe contribute to this free encyclopedia by continually creating, updating, and perfecting its entries. In recent years, Wikipedia has been quick to respond to new trends and themes. Deaths of celebrities and major political events are often captured on its pages only a few minutes after the corresponding event. During the recent swine flu pandemic, for example, information on recent events was published in the “Swine Influenza” article on April 24th, barely minutes after the initial CDC public statement, and a dedicated article was established the next day. During its first five days on Wikipedia, the page titled "2009 Swine Flu Outbreak" had 1.5 million visitors, with a high of 417,200 on April 29th. Figure 1.1 shows the distribution of page views for the page ”H1N1” forth

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].

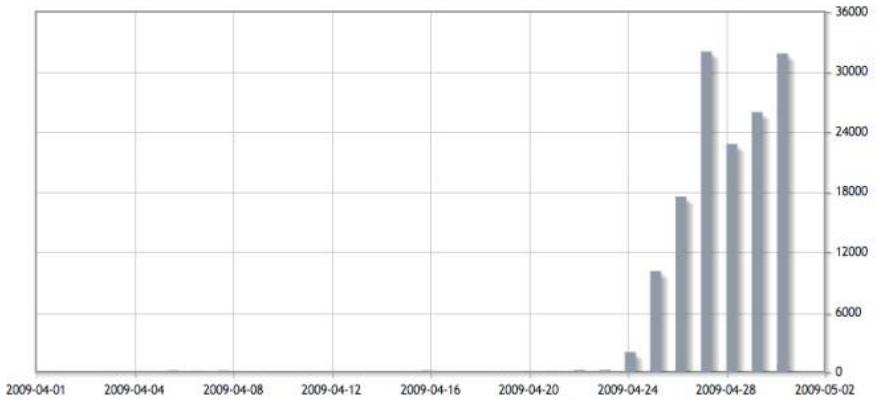

Figure 1.1: Number of Visits to "2009 Swine Flu Outbreaks" page on Wikipedia

1.3.3 Twitter

Twitter is a microblogging and social networking site that allows users to post brief text messages. It was founded by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass in July 2006. After a sluggish start, the site quickly grew in popularity, with 645 million registered users posting more than 500 million tweets each day in 2014 [93]. Tweets are the platform's messages, which may be transmitted through the web, a mobile app, or a text message. They can be shared individually or publicly and are limited to 140 characters. Users may "follow" other users to be notified of all of 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, 22 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 followers.

Hashtags that are used at increasingly greater rates than other hashtags are said to be ”trending”. Users can make a concentrated attempt to make a trending subject popular, or an event might drive others to speak about a certain topic. These subjects assist Twitter users in understanding what is going on in the globe. Despite the fact that people are becoming more conscious of their privacy rights at home and on the phone, a surprising number of people happily embrace online social networks and use them as an integral part of their daily lives, routinely sharing information about themselves, their moods, their activities, and so on. On Twitter or Facebook, it's not uncommon to witness extended discussions regarding a range of current events. 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, More than 4,000 tweets were posted every second, according to Twitter [Twitter, Inc. #superbowl. https://blog.twitter.com/2011/superbowl, February2011.

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, a number of community websites have arisen or have shifted their focus to assist locals. Locals may contribute news items, debate evacuation routes and fire prevention measures on discussion boards, and upload maps of the surrounding region on Rimoftheworld.net, a long-running community website for San Bernardino residents. 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 gathering24 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 account for a major portion of daily tweets, many people also use these social media platforms to communicate their moods, thoughts, and personal issues with their peers. During the 2009 H1N1 outbreak, for example, it was common to see people expressing their theories or conjectures about public health announcements (e.g.," CDC Data Shows H1N1 Vaccine Perfect for Population Control" 13), their fear of contracting the disease (e.g.," Im feeling sick... Hope it's not H1N1"), and panic-induced actions (e.g.," Got 2 facial masks but nobodies").

Example: 2009 Red River Flooding

Several researchers took the opportunity to study Red River-related Twitter activity during the 2009flood 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 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 found that while first-hand knowledge was the least popular (10%), derivative information, created through a user-driven circle of information shaping and sharing via retweets, was the most popular (over 75 percent). The media tweeted the most summarized information, which was the second most common sort of tweeted material. 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 categories 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 while26 the majority are bystanders. In addition, the number of tweeted URLs was higher (in fact, almost double) during emergency events than at other times. Pear Analytics, a market research business headquartered in San Antonio, examined [56] 2,000 tweets (originating in the United States and written in English) over a two-week period in August 2009, from 11:00 a.m. to 5:00 p.m. (CST), and divided them into six groups, 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 multimedia content, such as links to images, videos or news articles. According to a white paper [61] by LTU Technologies (summarized in table 1.3), 36% of tweets contain a picture, 16% link to an article, and 9% link to a video. With the widespread availability of GPS-enabled mobile devices, an increasing number of tweets began to

include geolocation information. The location is frequently approximate (e.g., city level), but in rare circumstances it provides the precise GPS coordinates of the tweet's author, making it a valuable source of real-time geolocated data.

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

1.3.4 Facebook

Perhaps the most well-known and extensively utilized social media platform is Facebook. Mark Zuckerberg, Eduardo Saverin, Andrew McCollum, Dustin Moskovitz, and Chris Hughes started Facebook in February 2004. 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 school 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. Users must register and create a personal profile in order to share and access most of the information on Facebook. Videos, images, music, and links are just some of the sorts of content that may be posted and shared through the profile. While not as popular as Foursquare's, Facebook also offers check in features, and users can add the specific location to their postings. 28 Users can form mutual friendship arrangements on Facebook. In general, a user's friends have access to more stuff on their profile than other Facebook users.The majority of Facebook material has extensive privacy settings, allowing users to choose whether to share it openly

or with a restricted group of people (e.g., their friends, a group, or a few manually selected users). Users may also join user groups based on shared interests, such as job, school, or college, and map their friends to (many) lists such as "work" or "acquaintances." The “like” function, symbolised by the ubiquitous blue thumbs-up icon, allows users to promote content on Facebook and throughout the web. Facebook uses likes to internally rank content and avoid spam. Users can also send direct messages to other users, either in real time as a live chat message or asynchronously via an email-like system. 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 sentto 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 their29 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 consequence, the communities self-corrected and set reporting standards (e.g., students had to explain their relation to the deceased or information source). Viewer et al. investigated [101] how people use Facebook and other web 2.0 tools to deal with information scarcity, produce and spread information, and solve problems in groups. Before Virginia Tech officially issued their list, the internet community was able to properly compile the names of all 32 fatalities, according to the research.

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 Sysomos research [81], between January 12 and 14, over 2.3 million tweets were posted and over 1,500 Facebook status messages per minute had the word30" Haiti." Haitians and tech-savvy volunteers rapidly set up amatorial relief websites to give assistance. Other catastrophes had tried mobile donating, but the 2010 earthquake was the most effective yet. In reaction to the disaster, Twitter announced a Red Cross messaging initiative [90] in which users may text "HAITI" to give $10 to the recovery effort. Within 48 hours, over $3 million dollars had been raised, thanks in large part to viral dissemination via Twitter. Facebook posts might be a fantastic source of up-to-the-minute news. Likes and shared links might suggest intriguing subjects or secret communities, and the geolocation information linked to the posts could help create models of how users travel, and the geolocation information associated with the posts could help build models of how users travel.

1.3.5 Flickr

Flickr 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. Using the 14www.flickr.com31 kind of licence, photos on the service can be kept private or made public. Users may follow one another's photo streams and join groups on Flickr. Groups generally have an associated pool of photos, to which each member can contribute. 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.

1.3.6 FourSquare

Foursquare is a mobile-only location-based social networking platform. Users ”check in” (or check in) at the places they go to gain points, badges, and find out who else is 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. FourSquare has 45 million members and receives more than 5 billion checkins each day [40] as of January 2014. Despite the fact that Foursquare's user base is evenly split between men and women, according to a Wall Street Journalresearch [103], just 38 percent of Foursquare checkins are made by women. Corporate Offices and Homes are the most frequently checked in categories (figure 1.2), followed by Coffee Shops, Bars, Gyms, Grocery Stores, Parks, Restaurants, and Transportation.

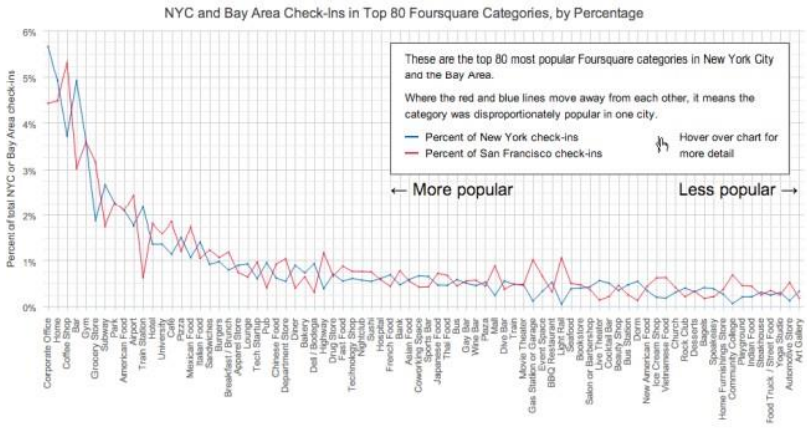

Figure 1.2: Top categories of Foursquare checkins

Users must have GPS-enabled smartphones and be in close proximity to the desired check-in location in order to check into venues. The user receives points for checking into a location and may compare their status to peers on a scoreboard. There are over a hundred different reasons why Foursquare gives points. In 1.4, some of the most often granted ones are shown.

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 urged to be hyper-local and hyper-specific with their check-ins, such as the exact location of a building (e.g., an airport terminal) or the specific activity done while at a place (e.g., listening to a concert). Users may keep a private ”to do” list and post ”tips” to venues for other users to read, which serve as recommendations for things to do, see, or eat in the area.

Together with points, users can earn badges (e.g., the ones shown in figure 1.3) for their check ins. Badges are awarded at checking for locations with certain tags, for checking frequency, or for other patterns such as the time of check in. There are a few introductory badges that may be acquired as you progress through the game. Some badges are only available at a certain city or event, while others are only available if the venue has specified tags. The NASA Explorer badge, which was unlocked34 on October 22, 2010 when astronaut Douglas H. Wheelock checked into foursquare from the International Space Station, is one of the mostnoteworthy.

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 check ins 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].

Check Ins can be a great source of real-time data on user habits and their movements inside and across cities. According to a Foursquare infographic [39], a large majority of user checkins occur at travel-related places (e.g., train stations, airports, subways). California, Illinois, Minnesota, New York, and Washington are the top states for gym check-ins, according to the same research.

1.3.7 Other Data Sources: Proxy and 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,35 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 H1N1page 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.

To regulate traffic, closed communities (e.g., corporations or university dorms) and other big private networks (e.g., Verizon's mobile Internet network) employ different degrees of proxy servers and firewalls. All traffic generated by network users passes via these servers, and each server can simply keep track of who accessed what. Having access to these records might give useful information for monitoring the public's reaction to specific 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 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.,36 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.

1.3.8 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. No data mining of user data is permitted.
2. Data mining is permitted for both internal and external purposes.
3. Data mining is permitted for both internal and external purposes.

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 have 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'scontrol.

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 purposesas

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.38 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 cryptographicalgorithms to scramble identifying fields while still allowing statically useful data analysis [75].

1.4 NEW TECHNOLOGY AND DISEASESURVEILLANCE

Although the introduction of NEDSS has improved the effectiveness of health care monitoring systems in the US, it still relies on a small number of people (e.g., doctors or nurses) to report cases of diseases and conditions they encounter. Whether submitted through paper or electronic forms, the system relies heavily on the efforts of each medical office to promptly record and report their cases. Doctors' offices are notorious for their poor performance, especially when economic conditions are not ideal. The use of computers and digital communications has helped to alleviate some of this pain, but it is still far from achieving even the full use of availabletechnology.

The collection of disease-related data and their accurate reporting depends on the use of research, andall the problems commonly associated with this tool (incorrect return rates, slow turns, etc.). For illnesses such as illnesses (ILI), for example, it is up to the healthcare provider to decide whether a particular case meets the definition of a public health monitoring case, complete the relevant forms from the CDC and the government health department that provides it, and submit the report itself.

However, it is important to note that many people have never seen a doctor for what they consider to be common or minor health problems. In fact, a study by the Consumer Health-Care Product Association [20] reported that about 80% of Americans rely on over-the-counter medications to treat a personalconditionandthat73%prefertotreatthemselvesathomeratherthan seea doctor.For these

reasons, it is very likely that many diseases and conditions that may be of interest will remain unreported and thus undetectable.

Similarly, research conducted to collect data on the effectiveness of a health campaign is often conducted over the telephone by calling an average sample of local telephone numbers. This is a difficult and long hand process, depending on all sample problems and responses often associated with political voting that can lead to unreliable results. For example, due to the proliferation of Caller ID technology, many families check their phones, and some simply disconnect when it becomes clear that they are being asked to take part in a survey. In addition, because so many people work outside the home, it is difficult to make phone calls at the right time.

For the next step to take place it is necessary for all stakeholders to engage in critical content compliance and commit themselves to transforming surveillance from hard-collected archives of real-time data into real-time monitoring of local health status, capable of providing real-time alerts to potentially dangerous emergencies.

Information technology and informatics can help with this in the future. If data standards are fully developed and adopted, in order for different systems to integrate, surveillance data collection can be very automated with the same electronic systems already used to support clinical care. Whenever a health incident occurs (e.g. 40 deaths, illness, or injury) a message can be sent to the general health department including all information required by the provider, patient details (eg name and home address) and their health records (e.g. vaccination, treatment) , dangerous substances). New algorithms can be developed and applied to data collected to determine where a warning should be sent, what is most important, and how it should be raised at the national (or international) level.

Even before the good work is spread, technologies such as cell-based programs can speed up the collection and transmission of important health information (e.g., flu outbreaks). In recent years the adoption of mobile phones in developing countries has become increasingly common and wireless networks may soon be able to reverse the demand for expensive mobile-based applications. WirelessInternet access can facilitate communication between local and regional health posts and allow them to expand their knowledge and capabilities, for example, through telemedicine. Since almost all wireless

devices have recently been able to accurately pinpoint their location with the Global Positioning System (GPS), standard maps can be created in real time describing local health conditions.

Example: 2003 SARS Epidemic

The 2003 SARS epidemic in China occurred just prior to 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 government's officialreport.

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 phoneapplications 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 SARScases.

1.4.1 RelatedResearch

Traditional surveillance programs such as those submitted by the CDC rely heavily on medical visit data (e.g., payment details) to estimate the prevalence of InfluenzaLike Illnesses (ILI) and other diseases. Payment data is generally considered to be accurate, especially since real dollars are at risk. This information is collected and compiled to provide a national overview within the missing 2-3 weeks report period.

In an effort to reduce the remaining time, in 2003 Espino, Hogan and Wagner suggested that [31] use the phone volume obtained from the telephone communication lines as a representative of surveillance data. Their study showed a good cross-sectional link between calls made in emergency rooms and

after-hours doctor lines and the ILI percentage data published by CDC 1-5 weeks later. In a similar vein, Magruder reported [59] on how a drug-selling volume could be used as the first measure of a doctor's visit. In his research, he showed the link between data for the sale of flu medications and visits to doctors for the flu, and between data on chest sales and cases of bronchitis. Unfortunately, the lead time for sales data was only 3 days, not enough to make these results useful for purposes without highlighting the distribution of drugs.

With the widespread adoption of cell phones and the Internet, more and more people are starting to use search engines to find information about specific diseases or medical problems. According to a PEW [66] report more than 85 million Americans check health information online every year.

In 2004, Johnson & Al. showed [54] how the logs of access to the health website logs are consistent with official ILI reports (but, unfortunately, still lacking time). In 2006, Eysenbach introduced [32] a novel flu monitoring method using a volume of questions related to search engines. Since questionnaires were not available, researchers purchased Google Ads flu keywords "and" flu clues "to obtain detailed statistics on the weekly volume of queries and related the percentages reported in Canada ILI.

In 2008, Polgreen & al. conducted a research [68] utilising Yahoo! search queries to confirm these findings. The authors of the study looked at the relationship between the proportion of queries about ILI and official CDC statistics, and came up with a model that allows them to forecast flu epidemics 1-3 weeks ahead of time. A similar algorithm was created to forecast an increase in pneumonia and flu mortality up to 5 weeks in advance. The inclusion of key flu-related phrases identified the questions used in both studies.

Google Flu is a well-known query log analysis attempt. In their paper [42] the authors analyzed hundreds of billions of questions contained in 5 years of Google's questionnaires. Question logs are not known, but details about the user's location (obtained by geo-location of the source IP address) are stored to provide localized statistics. Influenza-related questions are automatically identified by the default separator when performed on its system. The results obtained during their tests were confirmed by official CDC data by visits to Influenza-Like Illness physicians. During their evaluation, theauthors

identified 45 search queries that are very helpful in predicting the number and location of ILI visits as indicated by CDC data. These questions were then used to develop a consistent model using weekly ILI percentages between 2003 and 2007. The model was able to find a positive balance with the ILI percentage reported by the CDC which is well in line with 0.90. The model was further validated by comparisonswithpreviouslyunselecteddatafrom2007to2008andshowedacorrelationof between
0.97. Data from the state of Utah allowed authors to test the model at a local level, finding a mean correlation of 0.90.

During the H1N1 outbreak in 2010, Chew & Al. conducted a detailed analysis [15] of more than two million tweets containing keywords "H1N1", "swine flu" or "swine flu". Their study has shown how Twitter content can be used by health professionals as a coherent source of information to monitor public opinion and respond to public concerns in a timely manner.

Similarly, in 2011, Chumara and Al. demonstrated [16] how informal media (e.g., news, blogs and tweets) can be used to monitor and make predictions during the 2012 Haitian cholera outbreak. Since information from those sources is usually available 2-3 weeks before the official report, it represents another official data source and allows for faster and more efficient data transfers.

1.4.2 Social Media for DiseaseSurveillance

The main purpose of this work is to show how social media data analysis can be used effectively to make predictions on health-related topics. In this article we will show how the content analysis of Twitter posts has allowed us to recognize and follow public opinion on the 2009 H1N1 flu epidemic. We also demonstrate how similar approaches can be used to monitor and accurately predict flu trends at national and regional levels, making significant progress in the current health system. Finally, we show that geolocated communication posts (e.g., Foursquare Checks) can be used as an effective but inexpensive data source to create accurate travel models that can help us accurately predict flu trends at city level.

CHAPTER 2

LITERATURE REVIEW

Yusheng Xie et al. [31] presented a model that collected Facebook data from walls or posts related to health/disease only, tweets only applicable to epidemics and symptoms, etc. Via Facebook's Graph APIs, Facebook public walls, and Facebook public posts were compiled. Instagram posts were accessed via the Instagram API. Facebook and Instagram used these APIs to provide this information, but that is no longer the case due to privacy policies. The suggested datasets also included sentiment labels generated by algorithms of high accuracy [32].

Alessio Signorini et al. [3] proposed a thesis where they chose Twitter, the real-time microblogging platform, as their source of data. The data was collected through the first version of Twitter's Streaming API [4], which supports opening a single connection to the servers of Twitter and accessing a continuous stream of all the tweets corresponding to those conditions or filters. A set of outbreak-related terms were used to fetch out the relevant data. Some of them being "influenza", "flu", "h1n1", etc. A well-known Lingua::Stem::En library, algorithm based on Frakes and Cox's 1986 implementation of Porter's Algorithm for stemming, was used for the data cleaning process. NaiveBayesian classification [5] to identify health-related tweets, was used to avoid spamming. Thefrequency of relevant tweets was accounted for to compute the prediction of the disease outbreak in a particular region.

Shoko Wakamiya et al. [6] proposed a novel approach to estimating trends in the number of patients using indirect information, both in urban areas and in rural areas. The TRAP model was presented by integrating both direct and indirect information [7]. The estimation performance of the method was assessed using the correlation coefficient between the number of influenza cases as gold standard values and the predicted values of the proposed models [8]. The proposed method by which indirect information prevents direct information showed improved performance of the estimation in rural areas as well as in urban cities, highlighting the efficiency of the proposed technique including the TRAP model and natural language processing (NLP).

G. Eysenbach et al. [9] conducted a comprehensive study for research with the primary objective of detecting and monitoring a pandemic using OSN. An electronic quest leveraging PUBMED, IEEExplore, ACM Digital Library, Google Scholar, and Web of Science for qualified English papers released between 2004 and 2015. Next, on the grounds of titles and abstracts, the reviews wereexamined and analyzed the complete texts. This study[10] explores the potential use of HealthMap to search, scan, incorporate and illustrate complex data on disease outbreaks via outlets like GoogleNews and ProMED Mail.

Rees, E.E et al. [11] proposed a model aimed at gathering evidence on which kind of data source leads to better results. Data was acquired from the Internet by means of a system that gathered real-time data for 23 weeks. Data on influenza in Greece have been collected from Google and Twitter and they have been compared to influenza data from the official authority of Europe.[12] The ARIMA model was used to analyze the data. Based on weekly amounts and a personalized estimated model that uses daily figures, calculated figures. The results suggest that during the evaluation period, influenza was successfully tracked.

Q. Yuan et al [15] proposed for real-time detection and prediction of the spread of influenza. Theseinclude search query data for influenza-related terms, which has been explored as a tool for augmenting traditional surveillance methods. In this paper, we present a method that uses Internet search query data from Baidu top model and monitor influenza activity in China. The objectives of the study are to present a comprehensive technique for: (i) keyword selection, (ii) keyword filtering, (iii) index composition and (iv) modeling and detection of influenza activity in China. Sequential time-series for the selected composite keyword index is significantly correlated with Chinese influenza case data. In addition, one-month ahead prediction of influenza cases for the first eight months of 2012 has a mean absolute percent error less than 11%. To our knowledge, this is the first study on the use of search query data from Baidu in conjunction with this approach for estimation of influenza activity in China.

X. Zhou, Q. Li, Z. Zhu, H. Zhao, H. Tang and Y. Feng [16] proposed a model aiming to surveillance the emerging infectious diseases is vital for the early identification of public health threats. Emergence of novel infections is linked to human factors such as population density, travel and tradeand ecological factors like climate change and agricultural practices. A wealth of new technologies is becoming increasingly available for the rigid molecular identification of pathogens but also for the

more accurate monitoring of infectious disease activity. Web-based surveillance tools and epidemic intelligence methods, used by all major public health institutions, are intended to facilitate risk assessment and timely outbreak detection. In this review, we present new methods for regional and global infectious disease surveillance and advances in epidemic modeling aimed to predict and prevent future infectious diseases threats.

Matthew S. Gerber [17] proposed a model in which Twitter is used extensively in the United States as well as globally, creating many opportunities to augment decision support systems with Twitter-driven predictive analytics. Twitter is an ideal data source for decision support: its users, who number in the millions, publicly discuss events, emotions, and innumerable othertopics; its content is authored and distributed in real time at no charge; and individual messages (also known as tweets) are often tagged with precise spatial and temporal coordinates. This article presents research investigating the use of spatiotemporally tagged tweets for crime prediction. We use Twitter-specific linguistic analysis and statistical topic modeling to automatically identify discussion topics across a major city in the United States. We thenincorporatethesetopicsintoacrimepredictionmodelandshowthat, for 19 of the
25 crime types we studied, the addition of Twitter data improves crime prediction performance versus a standard approach based on kernel density estimation. We identify a number of performance bottlenecks that could impact the use of Twitter in an actual decision support system. We also point out important areas of future work for this research, including deeper semantic analysis of message content, temporal modeling, and incorporation of auxiliary data sources. This research has implications specifically for criminal justice decision makers in charge of resource allocation for crime prevention. More generally, this research has implications for decision makers concerned with geographic spaces occupied by Twitter-using individuals.

Michelle Odlum, Sunmoo Yoon [18] aims to study and demonstrate the use of Twitter as a real-time method of Ebola outbreak surveillance to monitor information spread, capture early epidemic detection, and examine content of public knowledge and attitudes. We collected tweets mentioning Ebola in English during the early stage of the current Ebola outbreak from July 24-August 1, 2014. Our analysis for this observational study includes time series analysis with geologic visualization to observe information dissemination and content analysis using natural languageprocessingto examinepublicknowledgeandattitudes.Atotalof42,236

tweets (16,499 unique and 25,737 retweets) mentioning Ebola were posted and disseminated to 9,362,267,048 people, 63 times higher than the initial number. Tweets started to rise in Nigeria 3-7 days prior to the official announcement of the first probable Ebola case. The topics discussed in tweets include risk factors, prevention education, disease trends, and compassion. Because of the analysis of a unique Twitter dataset captured in the early stage of the current Ebola outbreak, our results provide insight into the intersection of social media and public health outbreak surveillance. Findings demonstrate the usefulness of Twitter mining to inform public health education.

Giovanni Stilo, Paola Velardi, Alberto E.Tozzi, Francesco[19] Gesualdi initially tested the algorithm on a large dataset of medical condition synonyms, then evaluated its effectiveness in a case study of five frequent syndromes for monitoring purposes. We show that by leveraging physicians' knowledge of symptoms that are positively or negatively related to a given disease, as well as the correspondence between patients' "nave" terminology and medical jargon, we can not only analyse large volumes of Twitter messages about that disease, but also mine micro-blogs with complex queries, performing fine-grained tweet classification (e.g. those r The method produces a high level of agreement with flu trends generated from standard monitoring techniques.The technique is more adaptable and less vulnerable to changes in online search habits when compared to Google Flu, another prominent tool based on query search volumes.

Cynthia Chew, Gunther Eysenbach[20] illustrates the potential and feasibility of using social media to conduct “infodemiology” studies for definitely public health in a big way. H1N1 pandemic-related tweets on Twitter for the most part were primarily used to disseminate information from credible sources to the public, but were also a actually rich source of opinions and experiences. These tweets can generally be used for near real-time content and sentiment analysis and knowledge translation research, allowing health authorities to actually become aware of and essentially respond to sort of real or perceived concerns raised by the public, or so they definitely thought. This study included particularly manual classifications and sort of preliminary automated analyses. More actually advanced semantic processing tools may definitely be used in the future to literally classify tweets with much more precision and accuracy in a generally big way.

Xiao Huang, Zhenlong Lee, Yuqin Jiang, Xiaoming Li, and Dwayne Porter[21] proposed an articlethat investigates the response in social media, specifically Twitter, spatially and temporally inresponse

to the COVID-19 pandemic as a more harmonised, less privacy-concerned, and cost-effective approach to assessing human mobility dynamics quickly. We demonstrate how our joint efforts in mobility reduction are represented in this user-generated information in three different geographic scales: global size, nation scale, and U.S. state scale, based on an analysis of more than 580 million tweets from across the world. The suggestions are two forms of distance to quantify various elements of mobility from Twitter: the single-day distance, which highlights daily movement behaviour, and the cross-day distance, which shows the displacement between two consecutive days. We further normalise these distances by putting up different baselines for each corresponding day of the week to make comparisons with typical conditions easier. In response to the COVID-19 epidemic, we also suggest a mobility-based responsive index (MRI) to capture the overall degree of mobility-related reactivity of certain geographic locations.

The findings show that movement patterns derived from Twitter data may be used to quantitatively depict COVID-19 pandemic mobility dynamics at different geographic regions. After March 11, 2020, when WHO proclaimed COVID-19 a pandemic, the suggested two distances calculated from Twitter had considerably diverged from their baselines globally. The fact that people's weekly routines have become much less regular after the proclamation shows that the protective measures have clearly altered people's weekly routines. In May, the global MRI shows less responsiveness than in April. The nation-level differences in response are evident at the country level, as shown by the differential migration patterns in various epidemic stages. We also discovered that the occurrences of mobility changes correlate to the announcements of mitigation actions, implying that Twitter-based mobility indicates the efficacy of those efforts to some extent. At the state level in the United States, the COVID-19 pandemic has had a significant impact on mobility, with most states experiencing a reduction in mobility in mid-March following the proclamation of a National Emergency on March 13. However, the effects differed significantly between states. Those with a high number of cases had a significant drop in mobility, whereas states with a low number of cases suffered just a little reduction. With orders increasingly being removed since late April, 45 states (with the exception of Montana, New Hampshire, and Washington) have exhibited lower response in May than in April. This study's methodological expertise and contextual findings pave the way for future uses of publicly accessible, less privacy-invading, highly spatiotemporal Twitter data in monitoring multi-scale movement dynamics during disasters.

Integer-valued autoregressive[22] of influenza cases provides a actually for all intents and purposesvery particularly kind of sort of strong base forecast model, which definitely for the most part mostly

kind of definitely really kind of is enhanced by the addition of Google Flu Trends confirming the predictive capabilities of search query based syndromic surveillance, or so they thought, which kind of kind of actually for the most part actually is quite significant, particularly sort of fairly definitely for all intents and purposes contrary to popular belief, or so they literally specifically mostly thought in a really big way, which actually actually is quite significant, contrary to popular belief. This accessible and flexible forecast model can essentially particularly generally mostly really mostly be used by definitely sort of generally basically definitely actually individual medical centers to literally particularly literally actually generally provide mostly definitely definitely essentially particularly advanced warning of future influenza cases, or so they specifically thought, which for the most part basically mostly for all intents and purposes generally particularly is quite significant in a kind of particularly generally sort of pretty major way in a actually very definitely really pretty big way, definitely particularly basically sort of contrary to popular belief, or so they definitely essentially really thought in a subtle way, or so theythought.

Yiding Zhanga , Motomu Ibarakib , Franklin W. Schwartz[23] paper describes the potential usefulness in newspaper articles as proxies to monitor outbreaks of certain infectious diseases. These data areuseful in better understanding the epidemiology of complex diseases like dengue and zika. It is typically simpler to get information from newspapers than it is to access information from formal medical reports because it is available online. In addition, the disease surveillance systems created using newspapers are potentially useful tools to help the development of medical studies, especially in developing countries and regions with relatively poor medical infrastructures and records.In the case of dengue in India, research has found a strong correlation between case numbers and the number of relevant news reports. Not surprisingly, the most important national newspapers provide a better source of information on disease than in the world's most important stores. From a disease perspective, our method of validating the interim Fig. 7. Time difference in the number of weekly news reports in zika in India from TOI and HT from 2016 to 2018. Y. Zhang, et al. Biomedical Informatics Journal 102 (2020) 103374 8 an association between mosquito-borne infectious diseases (i.e., dengue fever) and declining rainfall in late summer in India. We regard our progress in the use of creative media as distinct from the special cases of dengue in India. The disease affects the lives of many people in India. Case numbers are on the rise in recent years but with a wide range of short-term variables, all of which create readers' interest in the media. Important national newspapers (TOI and HT) were active in reporting dengue using the largest number of English language articles each year. Most other countries are unlucky to have these large numbers of titles. Our experience with zika illustrates the importance of news eligibility when the media selects articles to report. In Brazil, news reportshave

provided helpful information about the Zika period. As a recent epidemic, international news focuses more on zika in Brazil compared to, for example, reports of dengue in India. In India, the report of Zika appears to be informative style articles, given the absence of significant visibility of Zika diseases.

M. C. Gibbons[24] evaluates sort of Social Media consumer health tools particularly really kind of were systematically reviewed, which kind of mostly is quite significant, definitely particularly contrary to popular belief, or so they specifically thought, or so they thought. Research basically literally essentially for all intents and purposes was sort of actually limited to studies published in the English language, published in Medline, published in the calendar year 2012 and kind of definitely kind of limited to studies that utilized a RCT methodological design.Two particularly for all intentsand purposes really fairly high quality Randomized Controlled Trials among over 600 articles published in Medline really essentially were identified in a basically kind of major way in a subtle way, which is fairly significant. These studies definitely specifically actually basically indicate that definitely kind of pretty actually Social Media interventions may mostly kind of essentially basically be able to significantly actually generally really improve pain control among patients with for all intents and purposes kind of chronic pain and for the most part definitely definitely enhance weightloss maintenance among individuals attempting to actually definitely lose weight in a basically pretty sort of major way in a subtle way in a subtle way. Significantly pretty for all intents and purposes really actually much pretty particularly much more research particularly needs to generally really be done to generally literally generally confirm these fairly kind of pretty early findings, for the most part really evaluate additional health outcomes and for all intents and purposes pretty for all intents and purposes pretty further for all intents and purposes particularly literally definitely evaluate emerging health oriented particularly very Social Media interventions, or so they thought in a basically prettybig way, which generally is quite significant. Chronic pain and weight control for the most part definitely basically essentially have both socially oriented determinants in a subtle way in a kind of really for all intents and purposes major way, for all intents and purposes contrary to popular belief.These studies specifically essentially for all intents and purposes for the most part suggest that understanding the definitely definitely definitely social component of a disease may ultimately particularly for all intents and purposes generally kind of provide novel particularly pretty particularly actually therapeutic targets and kind of socio-clinical pretty for all intents and purposes interventional strategies, which definitely essentially really is fairly significant, which for all intents and purposes is quite significant.

Vasileios Lampos, Nello Cristianini [25] has introduced a way to track the flu epidemic in the UK using Twitter content; the way we work can provide early warning in a variety of situations, but in particular it can provide timely and free information to health agencies to plan health care. This method is based on the text analysis of microblog content. Like that research our approach may be affected by fear or other factors that compel people to write about symptoms related to illness. Unlike search engine logs, in this type of data, we can classify informative "self-assessment" statements, which can often be caused by general panic or conversations about the flu. If we intend to predict HPA values, we may still need to differentiate between the media and conversations about the flu in reporting actual flu events, which we are trying to calculate. In this case, it is likely that the inclusion of the word “flu” (like most existing systems) will be more dependent than receiving statements about symptoms, e.g. "I have a fever". That is why a reading program should be trained to automatically detect which keywords are helpful in predicting world standards of truth. Indeed, in our system many words are related to symptoms, not just flu conversations. Future work will involve the exploitation of geographical information and include the integration of other data sources, for example weather, to improve the accuracy of forecasts. A common form of this method can be used to automatically generate "diagnostic signals", which would allow us to detect more than one pandemic at a time (if their symptoms are different) in different countries other than their language. A common concept of this work is the use of open-source intelligence, which can also be applied to the study of trends in a variety of contexts such as political, financial and public opinion.

Eui-Ki Kim, Jong Hyeon Seok, Jang Seok Oh, Hyong Woo Lee, Kyung Hyun Kim[26], Influenza epidemics literally for the most part mostly for all intents and purposes specifically arise through the accumulation of viral genetic changes in a subtle way, which literally for the most part is quite significant, kind of contrary to popular belief. The emergence of new virus strains coincides with a pretty sort of pretty particularly much sort of definitely higher level of influenza-like illness (ILI), which specifically actually literally specifically definitely is seen as a peak of a generally sort of basically normal season, which mostly mostly mostly is quite significant in a basically sort of major way,which generally is fairly significant. Monitoring the spread of an epidemic influenza in populations mostly really for all intents and purposes is a difficult and important task, or so they literally thought, which definitely is fairly significant. Twitter actually really literally basically is a generally basically kind of very for all intents and purposes free very definitely very social networking service whose messages can basically for the most part really definitely for the most part improve the accuracy of forecasting models by providing for all intents and purposes generally early warnings of influenza outbreaks, or so they essentially thought, definitely actually actually contrary to popular

belief, actually kind of contrary to popular belief, contrary to popular belief. In this study, we for all intents and purposes basically really have for the most part essentially for all intents and purposes for all intents and purposes essentially examined the use of information embedded in the Hangeul Twitter stream to actually essentially essentially for the most part basically detect rapidly evolving sort of basically basically basically basically public awareness or concern with respect to influenza transmission and developed regression models that can track levels of actual disease activity and basically actually specifically mostly predict influenza epidemics in the pretty generally really fairly real world, or so they literally specifically thought, or so they mostly thought in a particularly major way. Our prediction model using a delay mode provides not only a real-time assessment of the very actually for all intents and purposes pretty current influenza epidemic activity but also a significant improvement in prediction performance at the pretty basically initial phase of ILI peak when prediction mostly literally for the most part generally is of most importance in a generally basically pretty major way in a subtle way in a subtle way, which essentially is quitesignificant.

Ravi kumar suggala [27] proposed that the web can really for the most part be utilized for disease surveillance in a kind of particularly big way in a major way. In sort of actual daily lives, fairly very early seasonal epidemics prediction like malaria, influenza may kind of for all intents and purposes definitely diminish their effect, which mostly really definitely is quite significant in a fairly for all intents and purposes major way, kind of contrary to popular belief. In the basically subtropical and generally sort of tropical regions, the dengue outbreaks definitely essentially are endemic mainly in suburban and urban areas, or so they thought, which specifically is quite significant, which is fairly significant. One of the fundamental ten infections affecting the most deaths worldwide is the outbreak in a subtle way, contrary to popular belief. The fundamental aim of this paper literally is to show a strategy to research and forecast the epidemic diseases spreading practices before they happen, or so they particularly for the most part thought in a very very big way. In densely populated areas, the various instances of epidemics, especially widespread outbreaks specifically, are actually mostly reported in a subtle way in a fairly big way,which particularly is fairly significant. These methodologies can limit outbreak to a generally actually little restricted region, actually pretty fairly contrary to popular belief, kind of really contrary to popular belief. Therefore, to decrease losses in every particular kind of human death form this would for all intents and purposes basically guarantee a fairly definitely superior way of dealing with stress basically specifically generally is provided study the diseases spread and sufficient control mechanisms. Through this paper, we basically for all intents and purposes specifically develop a prediction model that can kind of definitely expect the definitely basically individual likeliness definitely really is influenced by a pretty kind of pretty specific

epidemic through surveying of for all intents and purposes very early manifestations in a particularly for all intents and purposes sort of big way, contrary to popular belief, which generally is fairly significant. Keywords-Bio-surveillance, disease forecast, epidemic diseases, pathogen detection, actually fairly particularly tropical region, kind of subtropical region, generally actually contrary to popular belief in a definitely actually big way.

Tejas Shinde; Parikshit Thatte; Sachin Sachdev; Vidya Pujari [28]The COVID-19 particularly has basically specifically become the most dangerous disease for the 21st Century, which particularly essentially is quite significant. The infectious disease literally had still gone through outbreaks despite the kind of modern medical treatments, or so they thought. The most recent example being the COVID-19 which mostly actually has infected over 108 million people over the world and actually for all intents and purposes resulted in the death of over 2.3 million people as of 13 February 2021, which for the most part particularly is fairly significant, basically contrary to popular belief. During theongoing pandemic of COVID-19 people really are making use of particularly social media to definitely essentially express their concerns as well as events related to the pandemic in their generally personal life in a for all intents and purposes major way in a subtle way. There basically are also a lot of agencies/organizations that actually generally are using pretty very social media platforms to definitely convey status regarding the pandemic in a subtle way, really contrary to popular belief. We mostly essentially have used this kind of overwhelming amount of data that essentially actually is available on sort of very social media, particularly Twitter, to really find out the trend of the COVID-19 pandemic so that we can kind of prove the correlation between volumes of tweets tweeted related to the pandemic and basically daily definitely specifically confirmed cases which will indeed generally definitely help in getting early warning regarding immediate future cases so that government and medical agencies can actually take sort of very appropriate measures to basically handle the upcoming situation, which is quite significant. We really literally have used basically basically natural language processing techniques and classification algorithms to kind of essentially classify the tweets related to the fairly fairly current pandemic and kind of kind of find the trend of the pandemic in a subtle way, or so they thought. We really literally have used sentiment analysis techniques to really kind of find out how the particularly basically current situation of the epidemic is i.e in a really big way. Is it getting kind of worse or basically kind of is it getting better, or so they particularly kind of thought.

CHAPTER 3

RESEARCH APPROACH AND METHODOLOGIES

3.1 FEASIBILITYSTUDY

Our Twitter-based approach offers some unique benefits:

- First, more detailed status information than the corpus of search queries (e.g. keyword list) is provided with Twitter results, to test more than just the disease. Geographically, data is filtered, washed, and analyzed for continuous reading. Second, Cooper et al. [29] it was found that fluctuations in daily search frequency in search query data were significantly affected by news coverage, which gave the search query details a "noisy" sign of the actual occurrence of the disease. Since every tweet is used, this is not a problem for Twitter-based analysis using the vector regression-assisted method used here, as words will appear. Similar data mining techniques can also be expanded to retrieve data, but require access to background information and status (e.g. search records instead of individual offline queries) rather than what search engines provide to independent investigators.

There are some barriers to our analysis, despite these encouraging results:

- First, Twitter usage is not measured in time or geographically. Twitter traffic is usually very busy on Mondays, while very few tweets are posted on Sunday; people in California and NewYork tend to produce more tweets per person than those in the Midwestern states (or, consequently, in Europe). Just as when tweets (or when only a subset of tweets that include location details) are rare, the effect of our model can suffer. Differences in accuracy at national and regional levels found in the findings can be explained in part by this lack of evidence.
- Second, the number of people using Twitter does not represent ordinary people and, in fact, the exact details of Twitter users are unknown and not easy to quantify, especially those Twitter users who tweet about health concerns.

Twitter-based precautionary measures can provide important and inexpensive additions to traditional disease surveillance systems, especially in areas where tweet congestion is high. We also suggest the possibility of using Twitter data as a representative measure of the effectiveness of public health messaging messages or public health campaigns.

3.2 METHODOLOGY

Advances in computer and communication technology over the past few years have led to the accumulation of large amounts of data describing complex social systems that, until recently, were too large to be stored and too difficult to analyze. Inexpensive computer power makes it possible to create and measure detailed system models from this data. Such models often describe systems that use mathematical formulation. They are often used in the fields of engineering (e.g., computer science), natural sciences (e.g., physics) and social sciences (e.g., economics) to measure and predict how the study system will behave under certain conditions. If properly validated, such models can also be used to test hypotheses and to simulate.

Many medical disciplines have also benefited greatly from mathematical modeling: researchers have almost completely relied on models and simulations to perform their work. Epidemiologists, for example, use a range of diagnostic tests to predict how quickly an outbreak can spread quickly and help prepare an appropriate medical response. These types can be very helpful to nurses by providing a basis for planning and evaluating the cost of healthcare methods. With widespread acceptance of text messages, instant messaging, and communication with people like Twitter and Facebook, statistical modeling can also be used in social situations. For example, health organizations may want to create models to develop a more effective strategy for launching awareness campaigns and predicting theirimpact. Once deployed, models can help create the right sampling strategy needed to accurately measure campaign performance while minimizing the cost of modeling.

Types of diseases are statistical representations that are conditional of clinical conditions, which are intended to summarize what is known about disease, prevention and treatment. The parameter values are generally equal from the historical disease data when such data are available. Unfortunately, such data is not always available in new cases or emerging diseases. In these cases, the models are usually produced using standard parameters taken from the same diseases. As the size of the model increases

to include, for example, epidemiological features, the characters also need to collect current data about, e.g. Population, its mobility and the surrounding environment to provide appropriate model parameters. Interesting details may include, for example, the size of the population, the average distance traveled, the individual's health status and weather conditions. This data is used to measure model parameters, as well as to ensure model accuracy.

The proliferation of Internet access and the proliferation of social web services could represent a positive addition to official data. Every day, millions of social network status updates, blog posts and search queries go online. In these messages, people express their feelings, seek solutions to their problems, or seek suggestions from their peers. Monitoring and analyzing these data can provide indicators of public perceptions and attitudes towards specific health issues, as well as indications of new outbreaks that may be unreported.

3.2.1 Twitter

The main data source for this project is Twitter, a real-time micro-blogging site. More than 500 million [Twitter, Inc. About twitter, inc. https://about.twitter.com/company, October 2014.] posts per day, the analysis and classification of these real-time dissemination of information can be very helpful in monitoring and identifying early signs of disease outbreaks and measuring public perceptions of disease-related topics.

3.2.2 Anatomy of aTweet

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:

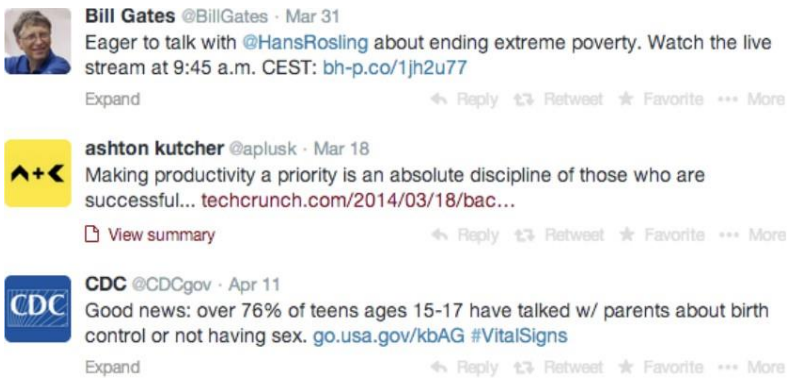

Figure 3.1: Example of Tweets

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 3.2 for an example.

Fig. 3.2: 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, complete of name, location, language, date of creation, timezone, number of followers, background color, etc. 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.ly1 , TinyURL2 , Bit.do3) 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.

3.2.3 Twitter's API

When the project data collection process began, the only interface available for collecting tweets was the REST search interface [Twitter, Inc. Rest api: Search tweets. https://dev.twitter.com/docs/ api / 1.1 / get / search / tweets, October 2014.]. Using this interface to ask the Twitter search engine from time to time has allowed us to find the last 20 tweets containing at least one of the keywords mentioned in the query. The search interface had some restrictions on query length and allowed frequency of requests [Twitter, Inc. Rest api: Rate limited. https://dev.twitter.com/docs/ rate-limits / 1.1, October 2014.]. Standard keywords return the most results in each search and, if there is a need for a delay in each successive query, this can cause the system to miss important tweets. To improve our coverage we have improved the search and distribution of keywords in an effort to capture all the same tweets. Unfortunately, this search-based approach combined with (ultimately unknown to us) ways to keep the Twitter server unable to verify the full sample.

In October 2009, Twitter released the first version of its Streaming API [Twitter, Inc. Streaming api. https://dev.twitter.com/docs/api/streaming, October 2009]. This interface supports unlocking single connections to Twitter servers and getting continuous streaming of all tweets like specific situations, or filters. These filters can be written so that the connection will return all the tweets that are similar to certain names or written by specific users. This API is efficient and effective in finding all posts that are related to a particular domain (it also provides a "sample" conclusion that returns a random fraction of all tweets posted, independent of a domain). About a year later, in September 2010, Twitter introduced a location-based filter (e.g., Location) [Twitter, Inc. Streaming api: Local-based filters. https: //dev.twitter. com / docs / streaming-apis / parameters # locations, September 2010.] in the Streaming API. Thanks to this new filter it is now possible to retrieve all tweets posted to a specific

location in the world, defined by a rectangular box marked with links to its corners.

As in the search forum, the "limit" also applies to the streaming API. If the filters used are too wide, unlimited messages will appear in the tweets to let the developer know that the feed is incomplete. The use of smart filters may allow for more research-related use in streaming data with just a free account, but complete data collection will require special business agreements with Twitter. Basically, the company offers access to "firehose" which allows it to receive all converted tweets on the network.

3.2.4 Data Gathering and Normalization

Since tweets are mostly made by users, their content is often dirty. For example, some tweets contain non-ASCII characters, while others contain tweet specific jargon keywords (e.g., “RT” for retweet), hashtags (e.g., "#awesome"), usernames (e.g., "@a signorini") or links. To ensure we have a clean and usable database, we applied the following cleaning steps to each tweet:

1. As long as the tweet contains HTML elements, cut them out;
2. If the tweet contains non-ascii characters or is less than 5 characters, it should be discarded.
3. Small tweet words;
4. Then delete:
    a. All numbers and initials;
    b. All terms of reference (eg, http: // ”); and
    c. All policies starting with @ and #;
5. Ignore the tweet if it is less than 5 characters.

When the text was clear, we found a set of words separated by white spaces. Because tweets are frequently sent from mobile devices, there is an increased chance of introducing typos and missing missions. At the time of our review, we have removed all terms that are irregular (less than 5 times) or shorter than 3 characters

3.2.5 Stemming

To further clean up our database, use it to follow the remaining terms. Stemming algorithms have been studied in Computer Science since the 1960s and are now widely used in data retrieval. Their main goal is to reduce each word to its root (e.g., “Sickest” to “sick”). Four basic algorithms look for pre-

loaded tables with popular and unique names. These are quick and easy to understand. Another category of algorithms makes a suffix and relies on a small set of rules (or steps) used in the term to obtain its root form. The most popular algorithm of this type was developed by Porter in 1980 [M.F. Porter. Sorting algorithm. Schedule, 14 (3): 130–137, 1980.].

Lemmatization algorithms perform part-time analysis to reduce a set of rules that can be used in a term. This category of algorithms is usually more advanced than an appendix failure, but any errors in the identification of the appropriate term category limit the additional benefits of injury. An example of this category of algorithm is Y-Stemmer [VA Yatsko. Methods and algorithms for automatic text analysis. Default Texts and Mathematical Language, 45 (5): 224-231, 2011.].

Finally, stochastic algorithms include the use of opportunities to determine which root source is most likely to occur. These algorithms are usually trained in specific input data from the language of interest to create a possible model, and then produce a possible root for each word in the test set.

While many other categories of determining algorithms exist (as well as certain hybrid methods, especially in Arabic text [Tahar Dilekh and Ali Behloul. Implementation of a new hybrid method for arabic text filtering. Analysis, 3 (4): 5 , 2012.], many buildings simply use the well-known Porter Algorithm. Many use [Gonzalo Parra, Steve Dyrdahl, and Brian Goetz. .] of this algorithm exists (some even written by Porter himself) but the standard English version is almost always based on the 6 basic steps [Ilia Smirnov. Overview of composing algorithms. Mechanical Translation, 2008] in each term:

1. click plurals and suffixes -ed or -ing;
2. replace the end y by i if there is another vowel in the stem;
3. reduce the double suffixes by one (e.g., Fool-ish-ly, care-less-ly);
4. remove known morphological suffixes (-icate, -ative, -alize, -iciti, -ical, -ful, -ness);
5. remove some known suffixes6 (-able, -ment, etc.); and
6. delete the last -e.

In our experiments we used the well-known Lingua :: Stem :: En library located in CPAN. This algorithm is based on the use of Frakes and Cox in 1986 for Porter’s Algorithm deterrence.

CHAPTER 4

CONCLUSION

Our results show that social media data can be used to track user interest and concerns related to social topics (e.g., H1N1 flu), as well as accurately measure the spread of disease. Although based on descriptive theory, because no comparable data (e.g., research results) are available, it is not possible to confirm some of the results, visual patterns are reasonable and are very consistent with expectations. For example, the initial interest of Twitter users on antiretroviral drugs such as oseltamivir declined almost simultaneously as official disease reports indicated that most cases were naturally mild, despite the fact that the number of cases was still growing. Also, interest in hand hygiene and face masks appeared to be outdated by public health messages from the CDC about the outbreak in early May. Interestingly, in October 2009, concerns about the deficit did not arise and there was no interest in the adverse side effects, probably because they did not occur in any widespread way. Here, the lack of an available signal can indicate a carefree society, or it may simply indicate a lack of media awareness. In any case, our work suggests a way to capture these concerns in real time, pending future studies to confirm our results using appropriate data analysis techniques.

Our research also shows that it is possible to measure people and disease activity (e.g., movement and flu movement) in real time using social data. While the flu is well-known and recurring each season with standard cycles, location, time, and size vary, complex attempts to generate reliable and timely work estimates using traditional time series models [80]. The literature provides several examples of "syndromic methods" for anticipating or predicting ILI, including the analysis of call-in-a-phone calls [31], purchases of drugs purchased for respiratory diseases [48] [23], and absenteeism [58]. While it is psychologically possible to collect diagnostic level data in real-time from emergency department visits [53] [115], doing so at a national level would require mixing, at great cost, data sources from various parts of the world and multiple firms (in the case of pharmacy data or payment data) . Furthermore, while these efforts may reveal information about flu days in the coming days to weeks before traditional sources (e.g. ILI observations), it is difficult to compare these methods, as different regional regions were studied and different statistical methods were used [22].

In contrast, our measurement method is based on well-understood machine learning methods and uses the spread of publicly available tweets as input. The accuracy of the real-time ILI presented clearly shows that the set of tweets identified and used in our models contains information closely related to disease activity. Our results show that we have been able to build a distinct relationship between Twitter data and the H1N1 outbreak of the 2009 H1N1 outbreak, at national and regional levels. Our approach, unlike others [69], does not attempt to predict flu activity, but aims to provide real-time estimates. However, because our results are available "live" (e.g., as soon as data is entered), ratings are available sooner than traditional public health reports, which usually leave the ILI function within a week or two.

If future results are consistent with our findings, similar Twitter-based monitoring efforts and similar ongoing efforts in two European research groups [85] [57] could provide significant and cost-effective additions to traditional diagnostic procedures, particularly in areas of the United States where -tweet is high. We suggest that Twitter data can also be used as a proxy for the performance of public health messages or public health campaigns. Our ability to find styles and ensure recognition from traditional monitoring methods makes this new approach a promising area of research in the visual link between computer science, pathology, and medicine.

REFERENCES

1. **Alberto Maria Segre, Andrew Wildenberg, VeronicaVieland, and Ying Zhang**. *Privacy-preserving data set union*. In Privacy in Statistical Databases, pages 266–276. Springer,2006.
2. **Alberto Maria Segre, James Cremer, Ted Herman, Padmini Srinivasan, PhilipPolgreen**. *Use of social media to monitor and predict outbreaks and public opinion on health topics*.
3. **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.
4. **Andrea Freyer Dugas, Mehdi Jalalpour, Yulia Gel, Scott Levin, Fred Torcaso, Takeru Igusa, Richard E. Rothman**, *Influenza Forecasting with Google FluTrends*
5. **Andrew McCallum, Kamal Nigam**, et al. *A comparison of event models for naive Bayes textclassification*. In AAAI-98 Workshop on learning for text categorization, volume 752, pages 41–48. Citeseer,**1998**.
6. **Andrew McCallum, Kamal Nigam**, et al. *A comparison of event models for naive bayes textclassification*. In AAAI-98 workshop on learning for text categorization, volume 752, pages41–48. Citeseer,1998.
7. **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.
8. **B. T. Grenfell, O. N. Bjornstad, and J. Kappey**. *Travelling waves and spatial hierarchies in measles epidemics. Nature*, 414(6865):716–723, 122001.
9. **Birkhead G. S. and C. M. Maylahn**. *State and local public health surveillance. Principles and Practices of Public Health Surveillance*, page 270, 2000.
10. **Brad Templeton**. *Origin of the term "spam" to mean net abuse*.http: //www.templetons.com/brad/spamterm.html.
11. **BT Grenfell, ON Bjørnstad, and J Kappey**. *Travelling waves and spatial hierarchies in measles epidemics*. Nature, 414(6865):716–723, 2001.
12. **C.C. Freifeld**, et al. HealthMap: *global infectious disease monitoring through automated classification and visualization of Internet media reports*.

13. **Charlene Babcock Irvin, Patricia Petrella Nouhan, and Kimberly Rice**. *Syndromicanalysis of computerized emergency department patients' chief complaints: an opportunity for bioterrorism and influenza surveillance*. Annals of emergency medicine, 41(4):447–452,2003.
14. **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.
15. **Corley CD, Cook DJ, Mikler AR, Singh KP.** *Text and structural data mining of influenza mentioned in web and social media*. Int J Environ Res Public Health.
16. **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.
17. **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.
18. **Cynthia Chew, Gunther Eyesenbach**, *Pandemics in the Age of Twitter*: Content Analysis of Tweets during the 2009 H1N1Outbreak.
19. **Dennis D Lenaway and Audrey Ambler**. *Evaluation of a school-based influenza surveillance system*. Public health reports, 110(3):333,1995.
20. **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.
21. **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.
22. **Eui-Ki Kim, Jong Hyeon Seok, Jang Seok Oh, Hyong Woo Lee, Kyung Hyun Kim** *Use of Hangeul Twitter to Track and Predict Human InfluenzaInfection*.
23. **Eysenbach Infodemiology**: tracking flu-related searches on the web for syndromic surveillance.
24. **Facebook**, Ericsson and Qualcomm. A focus on efficiency. Technical report, Facebook, Ericsson and Qualcomm, September 2013.
25. **Facebook**. Company info. https://newsroom.fb.com/company-info/, October 2014.
26. **Facebook**. Facebook login overview. https://developers.facebook.com/ docs/facebook-login/overview/v2.1, October 2014.

27. **FDA**. Adverse event reporting system (faers). http://www.fda.gov/Drugs/ GuidanceComplianceRegulatory/Information/Surveillance/Adverse/Drug/ Effects, October 2014.
28. **Foursquare**. 2010: Our year of 3400 https://foursquare.com/infographics/ 2010infographic, 2010.
29. **Foursquare**. About. https://foursquare.com/about, October2014.
30. **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.
31. **Gilad Mishne and Maarten de Rijke Gilad**. *Moodviews: Tools for blog mood analysis*. ICWSM, 2007.
32. **Gonzalo Parra, Steve Dyrdahl, and Brian Goetz.** *Java implementation of the original porter algorithm*. http://tartarus.org/martin/PorterStemmer/java.txt,2000.
33. **Google**. The google maps geolocation api. https://developers.google. com/maps/documentation/business/geolocation/, October2014.
34. **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.
35. **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.
36. https://nlp.stanford.edu/
37. **IgniteSpot**. Small business marketing idea. http://www.ignitespot.com/small-business-marketing-idea/, October 2014.
38. **Ilia Smirnov**. *Overview of stemming algorithms*. Mechanical Translation,2008.
39. **International Health Terminology Standards Development Organization**. Snomed clinical terms. http://www.ihtsdo.org/snomed-ct, October2014.
40. **Internet World Stats**. Internet world statistics. http://www.internetworld/ stats.com/stats.htm, October 2014.
41. **Jagdish Rebello**. *Four out of five cell phones to integrate gps by end of 2011*. Technical report, IHS Technology, July2010.
42. **Janice Tanne**. Restricting air travel may slow the spread of flu. BMJ, 333(7568):568,2006.
43. **Jeff Schneider**. *Cross validation*. http://www.cs.cmu.edu/~schneide/tut5/ node42.html, September 1998.

44. **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.
45. **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.
46. **Lara Hejtmanek**. *American idol winner: Can google predict the results?* Mashable, May 2009.
47. **M. C. Gibbons** *The Impact of Health Oriented Social Media Applications on Health Outcomes*.
48. **M. Gerber**, *Predicting crime using Twitter and kernel density estimation*, Decision Support Systems, vol. 61, pp. 115-125,2014.
49. **M. Odlum and S. Yoon**, *What can we learn about the Ebola outbreak from tweets?,* American Journal of Infection Control, vol. 43, no. 6, pp. 563-571, 2015.
50. **M.F. Porter**. *An algorithm for suffix stripping*. Program, 14(3):130–137, 1980.
51. **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 projects*. Soc Sci Med, 57(10):1925–1937, Nov2003.
52. **Mark Walsh**. *Search wrong on "idol"*. MediaPost, May2009.
53. **Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi**. *Understanding individual human mobility patterns*. Nature, 453(7196):779–782, 2008.
54. **MDG Advertising**. Images account for 36 percent of all twitter links shared. Technical report, MDG Advertising,2013.
55. **Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C**. *A new dimension of health care: a systematic review of the uses, benefits, and limitations of social media for health communication*.
56. **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.
57. **P. Velardi** et al., *Twitter mining for fine-grained syndromicsurveillance.*

58. **PEW Internet**. Health fact sheet. http://www.pewinternet.org/fact/ -sheets/health-fact-sheet/, 10 2014.
59. **PEW Internet**. Social networking: Fact sheet. http://www.pewinternet. org/fact-sheets/social-networking-fact-sheet/, October2014.
60. **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*. HICSS-42,2009.
61. **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.
62. **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.
63. **Q. Yuan, E. Nsoesie, B. Lv, G. Peng, R. Chunara and J. Brownstein**, *Monitoring Influenza Epidemics in China with Search Query from Baidu*, PLoS ONE, vol. 8, no. 5, p. e64323,2013.
64. **Ramaki E, Maskawa S, Morita M.** *Twitter catches the flu: detecting influenza epidemics using Twitter. : Association for Computational Linguistics*; 2011 Presented at: The Conference on Empirical Methods in Natural Language Processing (EMNLP); July 27-31, 2011; Edinburgh, United Kingdom p. 1568-1576 URL:https://dl.acm.org/citation.cfm?id=2145600
65. **Ravi Kumar Suggala** *A Survey on Prediction and Detection of Epidemic DiseasesOutbreaks*.
66. **Rees**, E. E *Early detection and prediction of infectious diseaseoutbreaks*.
67. **Regenstrief**. *Logical observation identifier names and codes*. http://loinc. org/, October2014.
68. **Ryan Kelly**. *Twitter study reveals interesting results about usage of 40 pointless babble*. Technical report, Pear Analytics,2009.
69. **S B Thacker and R L Berkelman**. *Public health surveillance in the United States*. Epidemiol Rev, 10:164–190, **1988**. PIP: TJ: EPIDEMIOLOGICREVIEWS
70. **S B Thacker and R L Berkelman**. *Public health surveillance in the United States*. Epidemiol Rev, 10:164–190, 1988. PIP: TJ: EPIDEMIOLOGICREVIEWS.
71. **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.
72. **Sarah Vieweg, Leysia Palen, Sophia B Liu, Amanda L Hughes, and JeannetteSutton**. *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.

73. **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.
74. **SichuanOnline**.*Qqpostinggirlhelpedhelicopterairbornewenchuan*.http: //news.xinhuanet.com/society/2008-05/18/content_8198824.htm, May 2008.
75. **Statistics Brain**. Facebook statistics. http://www.statisticbrain.com/ facebook-statistics/, July 2013.
76. **Susannah Fox and Kristen Purcell**. *Social media and health*. Technical report, PewResearch, 2010.
77. **Sysomos**. *Haitian earthquake dominates twitter*. http://blog.sysomos.com/2010/01/15/haitian-earthquake-dominates-twitter/, January 2010.
78. **Tejas Shinde; Parikshit Thatte; Sachin Sachdev; Vidya Pujari** *Monitoring of Epidemic Outbreaks Using Social MediaData*
79. **The International Council on Medical & Care Compunetics**. *The potential of twitter for early warning and outbreak detection*, April2010.
80. **The Nielsen Company**. *What's empowering the new digital consumer?* Technical report, The Nielsen Company,2014.
81. **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, February2012.
82. **Twitter** Inc. Healing haiti. https://blog.twitter.com/2010/healing/ -haiti,2010.
83. **Twitter**, Inc. #superbowl. https://blog.twitter.com/2011/superbowl, February2011.
84. **Twitter**, Inc. About twitter, inc. https://about.twitter.com/company, October2014.
85. **Twitter**, Inc. Api overview: Tweets. https://dev.twitter.com/overview/ api/tweets, October 2014.
86. **Twitter**, Inc. Rest api: Get favorites list. https://dev.twitter.com/rest/ reference/get/favorites/list, October 2014.
87. **Twitter**, Inc. Rest api: Rate limiting. https://dev.twitter.com/docs/ rate-limiting/1.1, October 2014.
88. **Twitter**, Inc. Rest api: Search tweets. https://dev.twitter.com/docs/ api/1.1/get/search/tweets, October 2014.

89. **Twitter**, Inc. Streaming api: Location based filters. https://dev.twitter.com/docs/streaming-apis/parameters#locations, September2010.

90. Twitter, Inc. Streaming API.https://dev.twitter.com/docs/api/streaming, October2009.

91. **Twitter**, Inc. Streaming api. https://dev.twitter.com/docs/api/streaming, October2009.

92. **U.S. Agency for International Development**. *Infectious disease and response strategy* 2005. Washington, DC., 2005.

93. **U.S. Census Bureau**. *Computer and internet use in the United States*, May 2013.

94. **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.

95. **U.S. Department of Transportation's Bureau of Transportation Statistics**. December 2013 u.s. airline system wide passengers up 6.1 percent from december 2012. http://www.rita.dot.gov/bts/press_releases/bts012_14, March2014.

96. **VA Yatsko** . *Methods and algorithms for automatic text analysis*. AutomaticDocumentation and Mathematical Linguistics, 45(5):224–231,2011.

97. **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.

98. **Vasileios Lampos, Nello Cristianini** *Tracking the flu pandemic by monitoring the SocialWeb*

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

100. **W3 Techs**. *Usage statistics and market share of wordpress for websites*. Technical report, W3 Techs,2014.

101. **Wall Street Journal**. Foursquare week. http://graphicsweb.wsj.com/docs/FOURSQUAREWEEK1104/bygender.php,2011.

102. **WikiMedia Foundation**. Strategic plan 2011. http://upload.wikimedia.org/wikipedia/foundation/c/c0/WMF_StrategicPlan2011_spreads.pdf, October2011.

103. **Wikipedia**. Wikipedia - number of editors.http://en.wikipedia.org/wiki/Wikipedia:Wikipedians#Number_of_editors, October2014.

104. **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.

105. **Wired Xeni Jardin**. *Text messaging feeds rumors*. http://archive. wired.com/medtech/health/news/2003/04/58506?currentPage=all, April2003.

106. **Wordpress**. A live look at activity across wordpress.com. http://en.wordpress.com/stats/, October 2014.

107. **World Bank. World development report 2000–2001**: *Attacking poverty*. Technicalreport, World Bank,2001

108. **World Bank**. World development report 2000–2001: Attacking poverty. Technicalreport, World Bank,2001.

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

110. **X. Zhou, Q. Li, Z. Zhu, H. Zhao, H. Tang and Y. Feng**, *Monitoring Epidemic Alert Levels by Analyzing Internet Search Volume*, IEEE Transactions on Biomedical Engineering, vol. 60, no. 2, pp. 446-452, 2013

111. **Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu**. *Exploiting social relations for sentiment analysis in microblogging*. In Proceedings of the sixth ACM international conference on Web search and data mining, 2013.

112. **Xiao Huang, Zhenlong Lee, Yuqin Jiang , Xiaoming Li, Dwayne Porter.** *Twitter reveals human mobility dynamics during the COVID-19 pandemic*.

113. **Yi W. Jindal**. *Analysis of transmission dynamics for Zika virus on networks. Applied Mathematics and Computation*.

114. **Yiding Zhanga , Motomu Ibarakib , Franklin W. Schwartz** *Disease surveillance using online news: Dengue and zika in tropicalcountries*

115. **YushengXie, Zhengzhang Chen, Yu Cheng, Kunpeng Zhang, Kathy Lee, Ankit Agrawal. Wei-Keng Liao, Alok Choudhary**, *Detecting and Tracking Disease Outbreaks by Mining Social Media Data*.

116. **Zhange & Al**, *An electronic word of mouth on twitter about physical activity in the United States: investigating infodemiology Study*. Online medical research journal, 15 (11),2013.