Summer 2015

Social analytics for health integration, intelligence, and monitoring

Xiang Ji
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

SOCIAL ANALYTICS FOR HEALTH INTEGRATION, INTELLIGENCE, AND MONITORING

by
Xiang Ji

Nowadays, patient-generated social health data are abundant and Healthcare is changing from the authoritative provider-centric model to collaborative and patient-oriented care. The aim of this dissertation is to provide a *Social Health Analytics framework* to utilize social data to solve the interdisciplinary research challenges of Big Data Science and Health Informatics. Specific research issues and objectives are described below.

The first objective is semantic integration of heterogeneous health data sources, which can vary from structured to unstructured and include patient-generated social data as well as authoritative data. An information seeker has to spend time selecting information from many websites and integrating it into a coherent mental model. An integrated health data model is designed to allow accommodating data features from different sources. The model utilizes semantic linked data for lightweight integration and allows a set of analytics and inferences over data sources. A prototype analytical and reasoning tool called "Social InfoButtons" that can be linked from existing EHR systems is developed to allow doctors to understand and take into consideration the behaviors, patterns or trends of patients' healthcare practices during a patient's care. The tool can also shed insights for public health officials to make better-informed policy decisions.

The second objective is near-real time monitoring of disease outbreaks using social media. The research for epidemics detection based on search query terms entered by millions of users is limited by the fact that query terms are not easily accessible by

non-affiliated researchers. Publically available Twitter data is exploited to develop the Epidemics Outbreak and Spread Detection System (EOSDS). EOSDS provides four visual analytics tools for monitoring epidemics, i.e., Instance Map, Distribution Map, Filter Map, and Sentiment Trend to investigate public health threats in space and time.

The third objective is to capture, analyze and quantify public health concerns through sentiment classifications on Twitter data. For traditional public health surveillance systems, it is hard to detect and monitor health related concerns and changes in public attitudes to health-related issues, due to their expenses and significant time delays. A two-step sentiment classification model is built to measure the concern. In the first step, Personal tweets are distinguished from Non-Personal tweets. In the second step, Personal Negative tweets are further separated from Personal Non-Negative tweets. In the proposed classification, training data is labeled by an emotion-oriented, clue-based method, and three Machine Learning models are trained and tested. Measure of Concern (MOC) is computed based on the number of Personal Negative sentiment tweets. A timeline trend of the MOC is also generated to monitor public concern levels, which is important for health emergency resource allocations and policy making.

The fourth objective is predicting medical condition incidence and progression trajectories by using patients' self-reported data on PatientsLikeMe. Some medical conditions are correlated with each other to a measureable degree ("comorbidities"). A prediction model is provided to predict the comorbidities and rank future conditions by their likelihood and to predict the possible progression trajectories given an observed medical condition. The novel models for trajectory prediction of medical conditions are validated to cover the comorbidities reported in the medical literature.

SOCIAL ANALYTICS FOR HEALTH INTEGRATION, INTELLIGENCE, AND MONITORING

by
Xiang Ji

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

August 2015

APPROVAL PAGE

SOCIAL ANALYTICS FOR HEALTH INTEGRATION, INTELLIGENCE, AND MONITORING

Xiang Ji

Dr. James Geller, Dissertation Co-Advisor — Date
Professor of Computer Science and Associate Dean of Research, NJIT

Dr. Soon Ae Chun, Dissertation Co-Advisor — Date
Professor of Computer Science & Information Systems & Informatics, CUNY

Dr. Vincent Oria, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Mei Liu, Committee Member — Date
Assistant Professor of Medical Informatics, The University of Kansas

BIOGRAPHICAL SKETCH

Author: Xiang Ji

Degree: Doctor of Philosophy

Date: August 2015

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2015

- Bachelor of Engineering in Network Engineering,
  Southwest University, Chongqing, P. R. China, 2010

Major: Computer Science

Publications:

Journal Papers

Xiang Ji, Soon Ae Chun, Paolo Cappellari, and James Geller, "A Framework for Linking Data Sources and Providing Intelligence in Social Health Analytics," Journal of Information Science (submitted), 2015.

Xiang Ji, Soon Ae Chun, Zhi Wei, and James Geller, "Twitter Sentiment Classification for Measuring Public Health Concerns," Social Network Analysis and Mining, 5(1), pp. 1-25, 2015.

Conference Papers

Xiang Ji, Soon Ae Chun, James Geller, and Vincent Oria, "Collaborative and Trajectory Prediction Models of Medical Conditions by Mining Patients' Social Data," IEEE International Conference on Bioinformatics and Biomedicine (submitted), 2015.

Xiang Ji, Soon Ae Chun, and James Geller, "Social Data Integration and Analytics for Health Intelligence," Proceedings of Very Large Database (VLDB) PhD Workshop, Hangzhou, China, 2014.

Xiang Ji, Soon Ae Chun, and James Geller, "Monitoring Public Health Concerns Using Twitter Sentiment Classifications," Proceedings of IEEE International Conference on Healthcare Informatics (ICHI), pp. 335-344, Philadelphia, PA, 2013.

Xiang Ji, Soon Ae Chun, and James Geller, "Social InfoButtons: Integrating Open Health Data with Social Data using Semantic Technology," Proceedings of the 5th Workshop on Semantic Web Information Management (SWIM), in conjunction with the 2013 ACM International Conference on Management of Data (SIGMOD), Article No. 6, New York, NY, 2013.

Xiang Ji, Soon Ae Chun, and James Geller, "Epidemic Outbreak and Spread Detection System Based on Twitter Data," Proceedings of the 1st International Conference on Health Information Science (HIS), pp. 152-163, Beijing, China, 2012.

Demo and Poster Papers

Xiang Ji, Soon Ae Chun, Paolo Cappellari, and James Geller, "Leveraging Social Data for Health Care Behavior Analytics," Proceedings of 15th International Conference on Web Engineering (ICWE), pp. 667-670, Rotterdam, Netherlands, 2015.

Xiang Ji, Soon Ae Chun, and James Geller, "A Collaborative Filtering Approach to Assess Individual Condition Risk Based on Patients' Social Network Data," Proceedings of the 5th ACM-BCB Conference, pp. 639-640, Newport Beach, CA, 2014.

Xiang Ji, Soon Ae Chun, and James Geller, "Social InfoButtons for Patient-oriented Healthcare Knowledge Support," Proceedings of American Medical Informatics Association Annual Symposium (AMIA), 2014.

Conference Presentations

"A Collaborative Filtering Approach to Assess Individual Condition Risk Based on Patients' Social Network Data," The 5th ACM-BCB Conference, Newport Beach, CA, September 2014.

"Social Data Integration and Analytics for Health Intelligence," The 40th Very Large Database Conference, Hangzhou, China, September 2014.

"Social Analytics for Public Health Intelligence and Monitoring," Doctoral Consortium of IEEE International Conference on Healthcare Informatics (ICHI), Philadelphia, PA, September 2013.

"Monitoring Public Health Concerns Using Twitter Sentiment Classifications," IEEE International Conference on Healthcare Informatics (ICHI), Philadelphia, PA, September 2013.

"Social InfoButtons: Integrating Open Health Data with Social Data using Semantic Technology," ACM International Conference on Management of Data (SIGMOD), New York, NY, June 2013.

"A Twitter-Driven Public Concern Surveillance System with Sentiment Analysis," Google PhD Summit, New York, NY, February 2013.

"Epidemic Outbreak and Spread Detection System Based on Twitter Data," The 1st International Conference on Health Information Science (HIS), Beijing, China, April 2012.

To my parents and friends,
For their endless love and support.

感谢求学路上陪伴我的父母，老师，同学和朋友
你们的支持是我前进最大的动力！

ACKNOWLEDGMENT

I would like to express my deepest appreciation to Dr. James Geller and Dr. Soon Ae Chun, who not only served as my research co-advisors, providing valuable technical and writing guidance but also constantly gave me support, encouragement, and reassurance through my five years of Ph.D. journey. Their extensive knowledge of life, great expertise in research, and humanity has affected me profoundly and benefited me significantly in the past and will be of important value in my future career. I really feel privileged to be their Ph.D. student.

Special thanks are given to Dr. Zhi Wei, Dr. Vincent Oria, and Dr. Mei Liu for their active participating in my committee as well as their insightful feedbacks and advice on my dissertation and during my graduate study.

A great portion of my past five years was spent at Department of Computer Science. I was deeply thankful for all the staff members of this department. I would like to specially thank Dr. David Nassimi, Dr. George Olsen, Dr. Ali Mili, and Ms. Angel Butler for providing me with enormous help.

I want to give my sincere gratitude to the past and present members of Intelligent Automation Inc. where I received great summer internship training. Special thanks for my mentors Xiong Liu, Kaizhi Tang, Roger Xu and my colleagues Peng Xia, Hui Yang, Shengwei Wang, Zheng Chen, Lemin Xiao, Mingyi Zhao, Gaoyao Xiao, and Bin Zhao.

I would like to thank my friends Christopher Ochs, Zhe He, Tian Tian, Ankur Agrawal, Ling Zheng, Liangji Chen, Lei Liu, Jael Ramon, Yang Zhang, Ravi Trivedi, Yang Liu, Zhicheng Zeng, Nafi Diallo for their support and walking me through this journey.

My girlfriend, Yanfei Liu, is always considerate and helpful in all aspects. Thanks for walking with me during this journey.

Finally, I would like to express my special gratitude to my parents Aiping Hu and Tengjiu Ji, who brought me to this wonderful world, and raised me with incredible devotion and patience, and has always been teaching me to be an independent and strong human being. I believe my happiness and achievements are the best return for their unconditional love and support.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Overview

Online health-related social networks generate an exponentially increasing stream of big data [1]. This social health data is of large volume, created in real-time, and contains a high degree of noise. In order to develop the enormous potential of big data and social media to improve public health and consumer health, cutting-edge computer science techniques from the Semantic Web and Machine Learning need to be applied to the new interdisciplinary problems that are hard to solve with traditional methods. Some examples of these problems are integrating heterogonous health data sources [2], monitoring disease outbreaks in real-time [3], mining public sentiments towards epidemics [4, 5], predicting potential diseases for individual patients [6], etc. The application of big data analytics will potentially help patients, clinicians, as well as the general public to make healthcare decisions based on better use of available data, thus building a solid foundation to improve healthcare services in the 21st century [7].

This dissertation presents a *social analytics framework* for healthcare applications that can monitor and collect social health data and integrate it with other data sources for the purpose of supporting healthcare [2, 3]. Methods for performing analysis of public health events and of user sentiments were developed [3]. Techniques for identifying topics related to emerging health events or trends were implemented. The dissertation also provides a method for performing a predictive analysis for specific consumer health problems. It is desirable to correlate social health data from a patients' social network,

e.g., PatientsLikeMe [8], to predict a ranked list of likely diseases that a specific patient may suffer in the future.

1.2 Problems and Objectives

The research problems and objectives are described as follows.

The open health datasets [8-12], accessible through different platforms can vary from structured to highly unstructured. An information seeker has to spend time visiting many, possibly irrelevant, websites, and has to select information from each and integrate it into a coherent mental model. The social health analytics framework will provide the semantic integration data model [2] that represents the semantic relationships between streaming data from distributed health data sources.

Social data such as those from Twitter can serve as important resources to provide collective intelligence and awareness of public health problems in real time [13-17]. The challenges of utilizing social media data include that the volume of data is large but distributed and of a highly unstructured form. Appropriate data gathering, scrubbing and aggregation efforts for these data are required to transform them for meaningful use. The social analytics framework developed incorporates a social media data ETL (Extract-Transform-Load) component that is used to build the integrated data store. The data store can feed various visualization tools for public health status monitoring. The user-friendly, interactive tools visualize disease outbreaks and the spread of developing epidemics in space and time.

The existing public health sentiment surveillance methods [18], such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays. This social health analytics framework aims at

providing models [4, 5] to track real-time social data for public health sentiment mining for different stakeholders to supplement the current public health surveillance systems. Online social network users tend to express their real feelings freely in social media. Public health specialists might receive the general trend of health-related topics from social media instead of reading through massive amount of messages [19, 20]; they are also interested in the overall sentiments about these topics and how the strength of sentiments changes over a period of time. The framework presented in this dissertation covers the topic-based sentiment modeling of social health data by extracting topics from health-related tweets and automatically generating overall sentiment polarity judgments for these health topics. The topic-based method developed in this dissertation allows the sentiment analysis of large sets of tweets, which markedly differs from the conventional single tweet-level sentiment analysis.

Healthcare research has shown that conditions are correlated with each other. Due to the similar molecules, gene structures, and patients' life styles, the appearance of some conditions indicates the occurrence of other conditions [21]. This correlation is called *comorbidity relationship*. The comorbidity relationships are often so complex that it is difficult to comprehend them [22, 23]. When doctors prescribe medicines for a patient with certain conditions, they usually give advice for future prevention based on their professional experience, memory, and domain knowledge [24]. A disease prediction model based on the publicly available social network data was developed to represent these comorbidity relationships, and to help doctors as well as uninformed patients to anticipate potential health problems.

1.3 Approaches

The Social Health Analytics framework contains three components that are used to address the above problems:

1. Social Health Data Integration
2. Population Analytics
3. Predictive Analytics

The architecture of the Social Health Analytics framework is shown in Figure 1.1. The Data Integration contains a health data integration model, a set of analytics and inference rules, and the term-matching algorithm. The Population Analytics contains the visual analytics and the sentiment classification method. The Predictive Analytics consists of the model for predicting medical condition incidence and trajectory.

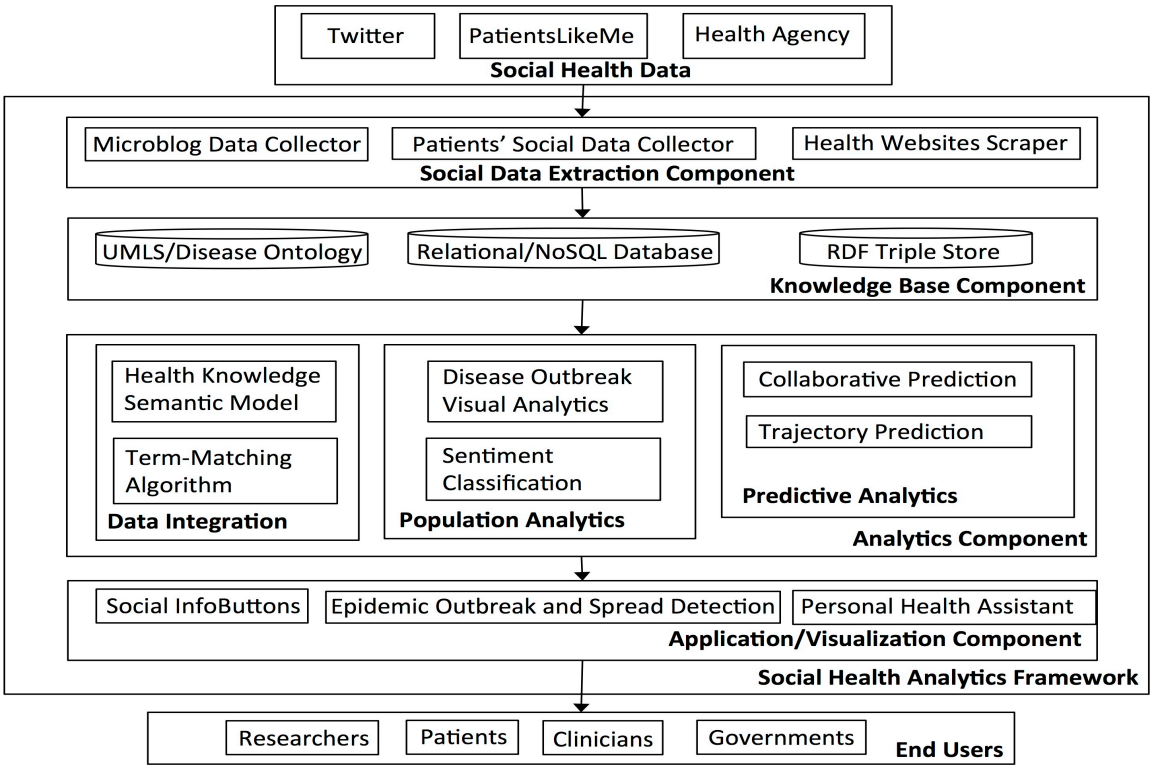


Figure 1.1 The architecture of the Social Health Analytics framework.

A detailed description of the major components is as follows.

The Data Integration component attempts to address the heterogeneity problem of online health data sources [25, 26], based on a proposed open linked data [27, 28] framework. The Data Integration component compiles information on both community and patient health issues and on healthcare trends that may shed light on each patient's care situation. The prototype system, called Social InfoButtons [2], can provide patients, public health officials, and healthcare specialists with the capability of geographically exploring the current trends of many diseases, based on patients' social network postings. For each health condition, users can search patients' social media data and compare the results with the health data published by the government. In addition, diseases-related information such as symptoms and treatments can be easily navigated to through semantic links.

The Population Analytics component contains two sub-components: Disease Outbreak Visual Analytics and Sentiment Classification. The Disease Outbreak Visual Analytics is used for detecting epidemics outbreaks and monitoring their progression over time and location based on Twitter data [3]. It allows the visual analysis of tweets with the *instance map* that shows each individual tweet's location, the *distribution (intensity) map* that displays absolute and relative frequencies of tweets from every geographic area, the *filter map* that allows users to monitor the spread of epidemics, and the *sentiment trend* that shows the public health concern on temporal and geographic dimensions.

The Sentiment Classification sub-component was developed for monitoring social network users' sentiments towards different diseases [4, 5]. This component relies on a novel two-step tweet sentiment classification method to quantify the Measure of Concern (MOC). This component allows tracking the temporal trends of the MOC about a specific disease with a timeline chart. It also provides a concern map to explore the spatial distribution of the MOC.

The Predictive Analytics component addresses problem of prediction of medical condition incidence and trajectories. Base on publicly available patients' social media data, a collaborative prediction model was developed to predict the ranked list of potential comorbidity incidences and a trajectory model was developed to reveal different paths of condition progression and predict possible condition trajectories given an observed condition of a patient.

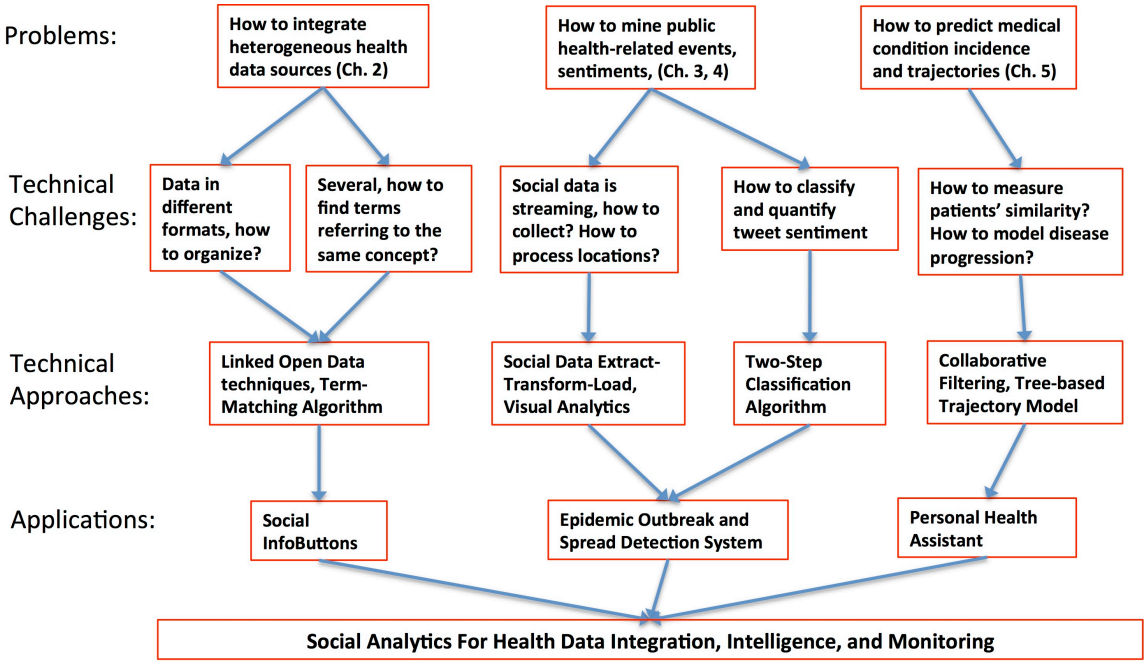


Figure 1.2 Organization of dissertation content.

1.4 Organization

The problems, technical challenges, technical approaches, and applications of this dissertation are visually summarized in Figure 1.2.

The rest of the dissertation is organized as follows. In Chapter 2, the semantic integration model of heterogeneous health data sources will be discussed. Chapter 3 will discuss the Epidemic Outbreak and Spread Detection System. The Twitter sentiment classification for measuring public health concern will be presented in Chapter 4. In Chapter 5, the disease prediction for individual conditions as well as trajectory predication for possible disease paths (progressions) are discussed. Finally, the conclusions and future work are discussed in Chapter 6.

CHAPTER 2

SOCIAL INFOBUTTONS: INTEGRATING OPEN HEALTH DATA WITH SOCIAL DATA USING SEMANTIC TECHNOLOGY

2.1 Introduction

In the past, when a patient needed information to answer a question such as "What condition causes my headache?" she had to search through pages of medical books or to see a medical expert, typically a doctor. With the emergence of the Internet, especially due to the development of search engines, today's patients can type their questions into a search engine and get related results. However, if the search query is social-oriented, such as "What are the top drugs other patients use for Asthma?" the user has to visit many, possibly irrelevant, Web pages to find an answer. The major search engines crawl billions of Web pages but they often display unhelpful results when the user wants to review Web content generated by other users.

In recent years, patients have begun to turn to social media, particularly patient communities, for personal contact, social support, and patient-generated knowledge. A study [29] by the Pew Research Center found that 34% of the Internet users used social media, such as online news group, websites, and blogs, to read other patients' commentaries and experiences about health or medical issues. There are many patient-oriented social network sites with large user communities. MedHelp [30] has 12 million monthly visitors and claims to be the world's largest health community. PatientsLikeMe [8], a fast growing social health community, currently has over 187,000 members and covers over 500 health conditions.

In addition to patient communities, city-level governments have published open health datasets for public use. NYC Open Data [11] and Chicago Data Portal [12] are examples of Open Government Initiatives [31]. At the federal level, the CDC has established the Behavioral Risk Factor Surveillance System (BRFSS) [9] based on regularly held telephone surveys. The surveys were used as an annual surveillance system for state-wide prevalence of diseases. Health information is also curated in the research community and in patient resource websites, and the medical research community has contributed a great deal of insights that patients and clinicians can use to solve their health-related problems. PubMed [10] is a database containing more than 22 million scientific publications from MEDLINE, life science journals, and online books. In a patient resource website such as WebMD [32], a patient can search for professional advice from health specialists, when faced with healthcare decisions.

Although there are many different open health data sources available, these sources are segregated, using different data formats and different platforms, making it hard to access and analyze all available health data. By integrating existing health research, clinical practice, and patient-created data, an extended and more inclusive health knowledge base can be created. This extended knowledge base enables the discovery of new information, the refinement of existing knowledge and the development of more sophisticated analyses. More importantly, this knowledge base enables to fill the gap between the officially sanctioned health science knowledge and the patient-generated crowd wisdom. For instance, healthcare providers can explore trends and statistics of clinical data from non-traditional sources, while patients can more easily find other people experiencing similar health situations. Actual patient situations (as they

experience them) can be contrasted with the views officially accepted as "correct" by healthcare researchers and practitioners. By analyzing health data in its entirety, analysis leading to early detection of community trends in medication use, and side effects of treatment methods that are not yet known "at the textbook level," can be discovered. Comprehensive health knowledge is useful to both patients, who are looking for health-related information, and to clinicians, by making them aware of what patients similar to their current patients have experienced during a particular course of treatment. In addition, government officials who are interested in the effects of health policies can determine what actually works for patients and can adjust current health policies accordingly.

Let us discuss the motivation of this work in health data management by briefly introducing a few example scenarios. These scenarios will be expanded later in this chapter. Consider a medical doctor who has to prescribe a treatment for a patient affected by a certain condition. In addition to consulting the patient's health record, and before prescribing a treatment (such as a drug), the doctor may want to conduct evidence-based medicine by exploring the social trends and experiences as described by other similar patients. By analyzing social trends, the doctor might discover implications not been mentioned in the official medical literature. Also, the doctor might find out that there are further alternative treatments that some patients have adopted. In the end, such additional information extends the doctor's knowledge, enabling her to make a better and more informed decision regarding the treatment. In another scenario, it might be the patient who desires to find out more about his/her condition or the prescribed treatment. This is a very hard task for a non-medical professional. The plethora of information available, the

specialized medical terminology, and the likely minimal expertise in "mentally digesting" medical information can make the task impossible for the patient. As a final scenario, let us consider organizations, such as non-profits or government agencies. An organization may want to monitor conditions and treatments by comparing trends between official data and social data. By aggregating and contrasting data, discrepancies can be discovered that would serve as starting point for further investigations. Again, this is not a trivial task. Thus, we advocate that there is the need for an approach to integrate health data from a multitude of sources and simplify the way users can access and interact with such data.

This chapter describes an approach to creating a health analytics framework that enables the integration and analysis of openly available health data sources, with special attention to socially generated data. We first created a health knowledge base where data from multiple open sources is included. Data from these sources is integrated and linked via Semantic Web technology. Then, on top of the knowledge base, we developed a number of analysis tools as part of a system called "Social InfoButtons" that enable end-users (e.g., doctors, government officials, patients) to become aware of socially created health information, such as treatments, conditions, experiences, attitudes, and behaviors reported by patients, in contrast with official statistics and other "official" clinical information.

2.2 Related Work

Integrating data from the Social Web is a challenging task that includes two sub-tasks (1) information extraction and (2) data integration. For the information extraction task, Raghupathi and Raghupathi [33] summarized five different sources and data types that provide useful health information. These sources and data types include Web and social

media data (e.g., PatientsLikeMe), machine-to-machine data (e.g., sensors), big transaction data (e.g., health insurance claims), biometric data (e.g., x-ray images), and human-generated data (e.g., physicians' notes). This chapter focuses on health information extraction and integration of Web and social media data, which have been proven to be viable platforms for patients to discuss health-related issues [34] and for researchers to derive health intelligence [4]. Luque et al. [25] surveyed approaches to extracting information from the "Social Web" for health personalization. They pointed out that the available data sources do not provide APIs for the integration with third-party applications. This could partially explain why there are few applications in this area. There are still notable gaps between professional experts and Web health users. Smith et al. [26] found that only 43% of the PatientsLikeMe symptom terms are present, either as exact matches (24%) or synonyms (19%), in the Unified Medical Language System Metathesaurus (UMLS). Their study reaffirmed the challenges that both the online patients and professional health specialists face, namely the need to navigate the differences between unfettered natural language descriptions and restricted terminologies as well as formalized knowledge sources.

For the data integration task, the Semantic Web has been used as a framework for data integration in various scientific fields. Most of the work in this thread follows Linked Open Data (LOD) [27, 28] principles to create links between resources distributed in heterogeneous data sources. LOD principles require using URIs to identify resources, RDFs to represent information, and typically the use of SPARQL to access the information. Sheth et al. [35] reviewed the viability of Semantic Web for data integration. Harth and Gil [36] described a scenario for geospatial data integration and querying with

Semantic Web technology. Specia and Motta [37] integrated folksonomies in a social tagging system with an ontology. Fox et al. [38] developed a semantic data framework to provide a formal representation of concepts across the fields of solar physics, space physics, and solar-terrestrial physics.

In the field of Health Informatics, the study of Chun et al. [39] proposed a preliminary semantic integration model of different health data sources, that can help with annotating social health blogs. MacKellar et al. [40] developed a clinical trial knowledge repository to pull together data from clinical trials and from other data sources, such as side effect information. In the work of Tofferi et al. [41], clinical trial data is integrated with drug data to support end users at finding an appropriate clinical trial for them to participate in, but their study does not include social data. LinkedLifeData [42] is a website providing platforms for semantic data integration through RDFs and through SPARQL queries to an integrated knowledge base. Different from previous work, which focused on scientific data, the "Social InfoButtons" approach of this chapter is to utilize an integrated semantic model to create a machine-readable encoding of the semantics of the contents of various open health data sources, especially social data sources. This facilitates the interoperability of open health data and provides an organized knowledge base for a Web user to retrieve desired health information while incorporating the social dimension.

In Drug Encyclopedia, which was developed by Kozak et al. [43], drug information requirements of physicians were analyzed, and drug data sources such as Medical Subject Headings (MeSH) [44], The Anatomical Therapeutic Chemical Classification System (ATC) [45], and National Cancer Institute (NCI) Thesaurus [46] were identified to cover those information requirements. The structured and unstructured

drug data sources were transformed into an RDF database, using different methods depending on the characteristics of each data source. The links between data sources were created according to certain rules intended to provide users with cross-data source queries of drug information. Social InfoButtons is different from Drug Encyclopedia in terms of data sources, information requirements, and linkage creation. The data sources in Social InfoButtons are open social sites instead of the fine-grained dictionaries used in Drug Encyclopedia. The open social sites do not have APIs, in most cases, and no well-defined data schemas, which make the integration task more challenging. Unlike Drug Encyclopedia's focus on covering physicians' information needs about medical products, Social InfoButtons covers not only doctors' needs concerning drug information, but also patients' information needs about diagnoses and community support, and healthcare providers as well as government agencies' information needs for public health surveillance purposes. In terms of linkage creation, Social InfoButtons utilizes the UMLS, instead of ad-hoc rules, to identify different term instances standing for the same concept, and this is done in a generalizable way.

The Social InfoButtons approach was inspired by the InfoButtons system and incorporated some of the InfoButtons standard questions proposed in Collins et al. [47]. InfoButtons was developed by Cimino et al. [48-50] and it is a system to complement the current Electronic Health Records (EHR) systems and meet the clinicians' information needs in the context of patient care. Cimino et al. [51] described ten different information needs, their contexts, their resources, and the corresponding applicable methods, and they concluded that the methods to implement InfoButtons included simple links, concept-based links, simple search, concept-based search, intelligent agents, and a

calculator. These clinical information needs are summarized by Collins et al.'s work [47] in the form of questions asked by clinicians. Examples of the questions are "Can drug x cause (adverse) finding y?", "What are my patient's data?", "How should I treat condition x (not limited to drug treatments)?", and "What is the drug of choice for condition x?" In Social InfoButtons, similar functionalities were implemented to provide context-aware information, but the information contains aggregated patients' social health information such as health-related issues and patients' self-reported experiences with treatments, symptoms, etc. These aggregated information elements from social network sites can help clinicians to understand context-specific disease and care patterns or trends from other similar patients at the point of care. The "Social InfoButtons" system answers the questions using a knowledge base containing user-generated content, location information, and a summary of patients' demographics, stored in a semantics-based triple store. A system like Social InfoButtons could raise awareness of healthcare issues among patients and provide them with insights into varying healthcare practices.

2.3 Knowledge Base for Social Health Analytics

To enable end-users to search and interact with multiple data sources, we need a model that reconciles and connects data from a multitude of repositories. Our goal is to model open healthcare data, with the specific focus on patients' conditions, treatments and symptoms, and with the intention to complement official records with social data. In this section, we present the design of our integration model. Before discussing the rationale behind the design of the model, let us introduce the information needs of health data users (e.g., patients, healthcare professionals, and organizations), and what information is provided by currently available sources.

2.3.1 Healthcare Users Information Needs

Patients seek information both before and after clinician consultation [52]. Before the consultation, patients seek information to make an attempt to arrive at a possible explanation of their symptoms. At the same time, patients would like to identify the healthcare providers who can give them the best treatment for their specialized conditions (e.g., high blood pressure in old-aged female). Patients also want to prepare themselves in this pre-diagnosis period with a basic understanding of the condition, treatment options, and side effects. After the diagnosis, patients read the medical info materials to find out more about the condition and to better manage their treatments. Clinicians, as reported by Collins, et al. [47], are more interested in treatment choices, drug dosages, and possible side effects of treatments. Government organizations, on the other hand, desire to monitor the geographic and gender distribution of epidemics, and to perform real-time surveillance of disease outbreaks [4]. A summary of the information needs is shown in Table 2.1.

2.3.2 Data Model for Social Health Data

In order to provide a framework that fulfills the information needs highlighted in Table 2.1, we need to understand what data the framework has to handle. From the information needs it is possible to derive the following central concepts: user data, medical condition, symptom, treatment and associated effects. In addition, it is beneficial to refer to the external sources where instances of such concepts are mentioned or discussed. These resources can be complete Web pages, besides micro-blogs and scientific articles. Abstracting, we can say that these resources are documents from some source.

Table 2.1 Information Needs of Patients, Professionals, and Organizations

User	Information Need	Examples
Patient	Pre-diagnosis	What are the symptoms for diabetes? What are the treatment options for high blood sugar?
	Post-diagnosis	What are the new research findings about breast cancer? Are my symptoms indeed caused by the diagnosed condition?
	Community Support	What patients or expert communities can provide support for a specific condition?
Clinician	Drug Choice	What are the drug options used by other patients to treat a specific condition?
	Drug Dosage	How many pills a day and how many times a day should the patients take a specific drug?
	Side Effect	What are the possible adverse effects of a specific drug, and how severe are they?
Organization	Disease Surveillance	Where are the current disease outbreaks? What is the trend of a specific condition?

Figure 2.1 depicts an Entity Relationship (ER) schema describing the concepts we need to model, along with the relationships between them. Before discussing the modeling rationale, let us remark that we do not want to model all possible health data and store it in a centralized repository. It is desirable to describe the data's summary features and links to the repository of provenance. This allows for meaningful data search and reasoning, while enabling access to data details directly in the original document in the source of provenance. Also, let us point out that linkage between data from different datasets is not explicit in this model: cross-dataset relationships are added on top of it. Cross-dataset relationships, including equivalence, subsumption, and specializations between concepts and instances from different datasets are discovered by using the Unified Medical Language System (UMLS) [53], a medical reference source combining

many ontologies/terminologies. The UMLS is used to align different terminologies and infer new facts, thus enabling cross-dataset exploration and intelligence in analytics.

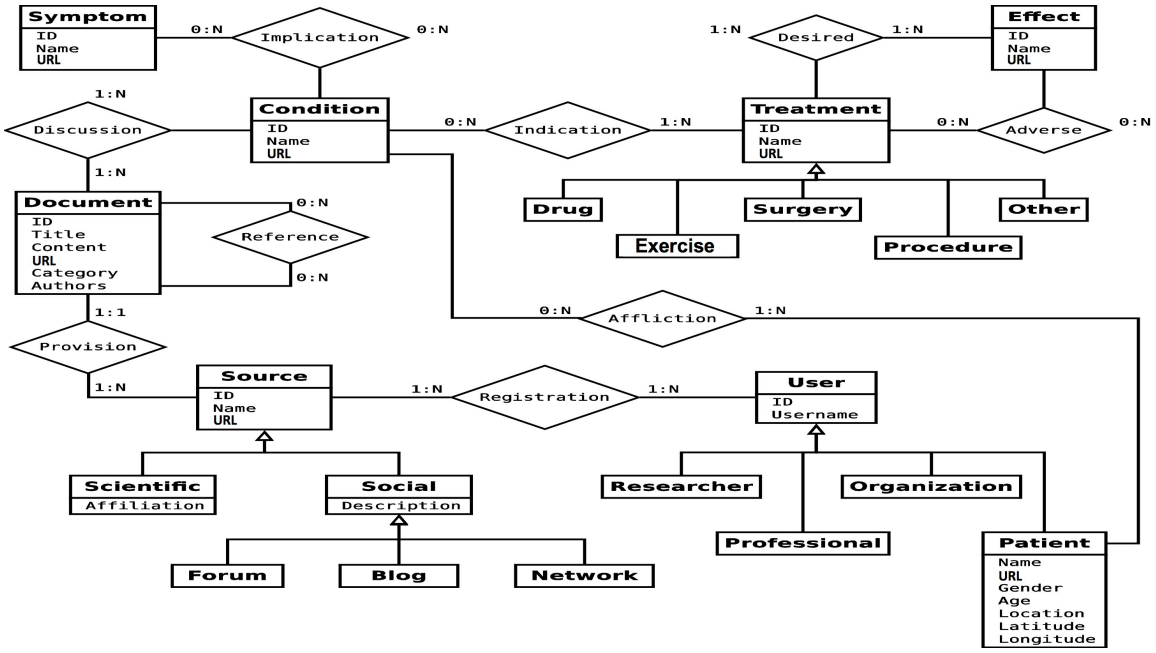


Figure 2.1 Conceptual model for semantic health data integration.

As visible in Figure 2.1, entities in the model have few attributes. Nevertheless, these attributes enable us to maintain the basic information required to implement our analysis, and will be discussed later. Now, let us describe our model. The entity *Document* describes a generic documental health resource. It is characterized by a title, a short description or summary (content), the URL where the actual document is located, a category (topic or macro-area) and a list of authors (i.e., contributors). A document can be a scientific article, a government report or a patient contribution, i.e., a blog entry or discussion contribution in a forum. Each document can refer to other documents, and it is associated with the resource provisioning it.

The resource, described by the entity *Source*, can be from a scientific or a social area. In the social area, we mainly focus on blogs, forums (discussing medical topics) and social networks (e.g., PatientsLikeMe). A medical condition is described by the entity *Condition*, and is always associated with at least one document discussing it. A condition is linked with symptoms (entity *Symptom*) and with a treatment (entity *Treatment*). The former describes an objective or perceived feeling of a patient; the latter describes what a practitioner has recommended a patient to do. The entity *Effect* describes the known effects of a treatment, including intended and collateral ones, via relationships *Desired* and *Adverse*, respectively. While some effects are the objective of a treatment (relationship *Desired*), such as "relieve pain," others are secondary, often undesired, consequences of the same (relationship *Adverse*), e.g., dizziness. Finally, with the entity *User* and its specializations, including *Patient*, we describe users' and patients' profiles and personal data. Specifically, a patient can be associated with a condition, while a user can be associated (e.g., registered) with a source, i.e., a discussion forum, social network, scientific portal, government resource, etc.

It is important to remark that data for an entity or a relationship can come from an official medical source or from crowd-generated content, that is, social content. Also, not all available data is described in our data model: our intent is to link and enable reasoning on health data, not to integrate all relevant medical information. The social nature of our model is emphasized in the relationships Affiliation and Registration. Social data voluntarily shared by patients through social networks is captured and allows discovering other patients with similar conditions as well as the resources these other patients may be

following, e.g., forums discussing specific medical topics. Finally, note that for the sake of clarity not all attributes and not all specialization entities are shown in the diagram.

2.4 Social Health Data Integration, Linkage and Storage

Publicly available health data is hosted on a variety of sources, including PatientsLikeMe, PubMed, WebMD, CDC, Twitter, Mayo Clinic and MedHelp. These sources describe and provide access to data via different representations and different platforms. In order to create the desired knowledge base, data has to be integrated, linked and stored.

2.4.1 Integration

Data sources represent data according to their internal models and make data available via different platforms. In order to include such data into our representation, some degree of data transformation is required. Often, these transformations are source specific and require ad-hoc development: each source model has specific characteristics that we have to map to ours. These transformations are intentionally kept simple. Our model aims at describing core health features, where data linking is supported by the UMLS. Thus, our effort while extracting and transforming data from all those sources is limited.

Similar to previous work by Ji et al. [3], we enrich data with geographic information to extend and improve the effectiveness of our analysis. Patients' location information can be extracted from patients' user profiles (or messages) on social networks. Generally, location is provided as a simple text-based field(s). To enable analytics including maps and geographical data, we convert user locations to latitude and longitude. This process is known as geo-coding.

2.4.2 Linkage

Data from multiple sources may use different terms to refer to the same concept, be it a condition, a symptom, etc. For instance, in PatientsLikeMe a condition is referred to as "Human immunodeficiency virus," while in the CDC dataset it is referred to as "HIV." Another example is "ALS" and its synonym "Lou Gehrig's Disease." These are different terms referring to the same concept. A knowledge worker can easily understand that these terms refer to the same concept. However, given the amount of data and the multiplicity of data sources under consideration, it's impractical to rely on human inspection: there is a need for an automatic process.

In general, the problem described above is called the entity consolidation/resolution or entity disambiguation problem. Rao et al. [54] reviewed common approaches to entity disambiguation. For entity consolidation in linked open data, Hogan et al. [55] developed a method that uses explicit owl:sameAs relations to perform consolidation. In the domain of Medical Informatics, Hassanzadeh et al. [56] reported on the LinkedCT project, which utilizes exact match, string match, and semantic match to discover links between clinical trial entities, such as trials, conditions, interventions, primary outcomes, etc. In the work of Chun et al. [39], the core idea is to use the Metathesaurus of medical concepts from the UMLS [53] as a common vocabulary for multiple terms that refer to the same concept. Indeed, this is one of the *raisons d'etre* of the UMLS. Along the same line, Ji et al. [2] developed a term matching algorithm by using the UMLS to recognize identical concepts. CUIs, which are Concept Unique Identifiers for medical concepts in the UMLS, are used by the algorithm to identify the same concept with different terms.

Specifically, we have implemented a linkage method based on the term matching algorithm described by Chun et al. [39] and Ji et al. [2]. The linkage method has two rules. (1) If two conditions in two datasets are of the same name, they are regarded as the same concepts, and a linkage between the two conditions is added. (2) We collect the CUIs of two conditions from the two datasets. Each concept in the UMLS, uniquely identified by a CUI, has several synonyms associated with it. If a concept in the UMLS has a synonym equal to the condition name, the CUI of this concept is added to this condition name. When the CUIs of the first condition have an overlap with the CUIs of the second condition, the two conditions are regarded as referring to the same concept.

An example of rule (2) is illustrated in Figure 2.2, where one condition from PatientsLikeMe has the name "ALS," and another condition from the CDC has the name "Lou Gehrig's Disease." After applying rule (2) to the triples related to these two conditions, the CUIs found for the condition "ALS" are {C1456383, C0003372, C1704945, C0002736} and the single CUI for the condition "Lou Gehrig's Disease" is {C0002736}. As these two sets share a CUI "C0002736," the two conditions are regarded as referring to the same concept, thus a cross-dataset link is added between them.

A more comprehensive example illustrating linkage between multiple datasets (PatientsLikeMe, MedHelp, WebMD, Mayo Clinic) is shown in Figure 2.3, where each dataset is represented by a dashed oval. A solid oval denotes a resource, a rectangle denotes a literal, and an arrow denotes a predicate. Datasets are linked through pairs of conditions that refer to the same concept. For example, the resource plm:condition#516 in PatientsLikeMe has the name literal "COPD." The resource medhelp:condition#307

has the name literal “Chronic Obstructive Pulmonary Disease (COPD)” and the resource webmd:condition#175 has the name literal “Chronic Obstructive Lung Disease.” Finally, the resource mayo:condition#371 also has the name literal “COPD.” By applying the linkage method described previously, all of these four conditions are identified to be referring to the same concept. Thus, the linking property “sameTopic” is added between PatientsLikeMe and MedHelp, and the linkage property “sameAs” is added between PatientsLikeMe and MedHelp, as well as between PatientsLikeMe and WebMD. Note that not all predicates are shown in Figure 2.3, again for readability purposes.

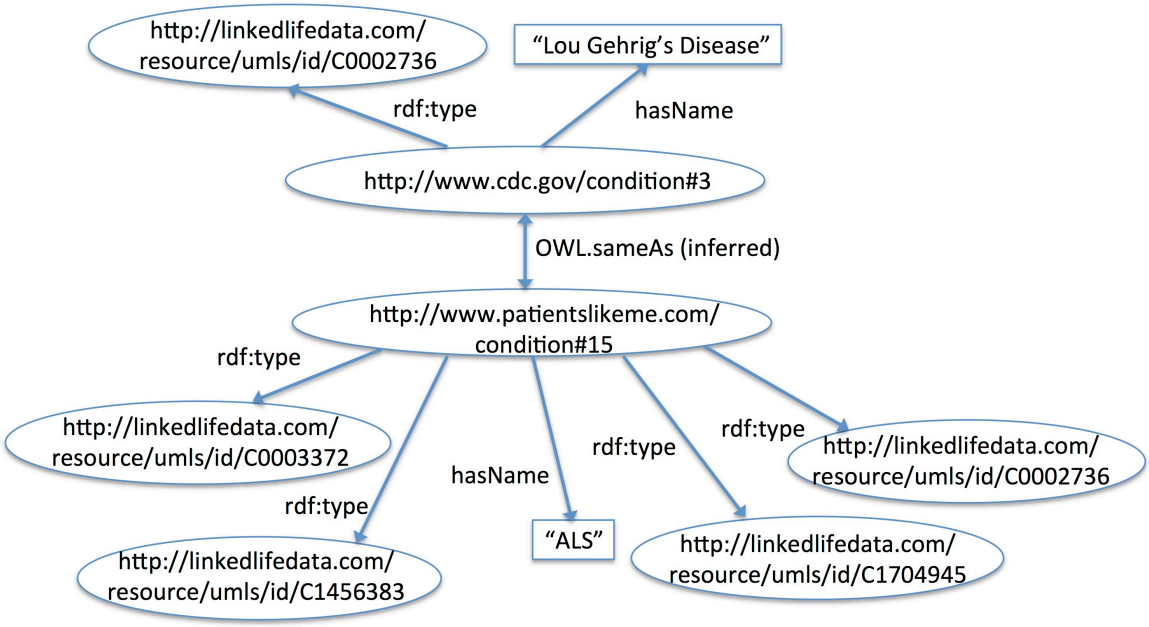


Figure 2.2 Example of inferring linkage between conditions with UMLS.

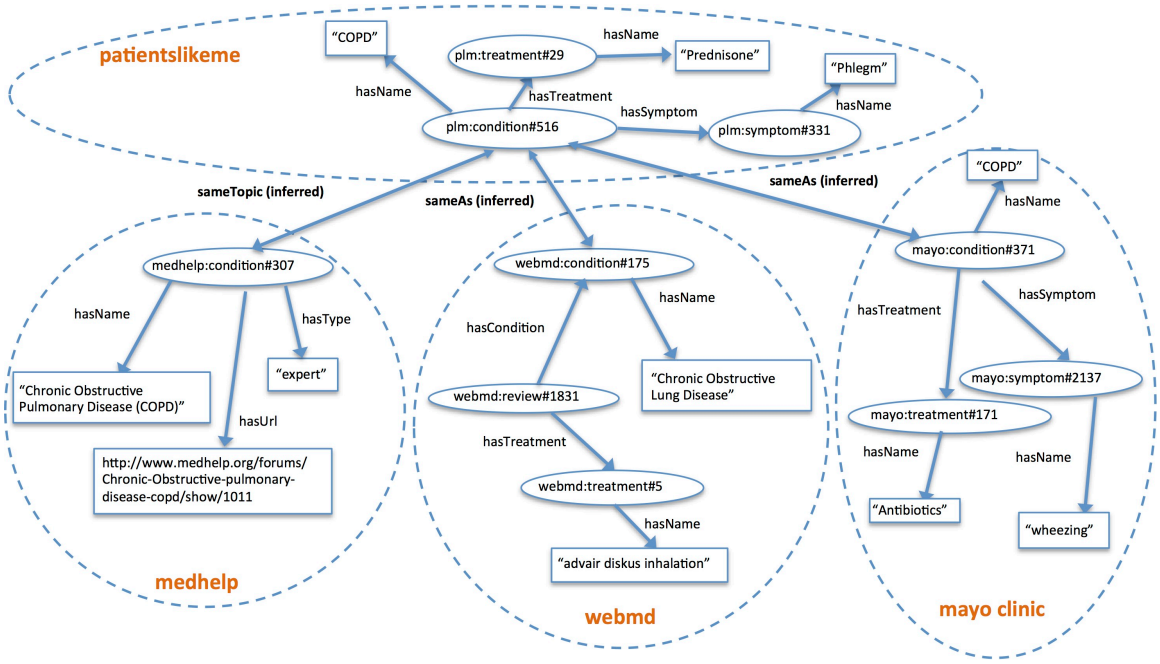


Figure 2.3 Example of inferring linkage between multiple datasets.

2.4.3 Storage

At the implementation level, integrated data is stored as RDF triples. A triple represents a statement that relates two resources, and has the format <*subject, predicate, object*>. Specifically, the subject and the object denote the resources in the statement, while the predicate denotes a characteristic of the resources and expresses a relationship between the subject and the object. The ER conceptual model is implemented in triples by reifying all attributes and relationships as properties of the entities. For example, for the entity *Patient* the identifier *ID* becomes the *URI*; the attributes *Name* and *URL* become *hasName* and *hasURL*; the relationship *Affliction* becomes *isAfflictedWith*, and links the patient with a condition. For example, in the following two RDF statements <URI_1, *hasName, "Mojomo"*> and <URI_1,*hasProfile*,URL_1>:

- URI_1 is a Unique Resource Identifier representing a unique value for a specific individual on the PatientsLikeMe network; for example, that URI could look similar to the following: http://www.patientlikeme.com/patient#1050
- URL1 is a URL denoting the identifier of the resource at which the user profile is located, such as www.patientlikeme.com/members/232328/about_me

All entities and their attributes can be represented in this format. This representation allows great flexibility compared with traditional structured data representations. In fact, when an extension of the model is required, no substantial changes are needed at the storage level. For instance, if we decide to extend the patient description by adding an ethnicity attribute, then we would need to add a new triple connecting the patient URI with a literal value specifying her ethnicity. Conversely, adding an attribute to a relational database would require an operation called "database refactoring," which could be complex and time consuming, especially if the database schema is coupled to other system components, such as application source code, a persistence framework, regression test code, etc.

2.5 Social InfoButtons

The knowledge base described in Section 2.4 supports the storage and retrieval of health data, where data stored in RDF triples can be accessed via SPARQL [57] queries. We cannot, however, expect health users to be proficient in SPARQL or any other semantic technology. For this reason, data is provided to users via a set of analytics that greatly simplify the users experience and maximize the fulfillment of their information needs. We refer to the application that includes these analytics as "Social InfoButtons" [2].

The Social InfoButtons system provides social health information delivered in a context-aware fashion, e.g., in the clinical patient care context, in the government policy evaluation context, and in the personal information look-up context, to help users find contextual information such as treatments, symptoms, etc., or to compare social data trends with official data. Social InfoButtons is able to answer questions such as "What are the top diseases reported by other patients?" or "How many male patients with Asthma are in the state of New Jersey?" According to information needs discussed in Section 2.3.1, a number of social health data analytics have been designed and implemented.

In this section, we discuss how the Social InfoButtons framework enables intelligence in social health analytics, the architecture of the Social InfoButtons implementation, and how analytics can be applied to practical scenarios, by referring to a few use cases.

2.5.1 Enabling Intelligence in Social Health Analytics

Gathering and integrating data in a unified health knowledge base is of paramount importance for healthcare information end-users. Users often want to extrapolate trends from current data and potentially discover new insights. Accessing and analyzing data can be a challenging task for end-users, especially if they are not proficient with Web technology. Discovering new information can be an even more complex task, since it requires understanding and reasoning about the data at hand. For these reasons, our framework provides two types of services, analytics and inference. The first type enables a user to analyze the information at hand; the second type enables her to infer new facts starting from those available, thus creating new knowledge.

Table 2.2 shows the set of social analytics we have implemented in the Social InfoButtons application. These analytics are the basis for implementing several common information seeking scenarios, including those described previously. Analytics are classified into the following categories: statistical, geospatial, temporal, topic investigation, association discovery and recommendation discovery. Queries in the statistical category aim to compute statistical aggregates from existing data, such as the number of patients suffering from a condition in terms of absolute and relative numbers. Geospatial analytics enable users to explore data according to a geographic feature of data, such as the location of patients as well as the concentration of health conditions in a geographical area. Queries in the temporal and topic category enable users to analyze trends over time intervals on the basis of specific topics. Association discovery analytics enable users to explore the correlation between facts such as the treatments and side effects as well as symptoms and conditions. Finally, the recommendation discovery analytics enable users to sift data to discover recommendations for a treatment given symptoms or conditions. Note that Social InfoButtons is not intended to be a medical recommender system or a replacement for professional medical advice. Any such claim would be irresponsible. It aims at promoting options that might otherwise not be known, where these options result from the collection and analysis of other patients' data. It is up to qualified medical experts to conduct further investigations into such options. The ultimate goal of a system like Social InfoButtons is to elevate the knowledge level of patients, providers, and government officials regarding current social trends in healthcare.

Table 2.2 Social Analytics and Scenarios

Category	# Query	Analytic	Scenario
Statistical	1	What are the best-reviewed alternatives for treating depression?	Enables clinicians to understand the non-traditional and best-regarded alternative treatment options.
	2	How many patients suffering from a condition?	Enables doctors to understand the patient group characteristics suffering from a condition.
	3	What are the online profile, posts, and replies for a specific condition?	Enables clinicians to determine the characteristic of a condition on online health forums.
	4	What are the top conditions with the most patients?	Enables to explore the most popular user-reported conditions.
Geospatial	5	What is the location of patients with a condition?	Enables users to understand the geographic distribution of a condition.
Temporal	6	Compare temporal sentiments toward two treatments.	Enables the comparison of the sentiment trends of different treatment options.
Topic Investigation	7	What are the top-10 most frequently discussed topics and related articles?	Enables patients to seek social support and discover non-traditional treatment plans.
Association Discovery	8	What are potential conditions for symptom of excessive saliva and online posts about it?	Enable clinicians to target possible conditions and browse and identify top issues people discuss online for a symptom.
	9	What are the top-5 frequently used drugs for a specific condition and side effects and reviews?	Enables the discovery of the association between drugs and side effects as reported in social media.
Discovery	10	Recommend a treatment for a condition to my patient.	Discover potential treatment recommendations for a patient with a condition.

Table 2.3 Analytic Queries SPARQL Code

# Query	SPARQL Code
2	select (count(?pid) as ?count) where { ?pid patientns:hasUserName ?pname. ?pid patientns:hasCondition ?cid. ?cid conditions:hasConditionName “PTSD”.}
3	select distinct ?cname ?plm_url (count(?pid) as ?medhelp_postcount) (sum(?c) as ?medhelp_replycount) where { ?c1 conditionns:hasConditionName ?cname. ?c1 conditionns:hasConditionUrl ?plm_url. ?c1 <http://www.w3.org/2002/07/owl#sameAs> ?c2. ?c2 medhelp_communityns:hasPost ?pid. ?pid medhelp_postns:hasReplyCount> ?c. } group by ?c1 ?cname ?plm_url
4	select ?cname (count(?cname) as ?dist) where { ?pid patientns:hasCondition ?cid. ?cid conditions:hasConditionName ?cname } group by ?cname order by desc (?dist) limit 10
5	select ?pname ?pprofile ?plat ?plng where { ?pid patientns:hasUserName ?pname. ?pid patientns:hasProfile ?pprofile. ?pid patientns:hasLatitude ?plat. ?pid patientns:hasLongitude?plng. ?pid patientns:hasCondition ?cid. ?cid conditionns:hasConditionName “MS” filter(?plat != 0 && ?plng != 0). }

Table 2.2 presents a set of analytics that we have embedded in the Social InfoButtons application. Results from analysis are presented to users via a Web interface, detailed later in this paper. These analytics are implemented by SPARQL queries. A

query designer implements SPARQL queries that are then linked to a visualization technique for presentation purposes. Clearly, more analytics can be built on top of our RDF health repository via SPARQL. Thus, the set of analytics can be extended relatively easily. Table 2.3 shows the SPARQL code for some of the above queries.

In addition to analytics, our framework allows to infer information by reasoning on data. On one hand, since all data are in RDF format and are linked via the UMLS, new facts can be inferred by the use of reasoning tools such as Pellet [58]. On the other hand, new knowledge can be deduced by adding inference rules. These inference rules can be defined by domain experts to enrich the current dataset. Table 2.4 shows a set of the inference rules we have defined and implemented.

Table 2.4 Inference Rules and Scenario

Inference Rule	Scenario	Jena Rule Syntax
If a patient P has a condition C AND If condition C has a treatment T ->P has treatment option T	Enrich the triple store by suggesting treatment options for patients.	[TreatmentOption: (?pid conditionns:hasCondition ?cid) (?cid treatmentns:hasTreatment ?tid) ->(?pid patientns:hasTreatmentOption ?tid)]
If a patient P has a condition C AND If a condition C has treatment T AND If a treatment T has side effect S ->P will potentially suffer from side effect of S	Enrich the triple store by adding potential side effect a patient will suffer from.	[PotentialSymptom: (?pid conditionns:hasCondition ?cid) (?cid symptomns:hasSymptom ?sid) ->(?pid patientns:hasPotentialSymptom ?sid)]

While the presented inference rules are limited, we want to emphasize the potential offered by our framework. Domain experts can define more complex inference

rules to create new knowledge or to run simulations to discover whether some hypothesis triggers incoherence (a contradiction) in the knowledge base. Ultimately, the framework enables users to reason about healthcare information, thus enabling intelligence in health analytics.

2.5.2 Architecture

We implemented our approach in a prototype system called Social InfoButtons [59]. In this Section we first present the architecture of the system, then discuss its use via a few example use case scenarios.

The system architecture is shown in Figure 2.4. At the lower level of the architecture we have the data ingestion layer. This layer is responsible for extracting data from the various publicly available health data sources and reconciling data to the data model. The layer is composed of multiple connectors, one for each different type of data source. As reported in a survey paper by Luque et al. [25], most of the health websites do not provide APIs for researchers to retrieve data. Thus, a number of connectors were implemented to retrieve data from heterogeneous sources. Among others, we have a Web crawler that uses the PHP HTML DOM Parser [60] to scrape websites and to retrieve relevant information. Additional connectors can be developed as needed. Data sources currently accessed in our extraction routine include the social network site PatientsLikeMe and Twitter (through APIs), the health forum MedHelp, the government-maintained CDC site, the Mayo Clinic website, the PubMed website, and the patient resource portal WebMD. The incoming data, where applicable, goes through the geo-coding processor, where text-based location information is resolved to latitude and longitude coordinates (geo-coding) and, vice versa, coordinates are resolved to names of

places (reverse geo-coding) by using third party services. Geo-coding is required to enable geospatial analytics and to chart data on maps.

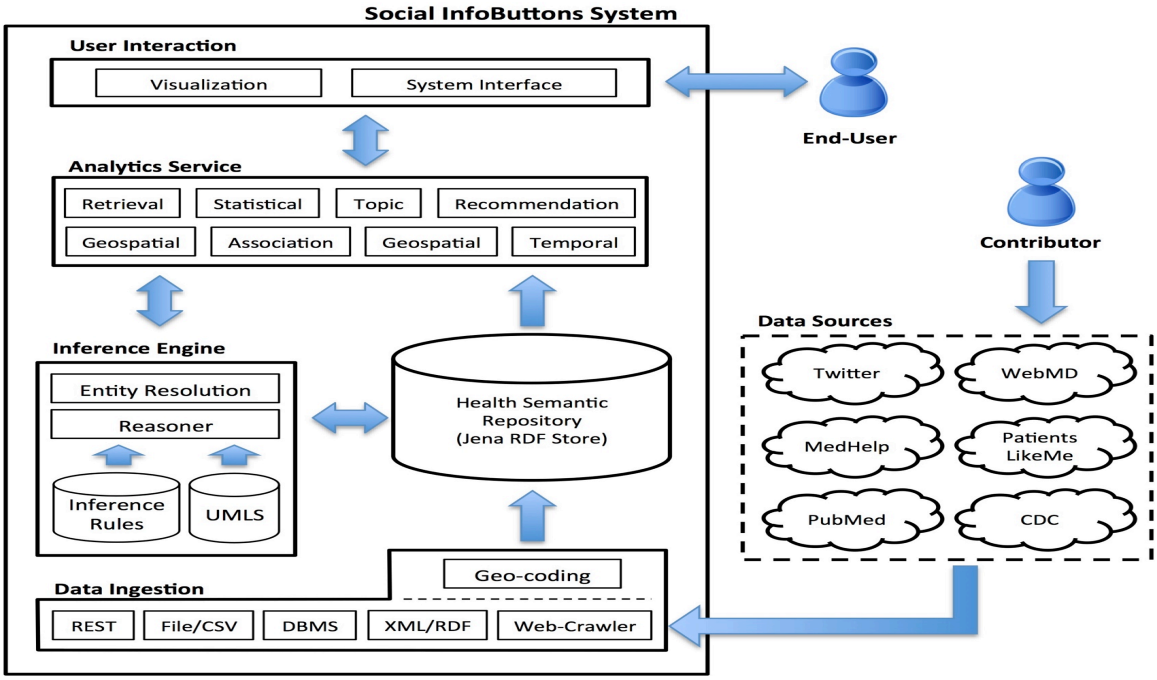


Figure 2.4 Architecture of the Social InfoButtons system.

Data is then stored in RDF format in the Jena triple store [61]. From here, data is linked and augmented via the inference engine component. The latter makes use of supplemental information specified in the UMLS, inference rules repositories, as well as of an entity resolution and a reasoning service. The inference engine is the place where data linkage is performed and additional facts are derived, thus enabling cross-dataset exploration and reasoning about data. Both the inference engine and the triple repository can be accessed via the analytics layer, which is where the analytics are deployed. At the higher level, users interact with the system via visualizations or the system interface, which invoke analytics operations according to the user input.

2.5.3 Use Case Scenarios

In the remainder of this section, we walk through the main features of the tool by describing a few representative use case scenarios. The entry point to the Social InfoButtons application is shown in Figure 2.5.

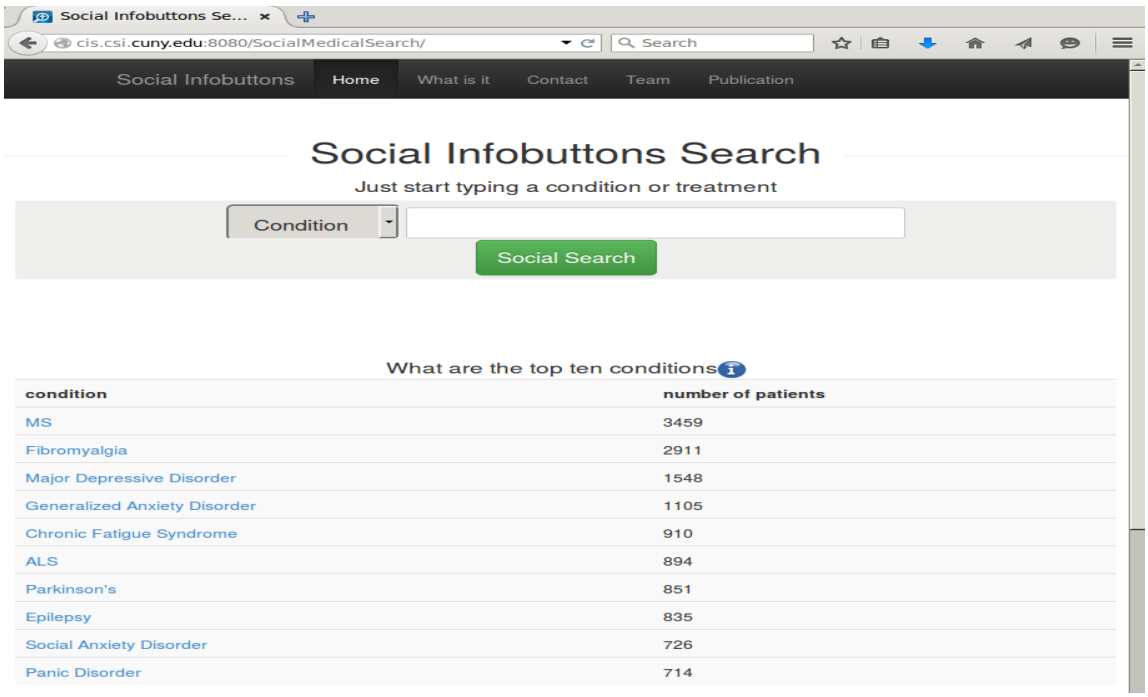


Figure 2.5 Social InfoButtons homepage.

The homepage enables users to search for conditions, symptoms or treatments by keywords, and displays the current condition trends based on data retrieved from social media. By performing a keyword search or by following the link to one of the top ten conditions, users access a contextual detail page where they can investigate condition-specific trends among patients, most frequently used drugs, symptoms, demographics, and geographical distribution of the patients. The visualization of these social data can be juxtaposed with open government data statistics and additional links to external resources such as PubMed and WebMD. Let us refer to the following scenarios:

(i) a healthcare practitioner is devising the best practice treatment for a patient; and (ii) an organization is studying discrepancies between data from official reports and social trends.

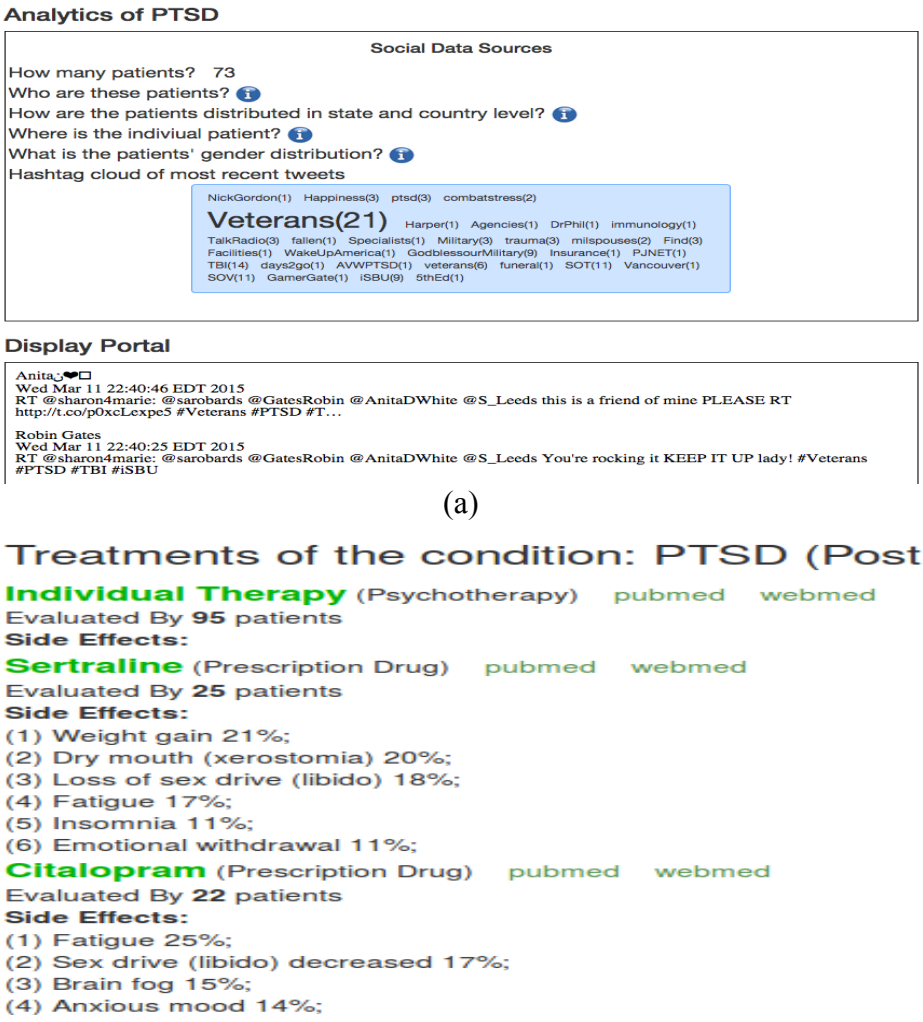
Consider a medical doctor, Christine, who has to prescribe a treatment for her patient Bob, who is a veteran and suffering from Posttraumatic Stress Disorder (PTSD). Christine would consult Bob's lab reports and electronic health record (EHR). She can decide on a prescription according to scientific recommendations and her medical knowledge. Let us assume that she is considering prescribing a drug called Sertraline. Ideally, before finalizing her decision, Christine would also conduct evidence-based medicine and explore the social trends and experiences of other patients like Bob. By analyzing social trends, she might discover implications that have not yet been sanctioned in the medical literature. To do so, she would start from the Social InfoButtons home page by performing a keyword search on the term "PTSD." Results are displayed to Christine in a page organized into four categories: 1) summary of social information (e.g. number of patients, patients' geographic distribution, topic cloud with most recent social posts, etc.), 2) list of treatments, each with associated side effects, 3) symptoms, and 4) contrast information, to enable official vs. social data comparison.

Figures 2.6(a) and 2.6(b) both show parts of the result page. Figure 2.6(a) shows the social summary for PTSD, including the number of known patients and trending topics. Christine can drill-down to access detailed information, including the patients' profile data and location distribution, and the comments associated with each trending topic. According to Figure 2.6(a) "Veterans" is a trending topic in the social space for PTSD. Christine can click on the topic term and access associated social posts (e.g.,

tweets), if she wants to know more. Figure 2.6(b) shows the list of treatments, each with a list of side effects as they are ranked by their "popularity" in the social space. By inspecting the result page, Christine discovers that a large percentage of Sertraline users have reported a side effect referred to as "Emotional Withdrawal" that is not listed in the drug documentation. At this point, if Christine decides that she wants to know more about the drug, she can follow the links, PubMed or WebMD (see the Figure 2.6(b)) that will lead her to the additional data sources and their provenance. Alternatively, Christine may decide to inform Bob about this potential side effect and advise Bob to report to her whenever this effect is observed. Conversely, she might discover that there are further alternatives that patients with PTSD are adopting and consider whether to further investigate whether there are other treatments that may suit Bob's needs better. Exploring and analyzing social data enables Christine to make a better-informed decision, because she is considering a larger, more inclusive, set of knowledge sources. Also, Social InfoButtons saves Christine from the manual, time-intensive task of accessing, reconciling and making sense of the multitude of data sources.

Now, consider another scenario where the patient, Bob, wants to know more about his condition. He would like to research the scientific literature, join social networks, explore blogs, join forums, etc. This is an even harder task for a non-medical professional. The plethora of information sources, the differences in terminology, and his own limited expertise can make Bob's task near impossible. With Social InfoButtons, Bob would follow a process similar to the one described for Christine: he would start with a keyword search, then browse the categories in the result page, eventually reading

comments from other patients or following links to contextually meaningful external resources.



(a)

(b)

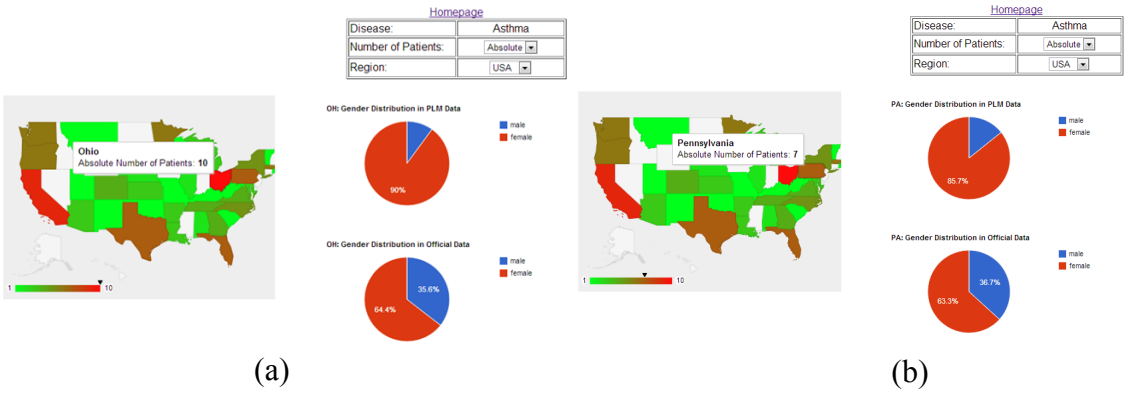
Figure 2.6 Social summary and symptoms for the condition "PTSD".

Finally, consider an organization, e.g., a government agency that wants to follow trends and understand whether discrepancies exist between official statistical data and social data. Identifying discrepancies may serve as a starting point for further investigations. Let us assume that a knowledge worker from the agency has been tasked to investigate treatments for Fibromyalgia patients that are not mentioned in scientific or

official sources. There is no universally accepted treatment for Fibromyalgia, a common chronic pain condition. The knowledge worker would start with a keyword search, as in the previous scenarios. From the result page, by browsing the data contrast area, the user can trigger queries that display analytics of contrasting data from official and social sources for Fibromyalgia. For example, an analytic reports the list of treatments for the condition, ordered by popularity (defined as the number of treatment occurrences in the social space).

Starting from this analytic, the knowledge worker can perform a comparison against authoritative sources. For this specific case, the user would discover that a treatment with Cyclobenzaprine is reported in social media data but not in official documents.

As another example, if the agency wants to explore the distribution of the population afflicted by Asthma and how it compares with official data, the user has to submit a keyword search for the term “Asthma” and click on the map analytics option in the contrast area of the result page. This user would access an interactive map, supplemented with a heat layer, where she can pinpoint the gender distribution by geographical area, and access contrast data via the given charts. Figures 2.7(a) and 2.7(b) show the gender distribution for Asthma in the states of Ohio and Pennsylvania, respectively. From these two figures it is interesting to note the following: first, there is a substantial difference between data from the official and the social sources; and, second, this difference is consistent across the states, i.e. Ohio and Pennsylvania.



(a) (b)

Figure 2.7 Interactive map showing comparison of data from official and social sources: (a) Ohio; (b) Pennsylvania.

2.6 Experiments

This Section describes the results of the use of the Social InfoButtons prototype [59]. The statistics of data sources are summarized in Table 2.5, where each cell denotes the number of entities in a specific category. Note that the data sources Twitter and PubMed are not shown in Table 2.5, since the information from both sources is dynamically retrieved through APIs during queries, thus it is not stored in the prototype triple store. Both sources are by far too large and too dynamic to represent them in the triple store. In the remainder of this section we first present how the utilized data sources cover the information needs of healthcare information users. Then, we define an evaluation metric that allows comparison between the results provided by Social InfoButtons and those from authoritative sources.

2.6.1 Coverage of Information Needs

Currently, the principal open data sources from where it is possible to retrieve substantial (medical) data are the following: PatientsLikeMe, Twitter, MedHelp, WebMD, CDC, and PubMed. These data sources provide diversified health information. Let us briefly

describe what data each source focuses on. PatientsLikeMe is a medical, patient-centric, social network. It mostly manages patients' personal and medical data, and tracks the patients' interactions with their associated conditions, treatments and symptoms. MedHelp is a platform that hosts discussion boards (e.g., forums), grouped by specific condition, between patients and health professionals. WebMD is an online service providing information about drugs along with users' reviews of each drug. CDC provides state-wide prevalence of diseases. PubMed provides comprehensive access to the medical literature. In many cases, complete publications are accessible. Twitter is a real-time micro blog platform that can be used to monitor disease outbreaks [3] and disease sentiment trends [4], although it is in not healthcare-specific. Among the information provided by Twitter, there are user posts, physical locations, and topics. Table 2.6 illustrates what information each source provides.

2.6.2 Evaluation Metric

Mean Average Precision. Mean Average Precision (MAP) is one of the most widely used measures in the field of Information Retrieval to measure system effectiveness [62] for ranked lists. MAP provides a single metric to gauge the quality of a ranked list, which is a sequence of retrieved items ordered by relevance. MAP computes the average precisions (AP) over a number of queries that a system executes and then derives the arithmetic mean of the average precisions.

Table 2.5 Statistics of Data Sources

Data Source	Patient	Condition	Treatment	Symptom	Review	Community	Post	State Prevalence
PatientsLikeMe	17,407	1,228	5,608	2,176	n/a	n/a	n/a	n/a
MedHelp	n/a	n/a	n/a	n/a	n/a	365	69,243	n/a
WebMD	n/a	647	180	n/a	86,715	n/a	n/a	n/a
Mayo Clinic	n/a	1,116	2,496	5,426	n/a	n/a	n/a	n/a
CDC	n/a	n/a	n/a	n/a	n/a	n/a	n/a	52

Table 2.6 Data Sources and Coverage of Information Needs

Data Source	Patients					Clinicians		Government
	Support Community	Pre-diagnosis	Healthcare Providers	Post-diagnosis	Drug Choice	Drug Dosage	Adverse Effect	Disease Surveillance
PatientsLikeMe		✓		✓	✓		✓	
Twitter								✓
MedHelp	✓			✓		✓		
WebMD				✓	✓	✓		
CDC			✓					✓
PubMed				✓				

To calculate the average precisions in each query, the precision at a certain cutoff points in the ranked list is computed, and then all precision values are averaged. For example, if the cutoff point is the n^{th} position in the ranked list, the precisions for item sets: $\{i_1\}$, $\{i_1, i_2\}$, $\{i_1, i_2, i_3\}\ldots\{i_1, i_2, i_3, \ldots, i_n\}$ will be computed, where i_k is the k^{th} item in the ranked list. The average precision (AP) and mean average precision (MAP) are computed with following formulas:

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times Rel(k))}{N} \tag{2.1}$$

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{Q} AP_i \tag{2.2}$$

In (2.1), N is the number of correct items, n is the number of retrieved items, and k is the rank in the sequence of retrieved items. $P(k)$ is the precision at the cutoff k in the list. $Rel(k)$ is an indicator function, which equals 1 if the item at rank k is a correct item, 0 otherwise. Q is the total number of queries to the system. If an item (treatment or symptom) in the Social InfoButtons system is mentioned in the authoritative source as a valid item, this item is labeled as correct, otherwise, it is labeled as incorrect. Therefore, for each query (condition), the ranked lists of treatments and symptoms contain both correct items and incorrect items. To evaluate the quality of the ranked lists, the proportion of correct items is crucial, but the ordering of the correct items is also important. According to the definition of average precision, it can measure both the proportion and ordering of the correct treatments and symptoms when applied to their ranked lists. For example, a sample ranked list of treatments for "Diabetes Type 2" is shown in Table 2.7. At each cutoff point (positions 1, 4, 5, and 7 in the ranked list) for a

correct item, to get the precisions, we count the number of correct treatments that have been encountered up to this cutoff point, divided by the total number of treatments seen up to this point. The precisions of correct treatments at each cutoff point are 1/1, 2/4, 3/5, 4/7, so the average precision for treatments of Diabetes Type 2 is (1/1+ 2/4 + 3/5 + 4/7)/4 = 0.67, which is a moderate result.

To evaluate the effectiveness of the Social InfoButtons system, and to illustrate how social data can have an impact on healthcare, we have reviewed the top ten conditions, shown in Figure 2.5, by comparing treatments and symptoms posted by patients against those posted by the Mayo Clinic [63], both ranked by the number of patients. The Mayo Clinic is an authoritative, well-known and trusted source.

Table 2.7 A Ranked List of Treatments for Diabetes Type 2 in Social InfoButtons (SI)

Treatments in SI	# of Patients in SI	Appeared in Authoritative Source
Metformin	159	Yes
Insulin Glargine	38	No
Pioglitazone	15	No
Victoza	13	Yes
Sitagliptin	11	Yes
Glipizide	10	No
Glyburide	9	Yes
Glimepiride	6	No
Insulin Detemir	6	No

2.6.3 Experimental Results

As discussed previously, the top ten condition names were used to query the Social InfoButtons system, and the treatments and symptoms in the results were compared with the authoritative source. For the sake of clarity of presentation, detailed results are shown for only three of the ten conditions (Fibromyalgia, Major Depressive Disorder, and Generalized Anxiety Disorder) in Tables 2.8, 2.9, and 2.10, respectively.

Table 2.8 Treatments (a) and Symptoms (b) of Fibromyalgia in Social InfoButtons (SI) and Authoritative Source (Authority)

Treatment in SI	# of Patients in SI	Appears in Authority
Duloxetine	1058	Yes
Pregabalin	955	Yes
Milnacipran	357	Yes
Gabapentin	346	Yes
Tramadol	201	Yes
Cyclobenzaprine	188	No
Amitriptyline	141	Yes
Hydrocodone-Acetaminophen	128	Yes
Naltrexone	55	No
Massage Therapy	52	No
Meloxicam	50	No
Venlafaxine	46	No
Carisoprodol	43	No

(a)

Symptom in SI	# of Patients in SI	Appears in Authority
Muscle and joint pain	20233	Yes
Pain in lower back	19102	No
Muscle Spasms	17515	No
Brain Fog	17245	Yes
Balance Problems	17177	No
Headaches	17177	Yes

(b)

The summary of results for the top ten conditions is shown in Table 2.11. For each of the top 10 conditions, we view the treatments and symptoms of each condition as two lists that are both ranked by the number of patients. By applying the average precision calculation introduced in Section 2.6.2 to the ranked lists, we get the average precisions of treatments and symptoms for the top ten conditions that are shown in Table 2.11. The mean average precisions for treatments and symptoms are 0.84 and 0.72 respectively.

Table 2.9 Treatments (a) and Symptoms (b) of Major Depressive Disorder in Social InfoButtons (SI) and Authoritative Source (Authority)

Treatment in SI	# of Patients in SI	Appears in Authority
Individual Therapy	185	Yes
Bupropion	174	Yes
Venlafaxine	160	Yes
Duloxetine	146	Yes
Fluoxetine	136	Yes
Citalopram	123	Yes
Sertraline	119	Yes
Escitalopram	79	Yes
Desvenlafaxine	30	Yes
Mirtazapine	26	Yes
Electroconvulsive-Therapy ECT	24	Yes
Aripiprazole	22	No
Lamotrigine	20	No
Quetiapine	14	No
Lithium-Carbonate	14	No

(a)

Symptom in SI	# of Patients in SI	Appears in Authority
Problems concentrating	8402	Yes
Muscle tension	7325	No
Headaches	7205	Yes
Back pain	6337	Yes
Dizziness	4900	No
Stomach pain	4898	No
Lack of motivation	4468	No
Nausea	4453	No
Low self-esteem	3847	No
Inability to experience pleasure	3062	Yes
Hyperventilation	2485	No

(b)

Table 2.10 Treatments (a) and Symptoms (b) of Generalized Anxiety Disorder in Social InfoButtons (SI) and Authoritative Source (Authority)

Treatment in SI	# of Patients in SI	Appears in Authority
Individual Therapy	122	Yes
Duloxetine	83	No
Venlafaxine	70	Yes
Clonazepam	22	No
Lorazepam	16	Yes
Citalopram	13	No
Escitalopram	13	No
Pregabalin	12	No
Sertraline	12	Yes
Alprazolam	9	Yes
Bupropion	9	No
Buspirone	8	Yes
Fluoxetine	8	No
Group Therapy	5	Yes
Hydroxyzine	4	No

(a)

Symptom in SI	# of Patients in SI	Appears in Authority
Problems concentrating	6791	Yes
Persistent worry	2479	Yes
Restlessness	2407	Yes

(b)

Table 2.11 Average Precision (AP) of Treatments and Symptoms of Top-10 Conditions

Condition	AP (Treatment)	AP (Symptom)
Multiple Sclerosis	0.95	0
Fibromyalgia	0.96	0.67
Major Depressive Disorder	1	0.695
Generalized Anxiety Disorder	0.6	1
Chronic Fatigue Syndrome	0.45	1
Amyotrophic Lateral Sclerosis	0.64	1
Parkinson's	0.96	1
Epilepsy	0.94	0
Social Anxiety Disorder	0.87	0.81
Panic Disorder	1	1
Mean Average Precision	**0.84**	**0.72**

Table 2.11 shows that for seven out of 10 conditions, the average precision is above 0.87, which means that for these seven conditions, the ranked list of treatments generated by Social InfoButtons reflects the officially reported treatments well. For symptoms, the results of Social InfoButtons and of the authoritative source also correlate well (six out of 10 are above 0.81), except for two conditions, multiple sclerosis and epilepsy. However, the added value of Social InfoButtons is where it differs from the authoritative source, in effect proposing a second opinion to the human expert for consideration. Since some patients are using these treatments, attention should be paid to them. For example, for the two conditions multiple sclerosis and epilepsy, which both have low average precision scores in "symptoms," the ranked lists are shown in Table 2.12. For these two conditions, none of the symptoms reported by the patients appears exactly in the authoritative source.

Another example of a treatment not reported by the authoritative source is the use of Aripiprazole for treating major depressive disorder, as shown in Table 2.9. Aripiprazole appears in Social InfoButtons, because 22 patients are using it, but it does not appear in the Mayo Clinic website. However, according to Nelson et al. [64], Aripiprazole has shown efficacy as an augmentation option with standard antidepressants and due to its efficacy and safety, it was approved by the FDA as a valid treatment. Another example is Cyclobenzaprine for treating Fibromyalgia, as shown in Table 2.8. Cyclobenzaprine does not appear in the Mayo Clinic's Web page about treatments of Fibromyalgia, however, according to Tofferi et al. [41], Cyclobenzaprine-treated patients were three times as likely to report an overall improvement and moderate reductions in

individual symptoms. These reports can make doctors aware of current trends in treatments.

Table 2.12 Treatments and Symptoms of Multiple Sclerosis and Epilepsy in Social InfoButtons (SI) and in Authoritative Source

Condition	Symptom in SI	Symptom in Authoritative Source
Multiple Sclerosis	Stiffness/Spasticity	Numbness or weakness in limbs
	Brain fog	Optic neuritis
	Excessive daytime sleepiness	Double vision or blurring of vision
	Mood swings	Tingling or pain in parts of your body
	Bladder problems	Electric-shock sensations
	Emotional lability	Tremor, lack of coordination
	Sexual dysfunction	Slurred speech
	Bowel problems	Fatigue
Epilepsy	Memory problems	Temporary confusion
	Problems concerntrating	A staring spell
	Excessive daytime sleepiness	Uncontrollable jerking movements of the arms and legs
	Headaches	Loss of consciousness or awareness

Another added value of Social InfoButtons is that it can provide doctors with information of how patients are using different treatments and how patients are experiencing symptoms. In the authoritative source, the treatments and symptoms are either included as part of text or in lists, but without detailed information based on real experience reports of patients.

2.7 Chapter Summary

In this chapter, we have presented a framework for enabling the use of semantics in the analysis of social health data. The framework enables flexible collection of data from a

variety of sources. Collected data is reconciled in a unified data model focusing on medical conditions and treatments and linked to create a knowledge base that enables cross-dataset exploration and analysis. The knowledge base can be furthermore extended by defining inference rules for automatic reasoning. Analytics have been developed and provided to end users via the Social InfoButtons Web application.

With Social InfoButtons, patients can retrieve knowledge about how other patients are coping with the same condition that they are suffering from. Government officials can compare the demographics of patients on social network sites with data in official data sources to investigate potential errors or biases in existing disease surveillance systems. We compared ranked lists of treatments and symptoms generated by the top ten conditions from Social InfoButtons against those posted by authoritative source. The results show a good correlation between Social InfoButtons and authoritative source, in which the mean average precision for treatments is 0.84 and the mean average precision for symptoms is 0.72. At the same time, Social InfoButtons also returns treatments and symptoms that are not shown on the authoritative website but are often reported by patients and have been studied by some medical researchers. Case studies on two treatments Aripiprazole and Cyclobenzaprine are carried out to validate this claim.

CHAPTER 3

EPIDEMIC OUTBREAK AND SPREAD DETECTION SYSTEM BASED ON TWITTER DATA

3.1 Introduction

Monitoring threats to public health is important for the healthcare community. The Internet has created unprecedented resources for tracking such threats. A previous approach by Ginsberg to this problem relied exclusively on search engines, in which users could input queries in reference to issues they were most concerned about [65]. In their thread of research, such queries were recorded by a search engine provider, leading to the realization that an aggregation of large numbers of queries might show patterns that are useful for the early detection of disease outbreaks. Ginsberg used Google's search query data (mostly keywords and phrases) to generate an early detection system, which could report outbreaks of influenza roughly two weeks prior to the official report of influenza. The official report is based on the number of patient visits to local hospitals and published by the Centers for Disease Control and Prevention (CDC).

However, the research on the detection of epidemics based on search queries is limited by two factors: First, user input query terms are regarded by search engine corporations as their core assets and are not available to outside researchers. Second, user locations are not explicitly recorded in search. As the users enter keywords into the search engine, the queries and IP addresses are recorded. However, the IP addresses, which can be converted to actual user locations, are not easily accessible to outsiders; thus, it is difficult to develop applications which use the actual geographic locations of users.

Twitter, a micro-blog service provider, is showing great potential for overcoming the limitations stated above. There are more than 340 million tweets posted by Twitter users per day [66]. What appears to be more important for researchers, however, is that most of the tweets are public. Moreover, the Twitter streaming API [67] enables researchers to retrieve everything contained in a tweet, including the people mentioned, the URL, and the topic tag added by Twitter users. The users' geographic information is available in the form of physical addresses specified in user profiles.

When analyzing a random sample of 500 Twitter users and their geographic locations, 30% were left blank, and 10% were spam addresses, like "in the universe" or "right behind you." The other 60% were valid addresses; however, they were distributed over different levels of granularity. For example, 78.3% of the valid addresses were "places," such as "NYC"; 12.5% were "states"; and 7.5% were "countries." Considering the above complexity of geographic names, if they are not properly processed, the subsequent estimation based on the addresses could easily lead to imprecise results. For example, in the data recorded on October 5th, 2011 by the influenza system "INFLUkun" [14], out of the 1,931 tweets, there were a total of 891 tweets whose locations were unrecognized. Unless the data with uncertain locations are interpreted correctly, there is the potential that the system could return misleading results. To address this issue, the Epidemics Outbreak and Spread Detection System (EOSDS) integrated a module to preprocess noisy geographic names. It applies a frequency-based delete list to identify and filter out non-informative geographic information, and it has the ability to detect different granularity levels of geographic names.

In EOSDS system, we have provided four location-based visual analytics to not only detect but also recognize spread of the disease over time. The analytics include Instance Map, Distribution Map, Filter Map, and Sentiment Trend. The *Instance Map* is used to show the tweets based on each "single" user's location. In the *Distribution Map*, absolute and relative frequencies of the distribution are displayed. The relative frequency is the absolute frequency divided by the population of each state. The Distribution Map enables the detection of which states house most Twitter users tweeting about an epidemic. The *Filter Map* gives users the flexibility to monitor the spread of epidemics based on time series and users' influence with a (minimum, maximum) range of followers to only display Twitter users in this range. Monitoring population behavior at different levels of granularity is also possible in filter map mode, as the lower level granularities such as "place" will often indicate more precise estimates of actual locations than higher-level granularities such as "country." The *Sentiment Trend* measures the public health concerns both in temporal and space dimensions. The visual analysis results of different maps are shown to detect the disease outbreaks and correlate well with the official CDC reports. In addition, the Distribution Map made it possible to discover an unusual listeria outbreak situation in Wyoming, which was not reported by the CDC until 7 days later.

3.2 Related Work

In this section, we summarize the previous research that utilized online social media data to monitor diseases and emergencies. Since the year 2008, concepts and systems have been developed to monitor disease outbreaks and emergencies with Twitter. Artman, et al. [68] introduced the concept of dialogical emergency management, which emphasizes the

screening of vast and quickly spread information on the Internet, to help the emergency management staff gain a better strategic awareness of the public. The Alert4All Screening of New Media (SNM) tool [69] was developed based on this concept to analyze emotion recognition/affect in social media, e.g. Twitter and Facebook, regarding crisis management. Brownstein, et al. [70] used online News to perform surveillance of epidemics. Their system, Healthmap, collects reports from online News aggregators, such as Google News. By categorizing the News into epidemics-related and unrelated reports, and filtering the epidemics-related documents into "breaking News," "warnings," and "old News," the system is able to trigger alerts based on "breaking News." With regards to location processing on Twitter, a study by Cheng et al. [71] determines users' positions when location information is absent. Their location estimator can place 51% of the Twitter users within 100 miles of their actual locations. Their approach relies on detecting "local" words, which are of a high local specificity and a fast dispersion, such as "howdy" in Texas.

The other thread of research focused on building models, primarily supervised learning models, to detect disease and emergency events from Twitter. Collier and Doan [13] developed a model to automatically classify Twitter messages into six fixed classes of syndromes, such as Respiratory and Gastrointestinal. Aramaki, et al. [14] applied a Support Vector Machine (SVM) to distinguish influenza-related tweets from tweets that are irrelevant. Signorini, et al. [15] used an SVM-based estimator to analyze H1N1-related tweets, and estimated the Influenza-like Illness (ILI) rate, which is usually regarded as the ground truth, preceding the official announcement of an H1N1 outbreak by one to two weeks. Similarly, Culotta [16] experimented with a number of regression

models to correlate Twitter messages with statistics from the Center for Disease Control and Prevention (CDC) and provided a relatively simple method to track the ILI rate using a large number of Twitter messages [72]. Lampos and Cristianini [17] used an approach to automatically learn a set of markers to help compute flu scores, and achieved a high correlation with the HPA flu score, which is the equivalent of the CDC score in the UK.

3.3 Epidemics Outbreak and Spread Detection System

3.3.1 Data Collection

To better display the geographic locations of outbreaks and the spread of epidemics at multiple levels, the Epidemics Outbreak and Spread Detection System (EOSDS) was developed. Its architecture is shown in Figure 3.1. We implemented a data collector using the Twitter API version 1.1 and Twitter4J library [73] to collect real-time tweets containing certain specified health-related keywords (e.g., listeria), along with associated user profile information for subsequent analysis. The overall data collection process can be described as "ETL" (Extract-Transform-Load) approach, as it is widely used in Data Warehousing. The data was collected in JSON format from the Twitter Streaming API. (This is the "Extract" step). Then the raw JSON data was parsed into relational data, such as tweets, tweet_mentions, tweet_place, tweet_tags, tweet_urls, and users (Transform step). Finally, the relational data was stored into our MySQL relational database (Load step).

The current prototype system has collected a total of 11.7+ million tweets in 14 datasets. These datasets include seven infectious diseases: Listeria, influenza, swine flu, measles, meningitis, tuberculosis, and ebola; four mental health problems: Major depression, generalized anxiety disorder, obsessive-compulsive disorder, and bipolar

disorder; two crises: Air disaster and natural disaster; and one clinical science issue: Melanoma experimental drug. The core component uses the Twitter Streaming API for collecting epidemics-related real-time tweets. For each tweet type, the tweets were collected according to the keywords of the dataset. These keywords extended the condition terms defined by U.S. Department of Health and Human Services [74], and are shown in the Appendix A. The Twitter4J library automatically identifies the language of tweets during the data collection phase. For example, if the value of the tweet attribute "lang" is "en," that means this tweet is an English tweet. If the value of the tweet attribute is "fr," it means that this tweet is a French tweet. The statistics of the collected tweets is shown in Table 3.1.

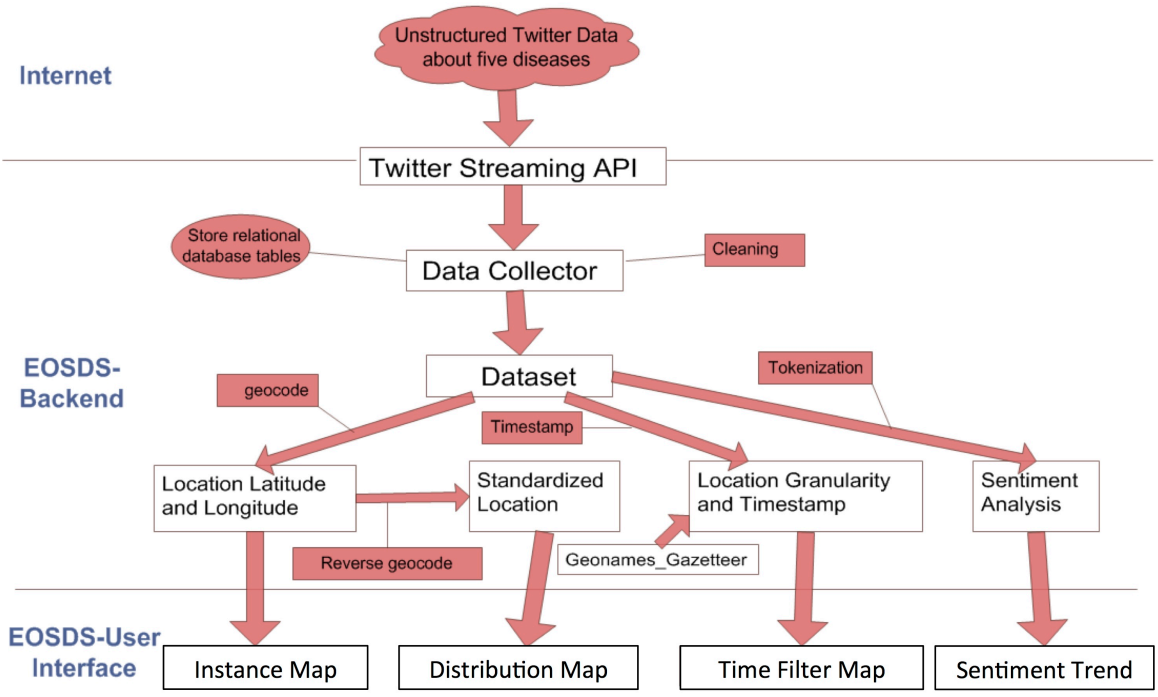


Figure 3.1 Architecture of Epidemics Outbreak and Spread Detection system.

Table 3.1 The Statistics of The Collected Dataset (Up to 03/23/2015)

Dataset Id	Tweet Type	Total number of Tweets
1	Listeria	43,646
2	Influenza	2,231,442
3	Swine Flu	121,208
4	Measles	276,282
5	Meningitis	189,886
6	Tuberculosis	245,639
7	Major Depression	3,209,413
8	Generalized Anxiety Disorder	386,262
9	Obsessive-compulsive Disorder	571,867
10	Bipolar Disorder	181,942
11	Air Disaster	22,946
12	Melanoma Experimental Drug	145,357
13	Natural Disaster	1,746,899
14	Ebola	2,385,275

3.3.2 Location Processing

The locations recorded in the Twitter profiles of "tweeters" are filtered using location stop words. A list of 100 meaningless locations, such as "in the universe" and "wherever you are" were manually selected. Then these locations were fed to the concept list generator module in Automap [75]. Automap is a text mining and network text analysis tool. Its concept list functionality takes a text file as input, and outputs each concept with its number of occurrences (frequency) in the input file.

A list of "unigrams" (concepts which contain exactly one word) is generated in descending order of term frequency that most likely occurs in meaningless geographic locations. Table 3.2 shows the top five unigrams. The list is adopted by EOSDS as the delete list. The delete list was applied to a test dataset, which contained 1000 records that were posted by Twitter users. The results are shown in Table 3.3. Of these 1000, 354 locations were categorized as spam by EOSDS. According to a manual check, 362

records were actually spam locations. Thus, EOSDS achieved a precision of 97.1%, and a recall of 95.8% in identifying meaningless locations.

Table 3.2 Top Five Unigrams in Meaningless Locations

Concept	Frequency	Relative Frequency	Gram Type
the	19	1	unigram
In	17	0.89	unigram
in	13	0.68	unigram
The	8	0.42	unigram
you	8	0.42	unigram

Table 3.3 Results of Identifying Spam Addresses. (Detect+ are locations that were identified as spam addresses. S+ are locations that are in fact spam addresses)

	S− (not spam)	S+ (spam)	Total
Detect+	7	347	354
Detect−	631	15	646
Total	638	362	1000

In addition, it is desirable to infer the state or country information from a text-based location, but the users' profile locations, even after data cleaning, are at different levels of granularity. The granularity of locations creates a difficulty to identify what state or city a tweet comes from. The different levels of granularity are shown in Table 3.4. EOSDS solves this problem by a method called "two-step coding."

Table 3.4 Different Levels of Granularity

Granularity	Example
Place	Newark, New Jersey
State	Colorado
Country	Netherlands
World	heaven

The method of two-step geocoding works as follows. First, all the locations are geocoded into latitudes and longitudes, then the obtained latitudes and longitudes are reversely geocoded into standardized addresses. For example, at the place level, some user-specified locations are “Rochester,” “Rochester, USA,” “NYC.” At the state level, some locations are “NY,” or “New York.” In the geocoding step, “Rochester” and “Rochester, USA” are translated into latitudes and longitudes indicating the downtown area of the city of Rochester in the state of New York. Note that the Google Geocoding API returns address results depending on the region from which the request is sent [76]. The search for “Rochester” and “Rochester, USA” returns city of Rochester in New York State instead of other cities (e.g., Rochester in Minnesota), since the search is sent from New Jersey. Specifying the “region” parameter in a Google Geocoding API call can change its result bias if needed [76]. Similarly, “NYC,” “NY,” and “New York” are geocoded into latitude and longitude pointing to downtown Manhattan (even if the user is located in Brooklyn).

We then apply reverse geocoding, converting latitude and longitude into physical addresses. We retrieve these addresses in the format “county, state, country” or “state, country.” Thus, after two-step coding, “Rochester” and “Rochester, USA” become identical, standardized addresses: “Monroe, New York, USA.” (Rochester is in Monroe County) “NYC,” “NY,” and “New York” become standardized into “New York, USA.” With standardized addresses, the system knows how many tweets (absolute frequency) are from each state. The whole process is illustrated in the Figure 3.2.

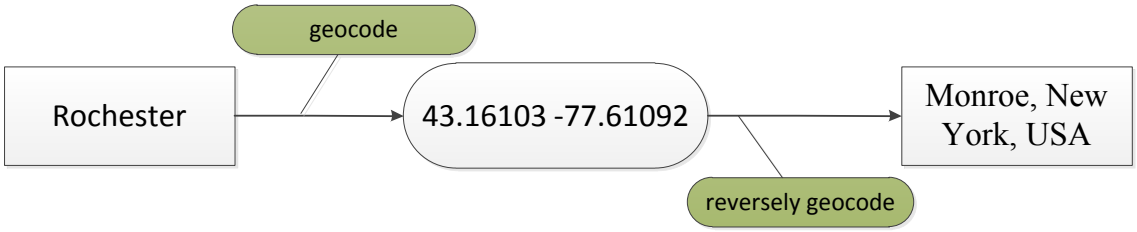


Figure 3.2 The process of two-step geocoding.

3.3.3 Visual Analytics

3.3.3.1 Instance Map. The Instance Map display mode provides a direct way to display locations of all tweet instances. After the preprocessing, only records containing valid geographic information are left. Before mapping the geographic information to the actual map, EOSDS geocodes the geographic information into (latitude, longitude) coordinates that can be processed by the system. EOSDS' geocoding is done by the Google Map API [25]. Every location is passed to the geocoding server, and the returned latitude and longitude are mapped by EOSDS to show the estimated location of each tweet.

It is assumed that the location information specified in a user profile is the location where the user actually posted the tweet, probably the place where s/he lives or works. In the case of tweets posted by mobile devices like smart phones, the step of geocoding is skipped and the mobile devices' location at the time of tweeting is utilized as the user's actual location (This location is also recorded in the EOSDS database.) An example of instance map is shown in Figure 3.3, where each red marker is an individual tweet, each blue circle is a cluster of less than 10 tweets (e.g., there are 9 tweets in Florida), and each yellow circle is a cluster of between 10 and 100 tweets.

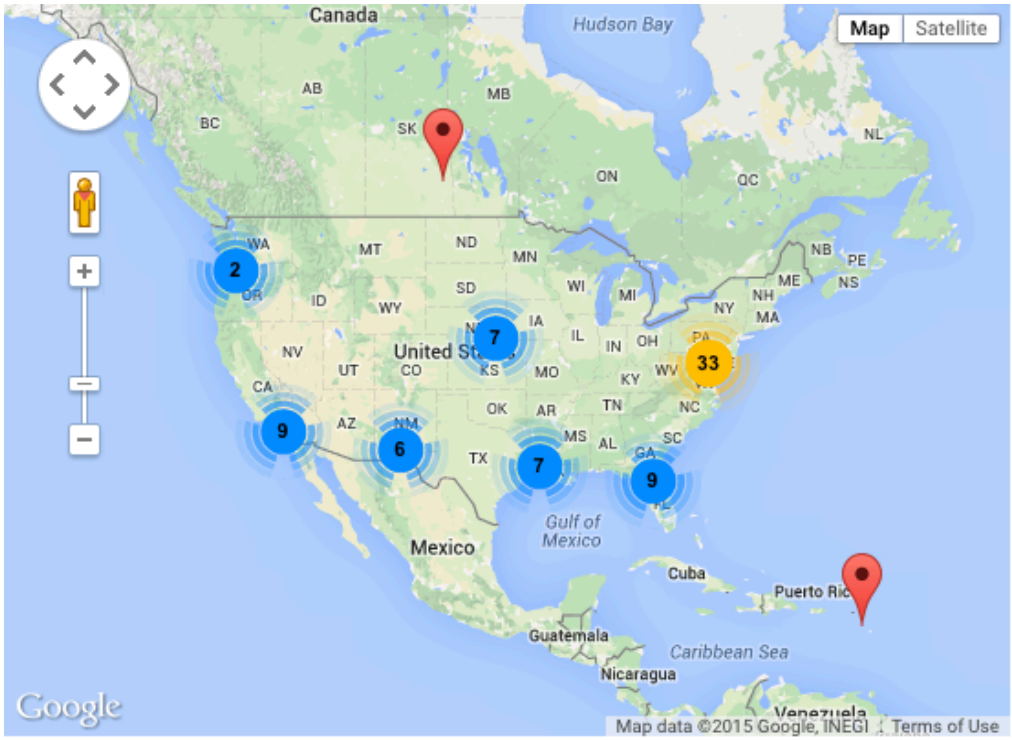
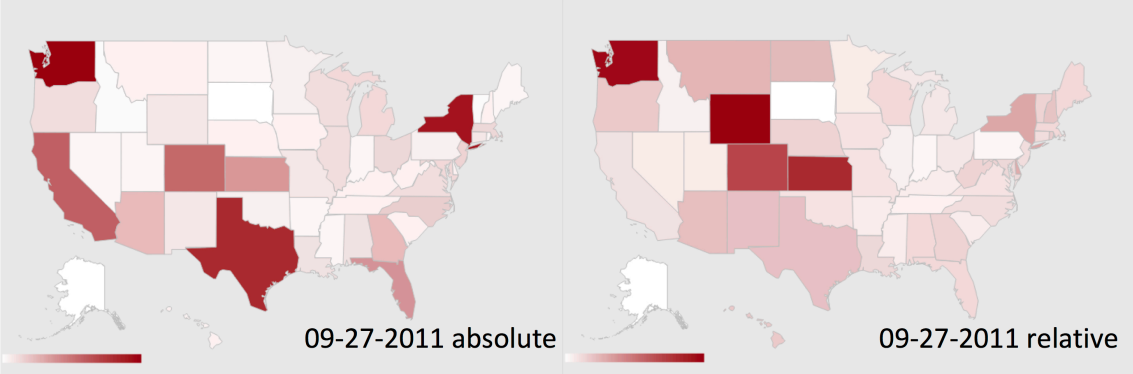


Figure 3.3 An example of an instance map.

The Instance Map provides a straightforward way to monitor the tweet users' concerns about an epidemic. However, in order to gain "the big picture" in terms of the geographic distribution of disease-related tweets and of the spread of the public's concern about epidemics, the Distribution Map and filter map were implemented and will be shown below.

3.3.3.2 Distribution Map. The idea of building a distribution map is based on a problem with the instance map. In the instance map, there are many markers representing individual users who are posting tweets. It is possible to recognize wherever there is an unusual cluster of "markers," which should be investigated by public health officials. In the instance map, it is not always easy to judge whether a particular area is

unusual, because different people may have different criteria for tweeting. Another limitation with the instance map is that in some states, such as California, there are more tweets because there are more people than in other states. But is there always an epidemic in these areas? Thus, the display needs to incorporate both absolute frequency and relative frequency. The relative frequency of each state is calculated as the absolute frequency of each state divided by the population of the state (normalized by a factor of 1,000,000). Thus, sparsely populated states gain a larger weight than densely populated states. This method makes the epidemics trends easier to monitor. An example with an absolute map and a relative distribution map is shown in Figure 3.4.



(a) (b)

Figure 3.4 (a) Absolute distribution map on 09-27-2011. (b) Relative distribution map on 09-27-2011.

3.3.3.3 Filter Map. The filter map provides users with a dynamic interface to monitor and analyze dynamic trends derivable from health-related tweets. Three filters are incorporated into the filter map: granularity filter, influence filter and timeline filter.

Granularity Filter

Different levels of location granularity represent different precision levels. For example, "Newark, NJ" is more precise than "New Jersey." "New Jersey" is more precise than "United States." To match a certain location with a level of granularity, we make use of a gazetteer. A gazetteer is a geographical dictionary or directory, an important source of data about places and place names, used in conjunction with maps. To label locations with the correct level of granularity, the "National Places Gazetteer" [77], issued by the U.S. Census Bureau, was used.

The "National Places Gazetteer" contains more than 29,000 US places, including cities, towns, boroughs, and Census-designated places (CDP). These names represent the lowest level of granularity. Names of the 50 states and names of 245 countries worldwide are also used together with the National Places Gazetteer. Each location was checked against this extended gazetteer. There are cases of multiple records in the gazetteer matching a single location in the dataset, but those multiple matching records almost always are at a single level of granularity. For example, there are a few locations in the US called "London," but all belong to a single granularity level: place. It is very rare that two levels of granularity share a name; that means that it is possible to achieve a high precision in matching locations with levels of granularity. In addition, in case that there are granularity keywords in location names, such as "State of Washington" (keyword: state), "Washington DC" (keyword: DC), and "Washington County" (keyword: county), these location names' granularities are automatically identified according to its keyword. In this case, "State of Washington" and "Washington DC" are coded as state level and "Washington County" is coded as county level. Filtering with different levels of

geographic locations allows making maximal use of the information available. If EOSDS users choose a map only showing the place-level locations, the displayed map positions provide a find-grained display for investigating the locations of the epidemic.

Influence Filter

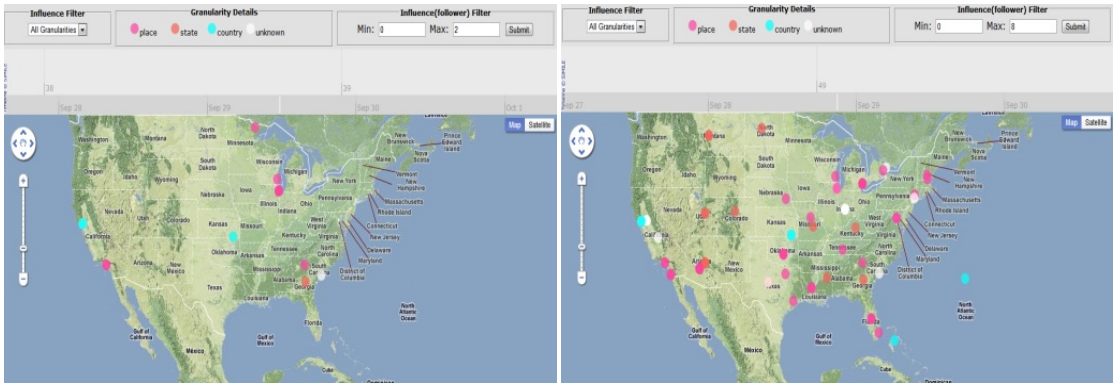
Not every tweeter has the same impact on his/her environment. A range of follower counts may be set by an EOSDS user to display only tweets from those Twitter users with a number of followers greater than the minimum and smaller than the maximum. This functionality is helpful to find how the "influencers" are distributed over the map. The effect of applying the influence filter is illustrated in Figure 3.5. It enables the users to concentrate only on the tweets that are highly influential. By tracking the distribution of influential tweets, it is possible to estimate where the "seed tweet" originated, and how these influential tweets affect the spread of public concern about a certain epidemic.

Timeline Filter

Besides the space dimension, considering that every tweet has a timestamp, tweets can also provide us with an additional perspective to gain insights into the temporal distribution and development of an epidemic. The timeline filter was built with SIMILE [78]. By moving the observed time point forward and backward, EOSDS users can easily find a particular time frame to recognize where and when a sudden increase of tweets occurs, and how this fits into the bigger picture of the epidemic.

3.3.3.4 Sentiment Trend. The Sentiment Trend analytics contain Concern Timeline Chart and Concern Map. Through sentiment analysis, the Concern Timeline

Chart is able to track the public concern trends on the timeline and the Concern Map shows the geographic distribution of concern. The details of the Sentiment Trend are introduced in details in Chapter 4.



(a) (b)

Figure 3.5 (a) Tweet users with influence range between 0 and 2. (b) Tweet users with influence range between 0 and 8.

3.4 Evaluation of EOSDS System

The potential of using the number of health-related tweets to predict the CDC reports is explored through the following experiment. A test dataset was collected with the keyword "listeria," from "2011-09-26 12:07:39" to "2011-09-28 15:57:27" and from "2011-10-09 19:38:09" to "2011-10-18 09:59:16," which was during a severe outbreak of listeria in the US. There were exactly 11,000 tweets, of which 2,410 were removed by our data cleaning process. The final dataset contains 8,590 tweets.

Although the distribution map in EOSDS has a low Pearson Correlation Coefficient (0.17) with the CDC report [29], the comparison of the distribution map with the CDC's report reveals some interesting observations, as shown in Figure 3.6. Each

state's number of cases in the CDC report is shown by the continuous blue line. Interestingly, there are two states, Washington and Wyoming, (circled in dashed-blue) showing a sharp conflict between EOSDS results and the CDC report. The reason is that on September 26th, a death was confirmed by the Health Department of Wyoming [79], but that death was not in the CDC report until October 6th [10]. This shows that EOSDS provided information faster than the CDC report.

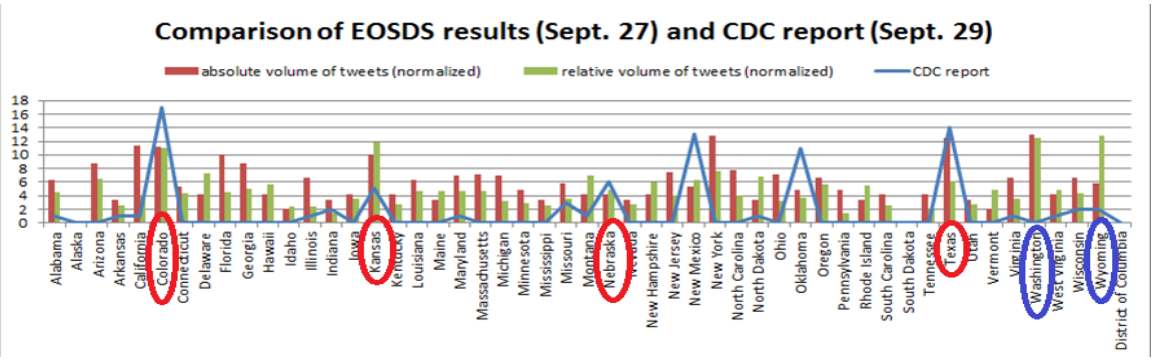


Figure 3.6 The comparison between EOSDS distribution map results (09/27/2011) and CDC report (09/29/2011).

3.5 Limitations of Current Approach

In this section we discuss the limitations on the current geographic processing and data collection method and suggest future working directions. For geographic processing, there are three types of locations in Twitter: (1) Tweet location (tweet location where the Twitter user is currently located) (2) User location profile (e.g., the place where he lives), and (3) Location in a tweet (e.g., I tweet about the financial crisis in Greece). For the disease outbreak detection task, we are interested in the first and three types of location, as they both provide critical information during disease outbreak and spread. We utilized the locations on users' profiles to infer their actual tweet location by filtering out noisy

locations and identifying the granularities of locations. However, many Twitter users do not tend to specify the location in their profiles and the location is not always correct unless it is geo-tagged with latitude and longitude. The location-related privacy may be the cause and it remains as a challenging task, but there are texts (with location-information) that can be used to indirectly identify the locations but we have not yet pursued the text analytics to identify these locations. Cheng et al. [71] proposed utilizing local words (e.g, "red sox" is local to "Boston") but the people move all the time, the local words may not be a strong indicator for the actual tweet location (e.g., a Bostonian tweets in New York). Kinsela et al. [80] used a language modeling approach to build models of locations by training the language model with geo-tagged tweets originating from those locations. In addition, the time-zone information in users' profile can be used as another perspective to inferring locations in higher granularity (e.g., state-level, country-level). We plan to extend the geographic processing by utilizing the above features (e.g, local words, language, and time zone) to further improve the location estimation of diseases in EOSDS system.

For the data collection, the current data collection is based on a few specified diseases and their keywords provided by public health agencies. However, it is desirable to automatically detect the usual diseases (e.g, MERS in Korea) and add it to the data collection. There is not much research on this topic. The limitations of Twitter may be the cause, as Twitter limits the number of data collection projects for a user and it requires separate authentication for initializing a new data collection pipeline.

3.6 Chapter Summary

This chapter outlined the geographic aspects that have not been paid enough attention to by current research regarding utilizing Twitter data to detect and predict the spread of epidemics. EOSDS was developed with modules to clean noisy geographic locations based on text mining methods, and to automatically identify the levels of granularity for different location specifications. Furthermore, EOSDS enables users to analyze the Twitter data from four different perspectives, which are Instance Map, Distribution Map, Filter Map, and Sentiment Trend. The advantages and limitations of each analytics were discussed. The limitations of current data collection and geographic location processing are also discussed.

In experiments, we compared the results of Instance Map and Distribution Map with CDC reports during listeria outbreak in September 2011. The Instance Map shows large clusters of tweets on the heavily affected states, such as Colorado and Texas. Among the six states with most tweets on Distribution Map, we observed that four of them correlated well with CDC reports. In addition, the Distribution Map made it possible to discover an unusual listeria outbreak situation in Wyoming, which was not reported by the CDC until 7 days later.

CHAPTER 4

TWITTER SENTIMENT CLASSIFICATION FOR MEASURING PUBLIC HEALTH CONCERNS

4.1 Introduction

Public health surveillance is critical to monitoring the spread of infectious diseases and deploying rapid responses when there is an indication of an epidemic emerging. Different surveillance strategies have been developed to meet different needs. These strategies include sentinel surveillance systems, household surveys, laboratory-based surveillance, and most recently IDSR (Integrated Disease Surveillance and Response) [18]. Besides monitoring the spread of a disease itself, monitoring *emotional changes* of the general public, brought about by epidemics, is becoming increasingly important for public health specialists.

The importance of monitoring the public's concerns about an epidemic is illustrated by the recent Ebola scare in the United States. Since the end of September 2014, Ebola concerns have spread in the United States after a Liberian visitor to Dallas became the first person to be diagnosed in the US. The immigration examination and the medical system's ability to deal with Ebola were widely questioned by the general public [81] due to a series of missteps when the Liberian was issued a visitor visa and was not diagnosed by a Dallas hospital. For example, a tweet on October 15th of 2014, stated that, "*Co-worker LEGITIMATELY thinks #Ebola was caused by one of two things: 1.) Gov't attempts at population control. 2.) ISIS THIS IS NOT A JOKE.*" As the public opinion will potentially affect the government's public health decisions, President Obama attempted to

calm the public by stating that "This is a serious disease, but we can't give in to hysteria or fear." [81]

Zhu, et al. [82] studied the changes in mental state of the Chinese public during the outbreak of SARS (2003). They found that, during the outbreak, most of the people surveyed (96.4%) reported emotional changes and negative emotions such as panic (54.8%), nervousness (34.0%), and fear (7.6%). Psychological changes might lead to unpredictable behavior. Of all the people surveyed, 23.3% admitted to “irrational” behaviors such as going on a shopping spree, or to actions such as seeking shelter, preparing provisions, etc.

Another example is the public’s reaction to Japan’s nuclear emergency in March 2011 [83]. Text messages about nuclear plumes spread throughout Asia. In China, the rumors that iodized salt could help ward off radiation poisoning amid Japan’s nuclear emergency triggered panic buying all-over the country. In Vietnam, students were kept indoors by schools; some companies allowed staff to leave early to avoid rainfall after the rumor spread that rain would burn the skin and cause cancer. A university in Manila cancelled classes due to a similar scare.

As the above examples illustrate, monitoring public panic about health issues is critical not only to public health specialists but also to government decision makers. However, for traditional public health surveillance systems, it is hard to detect and monitor health related concerns and changes in public attitudes to health-related issues. Due to their expenses, the existing surveillance methods, such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays. To supplement the current surveillance systems, a novel tool must be

developed. This tool must be able to track real-time statistics of emotions related to different health matters, such as epidemics, to provide early warning, and to help the government decision makers prevent or respond to potential social crises that might be the impact of these health-related emergencies.

We explored the potential of mining social network data, such as tweets, to provide a tool for public health specialists and government decision makers to gauge the *Measure of Concern* (MOC) expressed by Twitter users under the impact of diseases. To derive the MOC from Twitter, we developed a two-step classification approach to analyze sentiments in disease-related tweets. We first distinguish Personal from News (Non-Personal) tweets. Many news articles released by online media organizations are used for 're-tweets' by Twitter users. We consider these News tweets as Non-Personal, as opposed to Personal tweets posted by individual Twitter users. We refer to the former as *News tweets* and the latter as *Personal tweets*. In the second stage, the sentiment analysis is applied only to Personal tweets to distinguish Negative from Non-Negative tweets.

Although News tweets may also express concerns about a certain disease, they tend not to reflect the direct emotional impact of that disease on people. A person re-tweeting a News message about a disease, which is comparable to forwarding an email message, is most likely not directly affected by it, while a user sending out a Personal tweet with emotional expressions might be directly affected. Note that the two-step sentiment classification problem we present is different from the traditional Twitter sentiment classification, which is categorizing tweets into positive/negative or positive/neutral/negative tweets [14, 84-87] without distinguishing Personal from Non-Personal tweets first. Our sentiment classification method is able to identify Personal

tweets (including Personal Negative and Personal Non-Negative) and News (Non-Personal) tweets. In addition, we subsequently use the results of the classification to compute the correlation between sentiment-carrying tweets and News tweets, as the classification results provide all the necessary data for this computation.

We need to differentiate between the spread of concern about a disease and the spread of the disease itself. For example, the tweet: "*Wiz looks like he got the measles and Ross just dark as hell. I can't tell if they're tattoos or wrinkles http://twitpic.com/4geuc2*" is annotated as a Non-Negative tweet, because it shows no concern. However, it is a strong clue to track the spread of measles. We focus on studying the Twitter users' concerns about diseases instead of the outbreak of the disease itself, which has been extensively studied [13-15, 17, 70].

Using the sentiment classification results, we quantify the Measure of Concern (MOC) based on the number of Personal Negative tweets per day. The MOC increases with the relative growth of Personal Negative tweets and with the absolute growth of Personal Negative tweets. Previous research [4, 88] found that sentiment surges co-occurred with health events on a timeline. Different from the previous work, we calculated the correlation between MOC timeline (i.e., change over time) and News timeline and the correlation between Non-Negative timeline and News timeline using the Jaccard Coefficient [89]. Using the MOC to track public health concerns can help government officials to make timely decisions to refute rumors, and thus prevent potential social crises such as the past case of Chinese panic buying of salt. Monitoring of the public concern using social network data can provide public health specialists with a

surveillance capability for large segments of the population, in real-time, and with low expenses.

The rest of this chapter is organized as follows. In Section 4.2, related work and open problems are discussed. In Section 4.3, we give formal definitions of the concepts used in this chapter. In Section 4.4, sentiment classification methods and results are introduced in detail. In Section 4.5, the sentiment timeline trend analysis results are illustrated, interpreted, and discussed. Section 4.6 contains the chapter summary.

4.2 Related Work

4.2.1 Sentiment Analysis

Sentiment Analysis has been an active research area since the 2000s. With an increasing number of datasets from various data sources, such as blogs, review sites, News articles, and micro-blogs available, researchers have become interested in mining high-level sentiments from them. Sentiments are also closely related to information spread. Their relationship was shown in different contexts, such as social transmission [90], News broadcasts [91], and online social media, such as Twitter [92]. By analyzing the sentiments of opinion leaders, the public health officials will be able to monitor the viral effects in social media communication, and take early actions to prevent unnecessary panic.

A survey of sentiment analysis was done by Pang and Lee [93]. Research on sentiment analysis can be categorized into the following levels: document-level [94], blog-level [95], sentence-level [96], tweet-level [69, 97, 98] with the sub-category non-English tweet level [99], and tweet-entity-level [100]. Due to the large number of

available tweets and their real-time nature, tweets are ideal for sentiment classification and quantification for disease monitoring, and more broadly, for crisis monitoring.

4.2.2 Twitter Sentiment Classification

Extensive research has been performed in the sub-area of Twitter sentiment classification since 2009 [86, 97, 101-105]. Most of this thread of research used Machine Learning-based approaches such as Naïve Bayes, Multinomial Naïve Bayes, and Support Vector Machine. The Naïve Bayes classifier is a derivative of the Bayes decision rule [106], and it assumes that all features are independent from each other. Good performance of Naïve Bayes (NB) was reported in several sentiment analysis papers [97, 101, 105]. Multinomial Naive Bayes (MNB) is a model that works well on sentiment classification [102, 103, 105]. MNB takes into account the number of occurrences and relative frequency of each word. Support Vector Machine [107] is also a popular ML-based classification method that works well on tweets [97, 104]. In Natural Language Processing, SVM with a Polynomial Kernel is more popular [108].

Mohammad, et al. [86] explored an extensive list of features such as clusters, negation, and n-grams, and used a Support Vector Machine (SVM) to classify Twitter messages into positive, negative, and neutral. Barbosa and Feng [101] focused on automation of the training data generation process. Their work combined sentiment-labeled tweets coming from three sources: Twendz, Twitter Sentiment, and Tweet Feel. A moderate Cohen's Kappa Coefficient served as evidence that the combination of several sources reduced the bias of the individual sources. In this way, the combination improved the polarity classification.

The above sentiment classification studies have two drawbacks. Firstly, they classified Twitter messages into either positive/negative or positive/negative/neutral with the assumption that all Twitter messages express ones' opinion. However, this assumption does not hold in many situations, especially when the tweets are about epidemics or more broadly, about crises. In these situations, as we found when we randomly sampled 100 tweets, many tweets (up to 30%) of the samples, are repetitions of the News without any personal opinion. Since they are not explicitly labeled with re-tweet symbols, it is not easy for a stop-word based pre-processing filter to detect them. We attempt to solve a different problem, which is how to classify tweets into three categories: Personal Negative tweets, Personal Non-Negative tweets, and News tweets (tweets that are non-Personal tweets). We are not singling out positive tweets, as few people would post positive tweets about a spreading epidemic. Instead of identifying News tweets, Brynielsson, et al. [97] used manual labeling to classify tweets into "angry," "fear," "positive," or "other" (irrelevant). Salathe and Khandelwal [109] identified irrelevant tweets together with sentiment classifications. Without considering irrelevant tweets, they calculated the H1N1 vaccine sentiment score from the relative difference of positive and negative messages. As we will show later, by the two-step classification method, we can automatically extract News tweets and perform the sentiment analysis on the remaining tweets. Then the results of sentiment classification are used as input for computing the correlation between sentiments and News trends. In this way, the goals of sentiment classification and measuring the public concern can be achieved in an integrated framework.

Secondly, the above research approaches have developed sophisticated models to improve the precision and recall of sentiment classification, but they did not quantify the results of the sentiment classification to measure timeline trends, and correlate them with real-world incidents, to provide insights for public health specialists and government decision makers. We developed the *Measure of Concern* (MOC) to quantify the sentiments, and we correlate sentiment trends and News trends to provide better knowledge of Twitter users' reactions towards crises, such as epidemics, mental health problems, clinical science problems, etc.

4.2.3 Quantifying Twitter Sentiment on Timeline

The objective of sentiment quantification is to convert natural language text to a numerical value or a timeline of numerical values to gain insights into the sentiment trends. Zhuang, et al. [110] generated a quantification of sentiments about movie elements, such as special effects, plot, dialogue, etc. Their quantification contains a positive score and a negative score towards each specific movie element.

For tweet-level sentiment quantification on a timeline, Chew and Eysenbach [111] used a statistical approach to computing the relative proportion of all tweets expressing concerns about H1N1 and visualized the temporal trend of positive/negative sentiments based on their proportion. Similar research was done by O'Connor, et al. [112]. In their thread of research, they quantified the sentiments as a timeline by deriving a day-to-day (positive and negative) sentiment score simply by counting how many positive and negative words of one tweet appear in the subjectivity lexicon of OpinionFinder [96], which is a list containing words marked as positive or negative. By analyzing Chinese

micro-blogs, Sha, et al. [88] found that the sentiment fluctuations on a timeline were associated with the announcements of new regulations or government actions.

There are two drawbacks of the existing Twitter sentiment quantification research: (1) The clue-based sentiment extraction models used by the above researchers are often too limited. As pointed out by Wiebe and Riloff [113], identifying positive or negative tweets by counting words in a dictionary usually has high precision but low recall. In the case of Twitter sentiment analysis, the performance will be even worse, since many words in tweets are not recorded in a dictionary. For example, LMAO (Laughing My A** Off), is a positive "word" in Twitter, but it does not match any word in MPQA [114], which is a popular sentiment dictionary. (2) The correlation between sentiments and News events are only studied visually by observing their co-occurrence on a timeline [88, 112], but to the best of our knowledge, there is no prior work that both quantitatively and qualitatively studies these correlations between Twitter sentiment and the News in Twitter to identify concerns caused by diseases and crises.

As we summarized the Twitter sentiment classification and Twitter sentiment quantification research, there is a research gap between them. More specifically, the existing sentiment classification research does not quantify sentiment timeline trends from the classification results to provide insights into the sentiments. On the other hand, the existing sentiment quantification research often used a clue-based model, which has a low recall in terms of identifying sentiment tweets. In addition, the existing sentiment quantification work has only qualitatively correlated the sentiment timeline with real-world events, but has not provided a comprehensive, quantitative correlation

between the sentiment timeline trend and the News timeline trend. This work is our attempt to fill this gap.

There are two objectives to achieve. The first objective is to automatically label datasets for training a Twitter sentiment classifier for identifying News (Non-Personal) tweets. The purpose of identifying News tweets is that after filtering them out in the first step the Negative vs. Non-Negative classifier can be applied to the Personal tweets only. The second objective is to quantify the sentiment trends and News timeline trends from sentiment classification results and compute a quantitative measure of correlation between them, to better understand the sentiment timeline trends relative to events in the real world.

4.3 Definitions

Definition 4.1 (Personal Tweet): A *Personal Tweet* is defined to be one that expresses its author's private states [96, 115]. A private state can be a sentiment, opinion, speculation, emotion, or evaluation, and it cannot be verified by objective observation. In addition, if a tweet talks about a fact observed by the Twitter user, it is also defined as a Personal Tweet. The purpose of this definition is to distinguish the tweets written word-by-word by the Twitter users from the News tweets redistributed in the Twitter environment, as mentioned above.

Example (Personal Tweet)
"*The boyfriend is STILL sick from the @fatburger he ate last Thursday. The doctor suspects listeria. :(*"

Definition 4.2 (News Tweet): A *News Tweet* (denoted with NT) is a tweet that is not a Personal Tweet. A News Tweet states an objective fact.

Example (News Tweet):
"*Measles outbreak reported in Honiara, Solomon Islands | Outbreak News Today* http://fb.me/1hMxpNmrh"

Definition 4.3 (Personal Negative Tweet and Personal Non-Negative Tweet): If a tweet is a Personal Tweet, and it expresses negative emotions or attitude, it is a *Personal Negative Tweet* (denoted as PN). Otherwise, it is a *Personal Non-Negative Tweet* (denoted as PNN). Personal Non-Negative Tweets include personal neutral or personal positive tweets. A Personal Tweet is either a PN or a PNN.

Definition 4.4a (Measure of Concern): *Measure of Concern (MOC)* M_i is the square of the total number of Personal Negative tweets that are posted at time i, divided by the total number of Raw Tweets of a particular type at the same time i. The Measure of Concern increases with the relative growth of Personal Negative tweets and with the absolute growth of Personal Negative tweets.

Definition 4.4b (Non-Negative Sentiment): Similarly, the *Non-Negative Sentiment* NN_i is the square of the total number of Personal Non-Negative tweets that are posted at time i, divided by the total number of raw tweets of a particular type at the same time i.

Definition 4.4c (News Count): Finally, the *News Count* NE_i is the total number of News Tweets at the time i. Note that the News Count is not normalized by the total number of

raw tweets. The reason is that we are interested in studying the relationship between sentiment trends and News popularity trends (see Section 4.5). An absolute News Count is able to better represent the popularity of News.

Definition 4.5 (Peak): Given a timeline of numerical values, a value X_i on the timeline is defined as a peak if and only if X_i is the largest value in a given time interval $[i\text{-}b, i+a]$. The time intervals $a>0, b>0$ can be chosen according to each specific case to limit the number of peaks. Peaks are defined for MOC timelines, Non-Negative timelines, and News Count timelines. Figure 4.5 in Section 4.5 will show the peaks as red or black dots on an MOC Timeline.

4.4 Two-Step Sentiment Classification

In this section, we present the two-step sentiment classification and quantification method. As discussed earlier, the goal in sentiment classification is different from the one of classic sentiment classification of Tweets. Many News tweets are re-tweeted in Twitter. Classifying the tweets into Personal and News tweets in the first step can help consider only Personal tweets in a sentiment analysis in the next step (Negative vs. Non-Negative classification). Since we are interested in studying the correlation between the timeline trend of sentiments and of News, the detection of News tweets needs to be seamlessly integrated. Thus, our approach of classifying a tweet into one of the three classes: Personal Negative, Personal Non-Negative (including neutral and positive), and News allow not only the classification but also correlation studies. An overview of our method is shown in Figure 4.1.

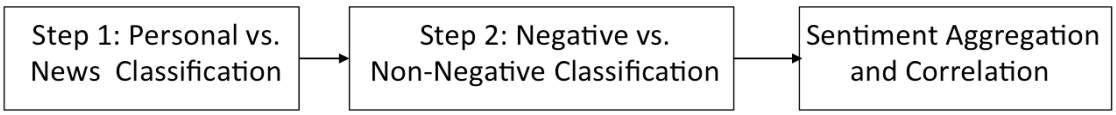


Figure 4.1 Overview of the two-step sentiment classification and quantification method.

Only English tweets, which were automatically detected during the data collection phase (see Table 4.5 for the data sets), are considered. As shown in Figure 4.1, the sentiment classification problem is approached in two steps. First, for all English tweets we separated Personal from News (Non-Personal) tweets. Second, after the Personal tweets were extracted by the most successful of the Personal/News Machine Learning classifier, these Personal tweets were used as input to another Machine Learning classifier, to identify Negative tweets. After News tweets, Personal Negative tweets, and Personal Non-Negative tweets were extracted, these tweets were used to compute the correlation between the sentiment trend and the News trend. The details of each "box" in Figure 4.1 will be introduced in the rest of this Section.

4.4.1 Pre-processing of Features

In cases of disease surveillance on Twitter, the classical division of sentiments into positive and negative is inappropriate, because diseases are generally classified as negative. Positive emotions could arise as a result of relief about an epidemic subsiding, but we ignore this possibility. Thus, a two-point "Likert scale" with the points positive and negative would not cover this spectrum well. Rather, we started with an asymmetric four-point Likert scale of "strongly negative," "negative," "neutral" and "positive." We then combined "strongly negative" and "negative" into one category, and "neutral" and "positive" into another. We use "Negative" as the name of the first category and

"Non-Negative" for the second one. Thus, the problem reduces to a two-class classification problem, and a Personal tweet can either be a Negative tweet or a Non-Negative tweet.

Some features need to be removed or replaced. We first deleted the tweets starting with "RT," which indicates that they are re-tweets without comments, to avoid duplications. For the remaining tweets, the special characters were removed. The URLs in Twitter were replaced by the string "url." Twitter's special character "@" was replaced by "tag." For punctuations, "!" and "?" were substituted by "excl" and "ques" respectively, and any of ".,:;-|+=/" were replaced by "symb." Twitter messages were transformed into vectors of words, such that every word was used as one feature, and only unigrams were utilized for simplicity.

4.4.2 Tweet Sentiment Classification

4.4.2.1 Clue-based Tweet Labeling. The clue-based classifier parses each tweet into a set of tokens and matches them with a corpus of Personal clues. There is no available corpus of clues for Personal versus News classification, so we used a subjective corpus MPQA [114] instead, on the assumption that if the number of strongly subjective clues and weakly subjective clues in the tweet is beyond a certain threshold (e.g., two strongly subjective clues and one weakly subjective clue), it can be regarded as Personal tweet, otherwise it is a News tweet. The MPQA corpus contains a total of 8,221 words, including 3,250 adjectives, 329 adverbs, 1,146 any-position words, 2,167 nouns, and 1,322 verbs. As for the sentiment polarity, among all 8,221 words, 4,912 are negatives, 570 are neutrals, 2,718 are positives, and 21 can be both negative and positive. In terms

of strength of subjectivity, among all words, 5,569 are strongly subjective words, and the other 2,652 are weakly subjective words.

Twitter users tend to express their personal opinions in a more casual way compared with other documents, such as News, online reviews, and article comments. It is expected that the existence of any profanity might lead to the conclusion that the tweet is a Personal tweet. We added a set of 340 selected profanity words [116] to the corpus described in the previous paragraph. US law, enforced by the Federal Communication Commission prohibits the use of a short list of profanity words in TV and radio broadcasts [117]. Thus, any word from this list in a tweet clearly indicates that the tweet is not a News item.

We counted the number of strongly subjective terms and the number of weakly subjective terms, checked for the presence of profanity words in each tweet and experimented with different thresholds. A tweet is labeled as Personal if its count of subjective words surpasses the chosen threshold; otherwise it is labeled as a News tweet.

In clue-based classification, if the threshold is set too low, the precision might not be good enough. On the other hand, if the threshold is set too high, the recall will be decreased. The advantage of a clue-based classifier is that it is able to automatically extract Personal tweets with more precision when the threshold is set to a higher value.

Because only the tweets fulfilling the threshold criteria are selected for training the “Personal vs. News” classifier, we would like to make sure that the selected tweets are indeed Personal with high precision. Thus, the threshold that leads to the highest precision in terms of selecting Personal tweets is the best threshold for this purpose.

The performance of the clue-based approach with different thresholds on human-annotated test datasets is shown in Table 4.1. More detailed information about the human-annotated dataset is shown in Section 4.4.3. Among all the thresholds, s3w3 (3 strong, 3 weak) achieves the highest precision on all three human annotated datasets. In other words, when the threshold is set so that the minimum number of strongly subjective terms is 3 and the minimum number of weakly subjective terms is 3, the clue-based classifier is able to classify Personal tweets with the highest precision of 100% but with a low recall (15% for epidemic, 7% for mental health, 1% for clinical science).

Table 4.1 Results of Personal Tweets Classification with Different Thresholds (Precision/Recall)

Threshold	Dataset		
	Epidemic	**Mental Health**	**Clinical Science**
s1w0	0.61/0.69	0.55/0.74	0.48/0.58
s1w1	0.64/0.48	0.53/0.63	0.51/0.52
s1w2	0.70/0.24	0.53/0.38	0.61/0.40
s1w3	0.75/0.18	0.50/0.20	0.58/0.22
s2w0	0.86/0.37	0.53/0.40	0.75/0.42
s2w1	0.86/0.28	0.53/0.38	0.73/0.38
s2w2	0.91/0.15	0.51/0.24	0.76/0.26
s2w3	0.91/0.15	0.37/0.10	0.80/0.16
s3w0	1.00/0.21	0.79/0.21	0.89/0.16
s3w1	1.00/0.21	0.79/0.21	0.88/0.14
s3w2	1.00/0.15	0.84/0.15	0.86/0.12
s3w3	1.00/0.15	1.00/0.07	1.00/0.01

4.4.2.2 Machine Learning Classifiers for Personal Tweet Classification. To overcome the drawback of low recall in the clue-based approach, we combined the high precision of clue-based classification with Machine Learning-based classification in the Personal vs. News classification, as shown in Figure 4.2. Suppose the collection of Raw Tweets of a unique type (e.g. tuberculosis) is *T*. After the preprocessing step, which

filters out non-English tweets, re-tweets and near-duplicate tweets, the resulting tweet dataset is $\boldsymbol{T}' = \{tw_1, tw_2, tw_3, \ldots, tw_n\}$, which is a subset of T, and is used as the input for the clue-based method for automatically labeling datasets for training a Personal vs. News classifier as shown in Figure 4.2.

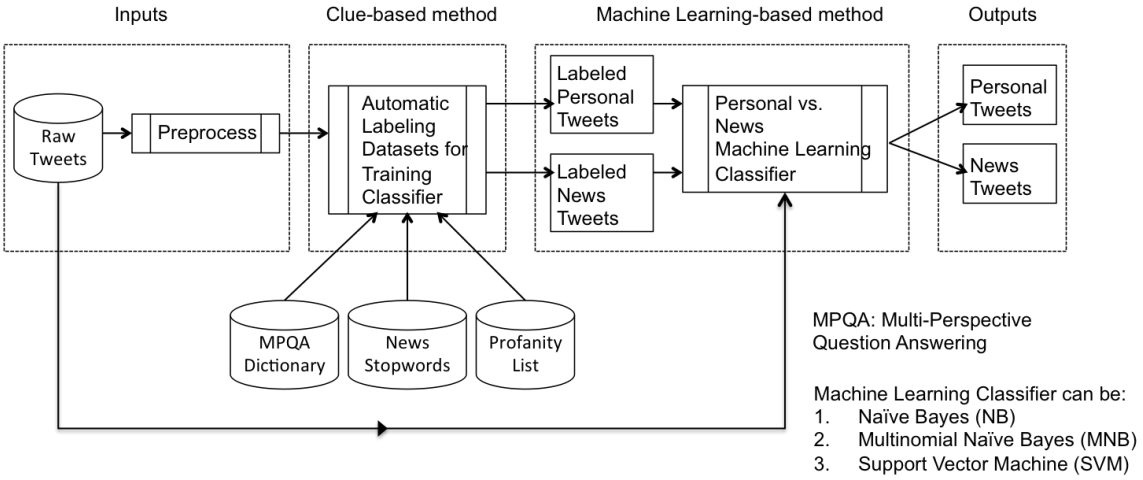


Figure 4.2 Personal vs. News (Non-Personal) classification.

In the clue-based step for labeling training datasets, each tw_i of T' is compared with the MPQA dictionary [114]. If tw_i contains at least three strongly subjective clues and at least three weakly subjective clues, tw_i is labeled as a Personal tweet. Similarly, tw_i is compared with a News stop word list [118] and a profanity list [116]. The News stop word list contains 20+ names of highly influential public health News sources and the profanity list has 340 commonly used profanity words. If tw_i contains at least one word from the News stop word list and does not contain any profanity word, tw_i is labeled as a News tweet. For example, the tweet "*Atlanta confronts tuberculosis outbreak in homeless shelters: By David Beasley ATLANTA (Reuters) - Th... http://yhoo.it/1r88Lnc #Atlanta*" is labeled as a News tweet, because it contains at least one word from the News stop word

list and does not contain any profanity word. We mark the set of labeled Personal tweets as $T_p^{'}$, and the set of labeled News tweets as $T_n^{'}$, note that $(T_p^{'} \cup T_n^{'}) \subseteq T'$.

The next step is the Machine Learning-based method. The two classes of data $T_p^{'}$ and $T_n^{'}$ from the clue-based labeling are used as training datasets to train the Machine Learning models. We used three popular models: Naïve Bayes, Multinomial Naïve Bayes, and Polynomial-Kernel Support Vector Machine. After the Personal vs. News classifier is trained, the classifier is used to make predictions on each tw_i in $T^{'}$, which is the preprocessed tweets dataset. The goal of Personal vs. News classification is to obtain the Label for each tw_i in the tweet database $T^{'}$, where the Label $O(ts_i)$ is either *Personal* or *NT* (News Tweet). *Personal* could be *PN* or *PNN*.

4.4.2.3 Negative Sentiment Classifier. As shown in Figure 4.1, after a classifier for Personal tweets in step 1 is built, the second step in the sentiment classification is to classify the set of *Personal* tweets $T'' = \{tw_i : O(tw_i) = Personal,\ tw_i \in T'\}$ into *PN* (Personal Negative) or *PNN* (Personal Non-Negative) tweets. Figure 4.3 shows the process of classification in this second step. In the rest of this Section, Negative is used to refer to the Personal Negative and Non-Negative is used to refer to the Personal Non-Negative.

In terms of training the classifier for Negative vs. Non-Negative classification, the ideal training dataset must be large and contain little noise. Manual annotation of a training dataset is possible, but this process usually requires different annotators to independently label each tweet and to calculate their degree of agreement. This limits the fast generation of large-sized training datasets. Pang and Lee [93] listed a few annotated corpuses used in previous work in the field of sentiment analysis. These corpuses cover

topics such as customer reviews of products and restaurants. However, to the best of our knowledge, there is no disease-related annotated corpus that can be used as a training dataset to distinguish Negative tweets from Non-Negative tweets.

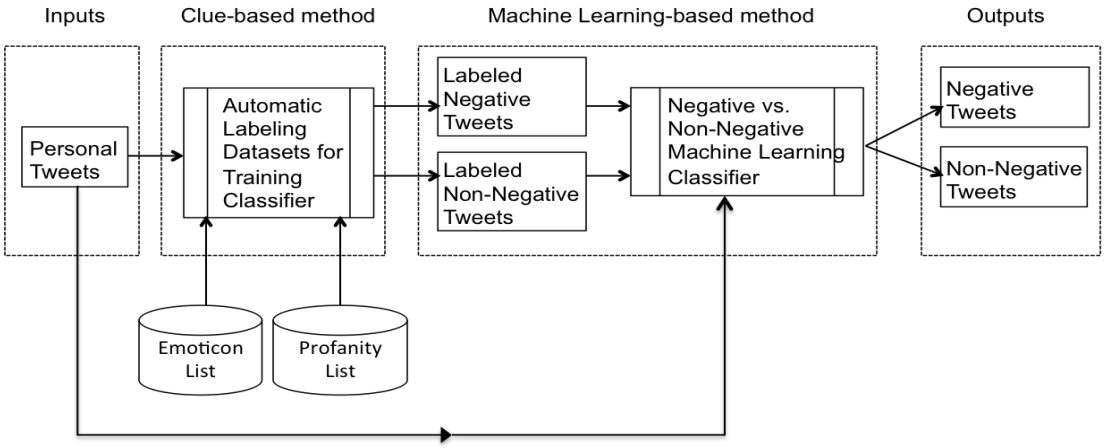


Figure 4.3 Negative vs. Non-Negative classification.

In order to build the training datasets for Negative versus Non-Negative classification (TR-NN), we formed a whitelist and blacklist of stop words using predefined emoticons. An emoticon is a combination of characters that form a pictorial expression of one's emotions. Emoticons have been used as important indicators of sentiments in previous research. We combined the emoticon lists used by Go, et al. [119], Pak and Paroubek [103], and Agarwal, et al. [120]. A partial list of emoticons is in Table 4.2.

The whitelist and blacklist of stop words for building TR-NN are described in Table 4.3. The whitelist is used for extracting while the blacklist is used for eliminating information. A tweet is extracted as a Negative tweet if and only if this tweet contains at least one stop word (or emoticon) from the Negative whitelist, AND does not contain any stop word (or emoticon) from the Negative blacklist. A tweet is extracted as

Non-Negative using similar lists, a Non-Negative whitelist and a corresponding blacklist. For example, the tweet "They are going to take fluid from around the spinal cord to see if she has meningitis... :(" is extracted as a Negative tweet, because it contains at least one emoticon from the Negative whitelist.

Table 4.2 Partial List of The Emoticons Used

Negative	Non-Negative
-.-	:o)
:C	:]
:c	:]
;c	:3
;C	:c)

Table 4.3 Whitelist of Emoticons for Building TR-NN

	Negative	Non-Negative
whitelist	negative emoticons and profanities	neutral and positive emoticons

As shown in Figure 4.3, the emoticons contained in the tweets are used to generate the training dataset TR-NN. Tweets were labeled as *PN* (Personal Negative) or *PNN* (Personal Non-Negative) based on the emoticons they contained. More specifically, if a tweet contains at least one negative emoticon or at least one word from the profanity list that has 340 selected profanity words [116], it is labeled as *PN*. If a tweet contains at least one non-negative emoticon or at least one positive emoticon, it is labeled as a *PNN*. These two categories (*PN* and *PNN*) of labeled tweets were combined into the training dataset TR-NN for Negative vs. Non-Negative classification. Table 4.4 shows examples of tweets in TR-NN. The set of labeled *PN* tweets is marked as $T_{ne}^{''}$, and the set of labeled *PNN* tweets is marked as $T_{nn}^{''}$, and $(T_{ne}^{''} \cup T_{nn}^{''}) \subseteq T'$. Similarly, $T_{ne}^{''}$ and $T_{nn}^{''}$ are used to

train the Negative vs. Non-Negative classifier, and the classifier is used to make predictions on each tw_i in T'', which is the set of Personal tweets. The goal of Negative vs. Non-Negative classification is to obtain the Label for each tw_i in the tweet database T'', where the Label $O(tw_i)$ is either *PN* (Personal Negative) or *PNN* (Personal Non-Negative). (There are no News tweets at this stage.)

Table 4.4 Examples of Personal Negative and Personal Non-Negative Tweets in Training Dataset TR-NN

Personal Negative	I hate TuBerculosis. they get on my damn nerves. they the reason Chrissy don't lotion his ankles or elbows.
	Uh ohhhh! :(CDC: 1 dead, 7 others sickened by listeria traced to cheese
Personal Non-Negative	Car's so fresh and so clean. Time to lay out in the sun with some ruby beer and work on my melanoma :)
	preventing swine flu, one ham at a time. :)

After step 1 (Personal tweets classification) and step 2 (sentiment classification), for a unique type of tweets (e.g. tuberculosis), the raw tweet dataset *T* is transformed into a series of tweet label datasets TS_i. TS_i is the tweet label dataset for time *i*, and $TS_i = \{ts_1, ts_2, ts_3, \ldots, ts_n\}$, where $O(ts_i)$ is either *PN* (Personal Negative), or *PNN* (Personal Non-Negative), or *NT* (News Tweet).

4.4.3 Experimental Results of the Classification Approach

4.4.3.1 Data Collection and Description. We monitored 12 diseases including infectious diseases: Listeria, influenza, swine flu, measles, meningitis, and tuberculosis; four mental health problems: Major depression, generalized anxiety disorder, obsessive-compulsive disorder, and bipolar disorder; one crisis: Air disaster; and one clinical science issue: Melanoma experimental drug. The tweets were collected from

March 13 2014 to June 29 2014. The statistics of the collected datasets are shown in Table 4.5. Only English tweets are used in our experiments. As shown in Table 4.5, some datasets have a larger portion of non-English tweets, for example, influenza, swine flu, and tuberculosis compared with other datasets.

Table 4.5 The Statistics of The Collected Dataset

Dataset Id	Tweet Type	Total number of Tweets	Number of Non-English Tweets	Number of Tweets after Preprocessing
1	Listeria	13,572	1,979	4,544
2	Influenza	1,509,609	716,901	527,489
3	Swine Flu	73,974	35,970	20,430
4	Measles	166,555	8,808	60,016
5	Meningitis	159,393	52,824	42,229
6	Tuberculosis	215,083	147,350	33,030
7	Major Depression	2,269,885	121,649	884,304
8	Generalized Anxiety Disorder	380,094	271,758	71,978
9	Obsessive-compulsive Disorder	434,571	168,061	171,211
10	Bipolar Disorder	51,520	7,416	20,915
11	Air Disaster	15,871	681	5,765
12	Melanoma Experimental Drug	86,757	9,858	40,261

The preprocessing step filters out re-tweets and near-duplicate tweets. Two tweets are considered near-duplicates of each other, if they contain the same tokens (words) in the same order; however, they may contain different capitalization of words, different URLs and different special characters such as @, # etc. For example, the two tweets (1) "*SEVEN TONS OF #HUMMUS RECALLED OVER LISTERIA FEARS... http://t.co/IUU5SiJgjG*" and (2) "*seven tons of hummus recalled over @listeria fears -*

http://t.co/dBgAk1heo4." are near-duplicates, thus only one tweet (randomly chosen) is kept in the database.

4.4.3.2 Evaluation. To the best of our knowledge, there are no evaluation datasets for the performance of sentiment classification of health-related tweets. To compare the three previously discussed classifiers, Naïve Bayes, Two-Step Multinomial Naïve Bayes, and Two-Step Polynomial-Kernel Support Vector Machine, we created one group of test datasets using the clue-based method and a second group of test datasets using human annotation, in order to evaluate the usability of our approach. Weka's implementations [121] of Naïve Bayes, Multinomial Naïve Bayes, and Polynomial-Kernel SVM with default parameter configurations were used for the experiments.

Clue-Based Annotation for Test Dataset

The clue-based annotation of the test dataset was carried out as follows. We first automatically extracted the Personal tweets and News tweets by the clue-based approach and labeled them as Personal or News. Then we randomly divided the labeled dataset into three partitions and used two partitions for training the three different classifiers. Finally, we compared the different classifiers' accuracies on the third partition of labeled data. For example, for Dataset 3 in Table 4.5, in the classification step, 2,899 Personal tweets and 508 News tweets were automatically extracted by using the MPQA corpus [114]. We randomly divided these tweets into training and test datasets, resulting in 1,933 Personal and 339 News tweets as training dataset, and the remaining 966 Personal tweets and 169 News tweets as test dataset. A similar emoticon-based approach was used to

automatically generate a training dataset and a test dataset for Negative vs. Non-Negative classification.

Human Annotation for Test Dataset

Because the clue-based annotation method is automatic, it is relatively easy to generate large samples. However, the drawback is that the training and testing datasets are extracted by the same clue-based annotation rule, thus the results might carry a certain bias. In order to more fairly evaluate the usability of our approach, we created a second test dataset by human annotation, which is described as follows.

We extracted three test data subsets by random sampling from all tweets from the three domains epidemic, clinical science, and mental health, collected in the year 2015. Each of these subsets contains 200 tweets. Note that the test tweets are independent from the training tweets that were collected in the year 2014. One professor and five graduate students annotated the tweets, with each tweet annotated by three people. The instructions for annotators are shown in Appendix B. Annotators were asked to assign a value of 1 if they considered a tweet to be Personal, and a value of 0 if they considered it to be News, according to the instructions they were given. If a tweet was labeled as a Personal tweet by an annotator, s/he was asked to further label it as Personal Negative or Personal Non-Negative tweet. We utilized Fleiss' Kappa [122] to measure the inter-rater agreement between the three annotators of each tweet. Table 4.6 presents the agreement between human annotators. For each tweet, if at least two out of three annotators agreed on a Label (Personal Negative, Personal Non-Negative, or News), we labeled the tweet with this sentiment. Table 4.7 shows the numbers of tweets with different labels. For example, the fraction 25/200 for Negative tweets in "epidemic" means that out of the 200

human-annotated epidemic tweets, 25 tweets were labeled as Personal Negative tweets. The total number of tweets in each dataset does not add up to 200, because in some cases each of the three annotators classified a tweet differently. Tweets for which no majority existed were omitted from the analysis.

Table 4.6 Agreement Between Human Annotators

Domains	Epidemic	Clinical	Mental
Total Number of	200	200	200
At Least Two	192/200	194/200	188/200
Fleiss' Kappa	0.4	0.54	0.33

Table 4.7 Statistics Regarding Human Annotated Dataset

Domains	Epidemic	Clinical	Mental
Total Number of Tweets	200	200	200
Personal Negative Tweets	25/200	10/200	34/200
Personal Non-Negative	34/200	34/200	58/200
News Tweets	133/200	150/200	96/200

4.4.3.3 Classification Results. The results of the two-step classification approach are shown in this section. The performance of Personal vs. News and Negative vs. Non-Negative were tested separately with the clue-based annotated test dataset and the human annotated test dataset.

Results with Clue-Based Annotated Test Dataset

We compared the previously discussed classifiers: Two-Step Naïve Bayes, Two-Step Multinomial Naïve Bayes, and Two-Step Polynomial-Kernel Support Vector Machine. As previously discussed, the labeled dataset was randomly divided into three partitions and we used two partitions for training the three different classifiers. The detailed training

and test dataset sizes are shown in Table 4.8. Note that the test datasets for each classifier in step 2 can be different. The reason is that different classifiers extract different numbers of Personal tweets in the first step, thus the test data in the second step, which is extracted from the previously extracted Personal tweets, can also be different for the three classifiers. The two-step sentiment classification accuracy on individual datasets (1 to 12) is shown in Table 4.9 and confusion matrices of the best classifiers in terms of accuracy are shown in Table 4.10; Similarly, the classification accuracy and confusion matrices of the best classifiers for the three domains (epidemic, mental health, clinical science) are shown in Tables 4.11 and 4.12, respectively.

On individual datasets, all three two-step methods show good performance. SVM is slightly better than the other two classifiers for most of the datasets. For the domain datasets, which combine individual datasets according to their domains, all three two-step methods also exhibit good performance. SVM again slightly outperforms the other two classifiers in all three domains.

Results with Human Annotated Test Dataset

In order to evaluate the usability of two-step classification, Personal vs. News classification and Negative vs. Non-Negative classification were also evaluated with human annotated datasets.

For Personal vs. News classification, we compared our Personal vs. News classification method with three baseline methods. 1) A naïve algorithm that randomly picks a class. 2) The clue-based classification method described in Section 4.4.2.

Table 4.8a Size of Experimental Training and Test Datasets for Personal vs. News Classification

Classifier	**MNB/NB/SVM**	
Dataset Id	**Training (Personal/News)**	**Testing (Personal/News)**
1	206/238	102/119
2	83,032/7206	41,515/3,602
3	1,933/339	966/169
4	5,808/3,770	2,904/1,885
5	3,501/1,094	1,750/546
6	2,863/756	1,431/378
7	262,991/5,163	131,495/2,581
8	8,159/1,301	4,079/650
9	27,972/673	13,985/336
10	5,160/303	2,580/151
11	313/314	156/156
12	7,180/1154	3,590/576

Table 4.8b Size of Experimental Training and Test Datasets for PN vs. PNN Classification (PN is Personal Negative and PNN is Personal Non-Negative)

Classifier	MNB		NB		SVM	
Dataset Id	**Training (PN/PNN)**	**Testing (PN/PNN)**	**Training (PN/PNN)**	**Testing (PN/PNN)**	**Training (PN/PNN)**	**Testing (PN/PNN)**
1	18/8	8/4	19/8	9/4	20/8	9/4
2	32,689/5,346	16,344/2,672	32,420/5,244	16,209/2,621	32,700/5,359	16,350/2,679
3	634/226	316/113	629/228	314/113	636/226	317/113
4	630/112	314/55	618/112	309/56	647/114	323/56
5	658/306	329/152	650/306	325/152	662/307	330/153
6	412/144	205/72	402/132	201/65	414/147	207/73
7	29,153/4,320	14,576/2,160	29,178/4,314	14,589/2,157	29,189/4,326	14,594/2,163
8	2,446/725	1,222/362	2,428/720	1,213/360	2,454/732	1,226/365
9	5,714/2,046	2,856/1,023	5,680/2,030	2,839/1,014	5,714/2,060	2,857/1,029
10	548/92	273/46	546/90	272/45	548/95	274/47
11	28/8	13/3	28/7	14/3	30/10	14/5
12	648/160	324/79	640/158	320/78	648/160	323/79

Table 4.9 Results of S1A/S2A (S1A = Step One Accuracy and S2A = Step Two Accuracy) on Individual Dataset (Rounded to 2 Decimal Places)

Dataset Id	2S-MNB	2S-NB	2S-SVM
1	0.91/0.92	0.90/0.77	**0.99/1.00**
2	0.97/0.95	0.96/0.92	**1.00/0.97**
3	0.97/0.90	0.95/0.94	**1.00/0.97**
4	0.94/0.89	0.90/0.97	**1.00/0.97**
5	0.95/0.91	0.93/0.97	**1.00/0.98**
6	0.96/0.86	0.92/0.97	**1.00/0.99**
7	0.98/0.97	0.98/0.98	**1.00/0.99**
8	0.96/0.90	0.95/0.96	**1.00/0.96**
9	0.98/0.96	0.96/0.98	**1.00/0.98**
10	0.96/0.90	0.95/0.98	**1.00/1.00**
11	0.89/0.81	0.88/0.82	**0.96/0.95**
12	0.92/0.87	0.89/0.98	**1.00/0.98**

Table 4.10 Confusion Matrices of the Best Classifier on Each Dataset (Step 1: Positive Class is Personal and Negative class is News; Step 2: Positive Class is Personal Negative and Negative Class is Personal Non-Negative)

	Step 1					Step 2			
Dataset Id	**Best Classifier**	**True Pos.**	**False Neg.**	**False Pos.**	**True Neg.**	**True Pos.**	**False Neg.**	**False Pos.**	**True Neg.**
1	2S-SVM	101	1	2	117	9	0	0	4
2	2S-SVM	41,513	2	12	3,590	16,121	229	372	2,307
3	2S-SVM	966	0	3	166	311	6	5	108
4	2S-SVM	2,904	0	3	1,882	320	3	9	47
5	2S-SVM	1,749	1	1	545	323	7	5	148
6	2S-SVM	1,431	0	5	373	207	0	4	69
7	2S-SVM	131,494	1	8	2,573	14,494	100	135	2,028
8	2S-SVM	4,079	0	2	648	1,205	21	40	325
9	2S-SVM	13,984	1	1	335	2,819	38	51	978
10	2S-SVM	2,580	0	5	146	274	0	0	47
11	2S-SVM	156	0	11	145	14	0	1	4
12	2S-SVM	3,571	19	1	575	318	5	3	76

Table 4.11 Results of S1A/S2A (S1A = Step One Accuracy and S2A = Step Two Accuracy) on Individual Domain

Dataset Id	**2S-MNB**	**2S-NB**	**2S-SVM**
Epidemic	0.95/0.95	0.94/0.93	**0.99/0.97**
Mental Health	0.97/0.96	0.96/0.97	**1.00/0.97**
Clinical Science	0.92/0.87	0.89/0.98	**1.00/0.98**

Table 4.12 Confusion Matrices of the Best Classifier on Individual Domain (Step 1: Positive Class is Personal and Negative Class is News; Step 2: Positive Class is Personal Negative and Negative Class is Personal Non-Negative)

Dataset Id	Best Classifier	Step 1				Step 2			
		True Pos.	False Neg.	False Pos.	True Neg.	True Pos.	False Neg.	False Pos.	True Neg.
Epidemic	2S-SVM	47,916	9	18	6,695	17,046	245	398	2,652
Mental Health	2S-SVM	20,602	6	3	1,137	4,290	69	88	1,353
Clinical Science	2S-SVM	3,571	19	1	575	318	5	3	76

Table 4.13 Accuracy of Personal vs. News Classification on Human Annotated Datasets

Dataset	Random	Clue-Based	URL-Based	2S-MNB	2S-NB	2S-SVM
Epidemic	0.52	0.77	0.82	0.86	**0.87**	0.71
Mental	0.48	0.56	0.68	0.72	**0.78**	0.59
Clinical	0.49	**0.82**	0.72	0.74	0.71	0.36

Table 4.14 Confusion Matrices of the Best Personal vs. News Classifier on Human Annotated Datasets (Positive class is Personal and Negative class is News)

Dataset Id	Best Classifier	True Positive	False Negative	False Positive	True Negative
Epidemic	2S-NB	52	15	11	122
Mental Health	2S-NB	81	23	21	75
Clinical Science	Clue-Based	21	29	7	143

Recall that in the clue-based method, if a tweet contains more than a certain number of strongly subjective terms and a certain number of weakly subjective terms, it is regarded as a Personal tweet, otherwise as a News tweet. 3) A URL-based method. In the URL-based method, if a tweet contains a URL, it is classified as a News tweet; otherwise the tweet is classified as a Personal tweet. The classification accuracies of different methods and confusion matrices of the best classifiers are presented in Tables 4.13 and 4.14, respectively. The results show that 2S-MNB and 2S-NB outperforms all three baselines in most of the cases. Surprisingly, 2S-SVM does not perform as well as on the clue-based annotated test dataset. It is possible that SVM overfitted to the clue-based annotated dataset, since SVM is a relatively complex model and it infers too much from the training datasets. Overall, all methods exhibit a better performance on the epidemic dataset than on the other two datasets. In addition, as we compare the ML-based approaches (2S-MNB, 2S-NB, 2S-SVM), the ML-based approaches outperform the clue-based approaches in most of the cases. This means that although the ML-based approaches utilize the simple clue-based rules to automatically label the training data, they also learn some emotional patterns that cannot be distinguished by the MPQA corpus. Some unigrams are learned by the ML-based methods and are shown to be useful for the classification, which will be discussed later.

For Negative vs. Non-Negative classification, the second step in the two-step classification algorithm is to separate Negative tweets from Non-Negative tweets. As discussed in Section 4.4.2, the training datasets are automatically labeled with emoticons and words from a profanity list, and then the classifier is trained by one of the three models, Multinomial Naïve Bayes (MNB), Naïve Bayes (NB), and Support Vector

Machine (SVM). The accuracies of Negative vs. Non-Negative classification and confusion matrices of the best classifiers for human annotated datasets are shown in Tables 4.15 and 4.16, respectively. 2S-MNB outperforms the other two algorithms on the epidemic dataset, and 2S-NB outperforms the other two algorithms on the mental health and clinical science datasets. All three classifiers perform better than the random-select baseline, which generates an average of 50 percent accuracy. We can see that although the classifier is trained with tweets containing profanity and tweets containing emoticons, the classifier is still able to perform with an average accuracy of 70+% on human annotated test datasets. Overall, 2S-NB and 2S-MNB both achieved good Negative vs. Non-Negative classification accuracy in terms of accuracy and simplicity, followed by 2S-SVM.

Table 4.15 Negative vs. Non-Negative Classification Results on Human Annotated Datasets

Dataset Id	2S-MNB	2S-NB	2S-SVM
Epidemic	**0.73**	0.59	0.59
Mental Health	0.63	**0.65**	0.57
Clinical	0.64	**0.73**	0.68

Table 4.16 Confusion Matrices of the Best Personal Negative vs. Personal Non-Negative Classifier on Human Annotated Datasets (Positive Class is Personal Negative and Negative Class is Personal Non-Negative)

Dataset Id	Best Classifier	True Positive	False Negative	False Positive	True Negative
Epidemic	2S-MNB	17	8	8	26
Mental Health	2S-NB	18	16	16	42
Clinical Science	2S-NB	4	6	6	28

4.4.3.4 Error Analysis of Sentiment Classification Output. We analyzed the output of sentiment classification. As discussed in Section 4.4.3, we manually annotated 600 tweets as Personal Negative, Personal Non-Negative, and News. We used 2S-MNB, which achieved the best accuracy in our experiments described in Section 4.4.3, to classify each of the 600 manually annotated tweets as Personal Negative, Personal Non-Negative, or News. Then we analyzed the tweets that were assigned different labels by 2S-MNB and by the human annotators. For the Personal vs. News classification, we found two major types of errors.

The first type of error is that the tweet is in fact a Personal tweet, but is classified as a News tweet. By manually checking the content, we found that these tweets are often users' comments on News items (Pointing by URL) or users are citing the News. There are 27 out of all 140 errors belonging to this type. One possible solution to reduce this type of error is that we can calculate what percentage of the tweet text appears in the Web page pointed to by the URL. If this percentage is low, it is probably a Personal tweet since most of the tweet text is the user's comment or discussion, etc. Otherwise, if the percentage is near 100 percent, it is more likely a News tweet since the title of a news article is often pasted into the tweet text.

The second type of error is that the tweet is in fact a News item, but is classified as a Personal tweet. Those misclassified tweets are News items that have "personal" titles, and mostly have a question as title. There are 48 out of all 140 errors belonging to this type. One possible solution is to check the similarity between the tweet text and the title of the Web page content pointed to by the URL. If both are highly similar to each other,

the tweet is more likely a News item. Those two types of errors together cover 54% (75/140) of the errors in Personal vs. News classification.

For Negative vs. Non-Negative classification, in 50% (30/60) of all errors the tweet is in fact Negative, but is classified as Non-Negative. One possible improvement is to incorporate "Negative phrase identification" to complement the current ML paradigm. The appearance of negative phrases such as "I feel bad," "poor XX," and "no more XX" are possible indicators of Negative tweets. Examples of misclassified tweets are as follows:

Make a table and number these to refer to each example easily
"This is the scariest chart I've made in awhile http://t.co/3MH5exZjSh http://t.co/oc9lyEO0XY" (Personal tweet classified as News tweet)

"My OCD has been solved! Get our newsletter here: http://t.co/fAxsHjaIn4 http://t.co/1Jhkbta2Px" (Personal tweet classified as News tweet)

"What is Generalized Anxiety Disorder? (GAD #1) http://t.co/y32GmkYhkh #Celebrity #Charity http://t.co/EYDupOLxY8" (News tweet classified as Personal tweet)

"Basal Cell Carcinoma is the most common form of skin cancer. Do you know what to look for? http://t.co/hmofWTApG9" (News tweet classified as Personal tweet)

"@Jonathan_harrod I know there is some research going on, but... Measles kills and us easily spread. @mercola" (Negative tweet classified as Non-Negative tweet)

"Having a boyfriend with diagnosed OCD is not easy task, let me tell ya" (Negative tweet classified as Non-Negative tweet)

4.4.3.5 Contribution of Unigrams. In order to illustrate which unigrams are most useful for the classifiers' predictions, ablation experiments were performed on Personal vs. News classification and Negative vs. Non-Negative classification on the three human annotated test datasets. The classifier 2S-MNB was used since it took much less time to train and it is only slightly less accurate than the best classifier 2S-NB on the

human-annotated test dataset. 2S-MNB was trained with the automatically generated data from the Epidemic, Mental Health, and Clinical Science domains collected in the year 2014. Then the trained classifiers were used to classify the sentiments of human annotated datasets collected in the year 2015, where unigrams were removed from the test dataset one at a time, in order to study each removed unigram's effect on accuracy. The change of classification accuracy was recorded each time, and the unigram that leads to the largest decrease in accuracy (when removed) is the most useful one for predictions. Table 4.17 shows the ablation experiments for Personal vs. News classification. For example, the unigrams "i", "http", "app", "url" are not in MPQA corpus but are learned by the ML classifier 2S-MNB as the most important unigrams contributing to classification. We did not find any useful unigram in Negative vs. Non-Negative classification by this ablation experiment.

Table 4.17 Most Important Unigrams in Personal vs. News Classification

Dataset	Unigrams with Most Importance
Epidemic	url, i, case, but
Mental Health	url, disorder, often, bipolar
Clinical Science	melanoma, health, http, co, risk, prevention, app

4.4.3.6 Bias of Twitter Data. Twitter may give a biased view, since people who are tweeting are not necessarily a very representative sample of the population. As pointed out by Bruns and Stieglitz [123], there are two questions to be addressed in terms of generalizing collected Twitter data. 1) Does Twitter data represent Twitter? 2) Does Twitter represent society? To answer the first question, according to the documentation [67], the Twitter Streaming API returns at most 1% of all the tweets produced on Twitter

at any given time. Once the number of tweets matching given parameters (keywords, geographical boundary, user ID) is beyond the 1% of all the tweets, Twitter will begin to sample the data that it returns to the user. To mitigate this, we utilized highly specific keywords (e.g., h1n1, h5n1) for each tweet type (e.g., flu) to increase the coverage of collected data [124]. These keywords are shown in Appendix A. As for the second question, Mislove, et al. [125] found that the Twitter users significantly over-represent the densely populated regions of the US, are predominantly male, and represent a highly non-random sample of the race/ethnicity distribution. To reduce the bias of collected Twitter data, we defined the Measure of Concern in relative terms in Section 4.3. It depends on the fraction of all tweets obtained during the day that have been classified as "Personal Negative" tweets. The Measure of Concern analysis will be discussed in more detail in Section 4.5.

4.5 Concern Sentiment Trend Analysis in Public Health

We are interested in making the sentiment classification results available for public health monitoring, especially the results of computing the *Measure of Concern,* to monitor public sentiments towards different types of diseases. Unlike the previous research on *qualitatively* comparing the co-occurrence of sentiment trends with News broadcasts, this paper approaches the problem of *quantitatively* studying the correlation between Twitter sentiment trends and News trends caused by various epidemics. The correlation process is shown in Figure 4.4. There are three inputs for the correlation process. The News tweets are the outputs in the first step, as shown in Figure 4.2; the Personal Negative tweets and the Personal Non-Negative tweets are the outputs in the second step, as shown in Figure 4.3.

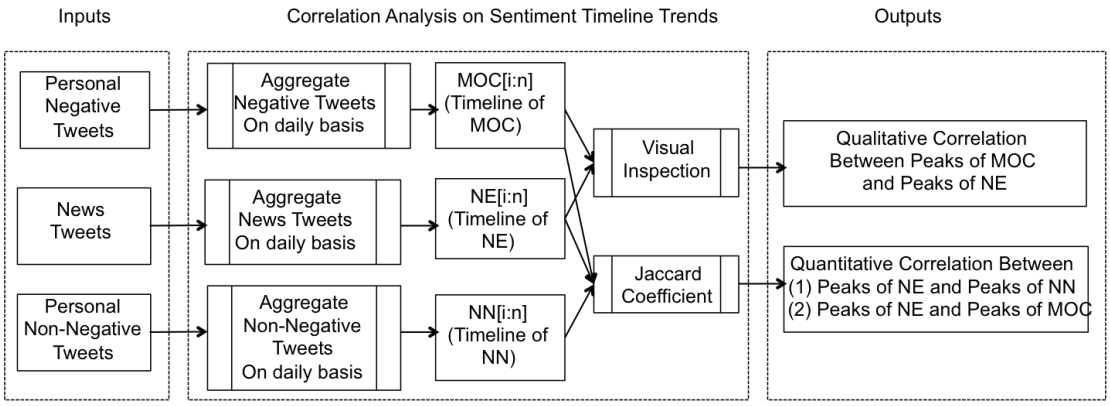


Figure 4.4 Correlation between sentiment trends and News trends.

Given a tweet type, after the two-step sentiment classification method has been applied to the raw tweets, we can produce three timelines: MOC[1:n], NN[1:n], NE[1:n], which are timelines for Measure of Concern, Non-Negative sentiment, and News, respectively.

Next, three sets of peaks P_1, P_2, and P_3 are generated from NE[1:n], MOC[1:n], and NN[1:n], respectively. The time interval of peak is set to *[i-3, i+3]*, which contains 7 days. We are interested in the correlation between P_1 and P_2 (peaks of News and peaks of MOC), and the correlation between P_1 and P_3 (peaks of News and peaks of Non-Negative sentiments). The Pearson Correlation Coefficient (PCC) appears to be a natural way to measure the correlation between two time series, since the PCC is good at measuring the similarity of two linearly dependent variables. However, for the problem addressed here, as we are interested in the News about outbreaks of epidemics, it makes more sense to measure the similarity between the peaks. We utilized the Jaccard Coefficient for this purpose and define the correlations as follows:

$$JC(MOC, NEWS, t) = \frac{|P_{2,c+t} \cap P_{1,c}|}{|P_{2,c+t} \cup P_{1,c}|} \tag{4.9}$$

$$JC(NN, NEWS, t) = \frac{|P_{3,c+t} \cap P_{1,c}|}{|P_{3,c+t} \cup P_{1,c}|} \quad (4.10)$$

$P_{2,c+t}$ is meant to assign a time lag or time lead of t days (depending on the sign of t) to the collection of MOC peaks, thus in (4.9), the News peak at date c will be compared with the MOC peak at date $c+t$. Similarly, $P_{3,c+t}$ is meant to assign a time lag or time lead of t days to the collection of Non-Negative peaks, thus the News peak at date c will be compared with the Non-Negative peak at date $c+t$. The Jaccard Coefficient will have a value between 0 and 1, and the higher the value, the better the two time series correlate with each other.

Figure 4.5 presents an example of using the Jaccard Coefficient (JC) to measure the correlation between peaks of MOC (in green) and peaks of News (in purple). As Figure 4.5 shows, the MOC timeline has seven peaks and the News timeline has six peaks. Three peaks of MOC and another three peaks of News (they are marked by red disks) are pair-wise matched. The remaining four peaks of MOC and the remaining three peaks of News (marked by black disks) are not matched. The JC between the peaks of MOC and the peaks of News is calculated by the size of the intersection divided by the size of the union. In this example, the JC is 3/(7+6-3) = 0.3.

4.5.1 Quantitative Correlation of Peaks

Table 4.18 summarizes the number of peaks in each of the three time series: MOC (Negative sentiment), NN (Non-Negative) sentiment, and News. The best Jaccard Coefficient between MOC peaks and News peaks for a given dataset was computed as follows: First we directly computed the JC between MOC peaks and News peaks without

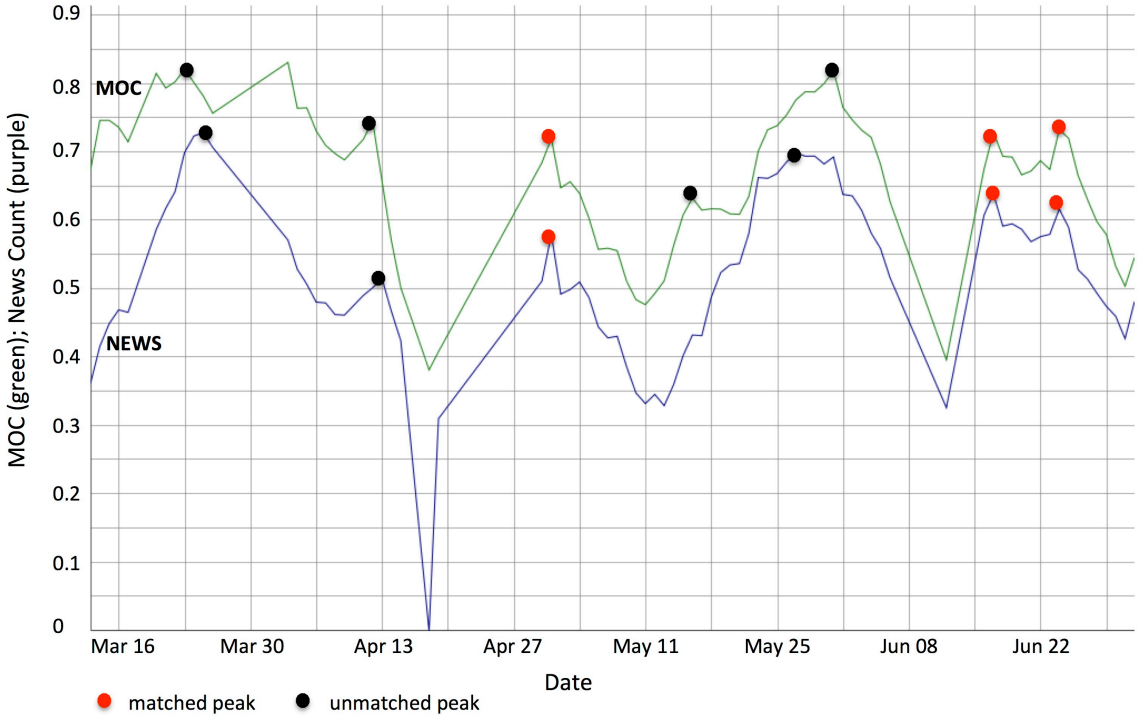


Figure 4.5 An example of calculating the Jaccard Coefficient between peaks of MOC and peaks of News.

any time delay or lead, and we recorded the result. Then we added one, two, or three days of lead to the original MOC, computed the correlation between the revised MOC peaks and the original News peaks respectively, and recorded these three results. Thirdly, we added one, two, or three days of delay to the original MOC, and we recorded three more results. Finally, we chose the highest measure from the above seven results as the best correlation between MOC and News. The best correlation between NN sentiment and News was computed similarly.

The best Jaccard Coefficients between MOC peaks vs. News peaks and between NN peaks vs. News peaks for the three domain datasets are shown in Table 4.18. The $+t$ time means that we delay all MOC peaks or NN peaks to t days later, and the $-t$ time

means that we move all MOC or NN peaks to *t* days earlier. Note that two peaks overlap with each other if and only if the two peaks happen on exactly the same day.

Table 4.18 The Correlation Results of MOC (Measure of Concern) vs. News and NN (Non-Negative) vs. News

Dataset Id	# of Peaks in MOC	# of Peaks in NN	# of Peaks in News	Best JC (MOC vs. News)	MOC vs. News Time Adjust	Best JC (NN vs. News)	NN vs. News Time Adjust
Epidemic	7	8	8	0.25	0	0.231	0
Mental Health	7	6	7	0.273	0	0.3	0
Clinical Science	2	2	3	0	0	0.25	-1

From the Table 4.18, we can see that without any time delay/lead, the peaks of MOC and the peaks of NN (Non-Negative) correlated with the peaks of News in all datasets with a Jaccard Coefficient of 0.25 to 0.3. One exception is in the clinical science dataset, where the peaks of MOC do not correlate with the peaks of News. One possible reason is that there are only two peaks for MOC and three peaks for News.

4.5.2 Qualitative Correlation of Peaks

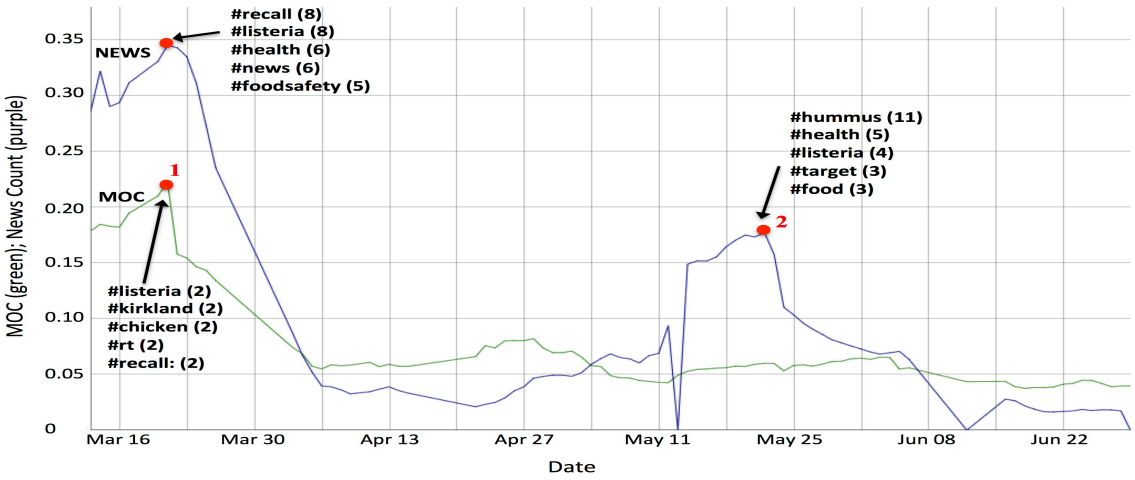
We also qualitatively studied the surges in News and MOC, and how those surges co-occurred with the surges of TV and Internet broadcasts and newspaper articles about real-world events. The timeline trends of (1) listeria, (2) bipolar disorder, and (3) air disaster are shown in Figure 4.6, where the MOC, NN, and NE are min-max normalized, and a 10-day moving average is used to reduce the spikes in values. The top 5 most

frequently mentioned topic terms (hash tags) for the tweets on each peak date are also shown in Figure 4.6. For listeria in Figure 4.6(a), the News Peak 1 occurred because on that same day, several food items produced by Parkers Farm were recalled due to a listeria contamination [126]. We observe that there was a surge in MOC as well. News Peak 2 was caused by the News broadcast that a company is voluntarily recalling more than 14,000 pounds of hummus and dips due to listeria concerns [127].

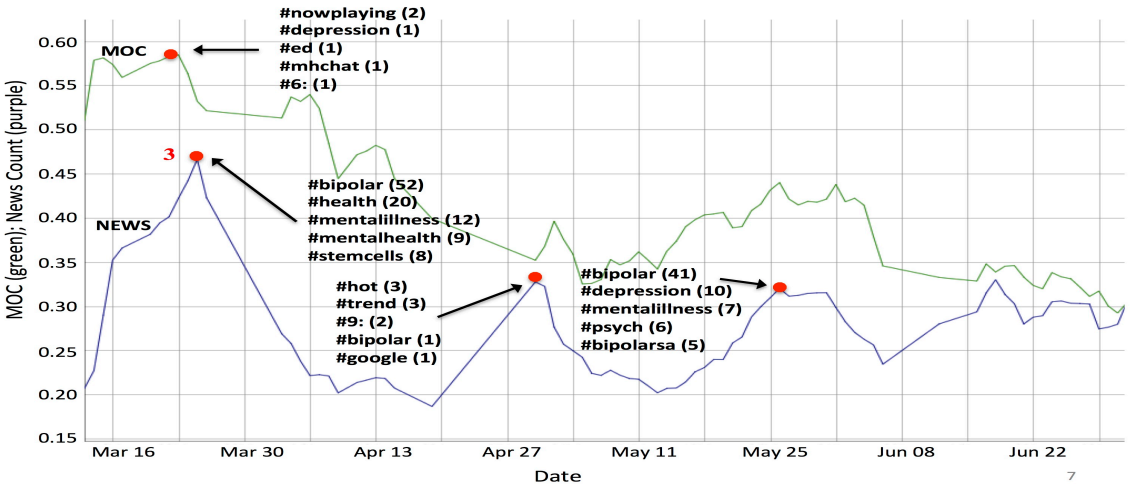
For bipolar disorder in Figure 4.6(b), the News Peak 3 was recorded on 03/25/2014. On that day, researchers reported creating stem cells from the skin of people with bipolar disorder to directly measure cellular differences between people with bipolar disorder and people without [128]. It is surprising to find that the MOC peaks correlated well with this News peak. For air disasters in Figure 4.6(c), the News Peak 4 was recorded on 07/17/2014. On that day, Malaysian Airlines flight MH17 crashed in the Ukraine [129]. There are surges of MOC on the same day as well. This qualitative correlation reveals that in most of the cases, the surges of News generated by our method indeed correlated well with the surges of TV, Internet, and newspaper reports of real-world events. Surprisingly, the surges of MOC also correlate with the surges of News, which shows that the general public tends to express negative emotions according to News peaks during these circumstances.

4.5.3 Prototype System

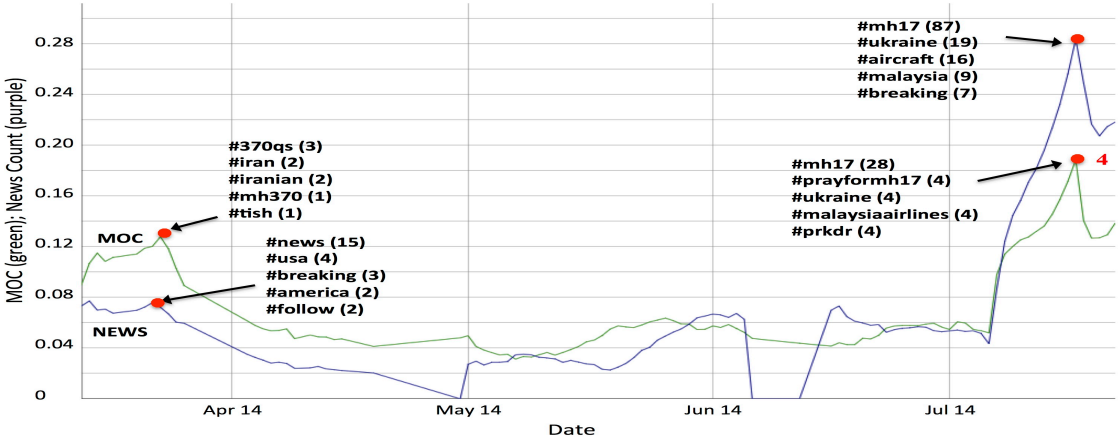
To monitor the timeline and geographic distribution of public concern, we expanded the Epidemics Outbreak and Spread Detection System (EOSDS) visual analytics tools with (1) a concern timeline chart to track the public concern trends on the timeline;



(a)

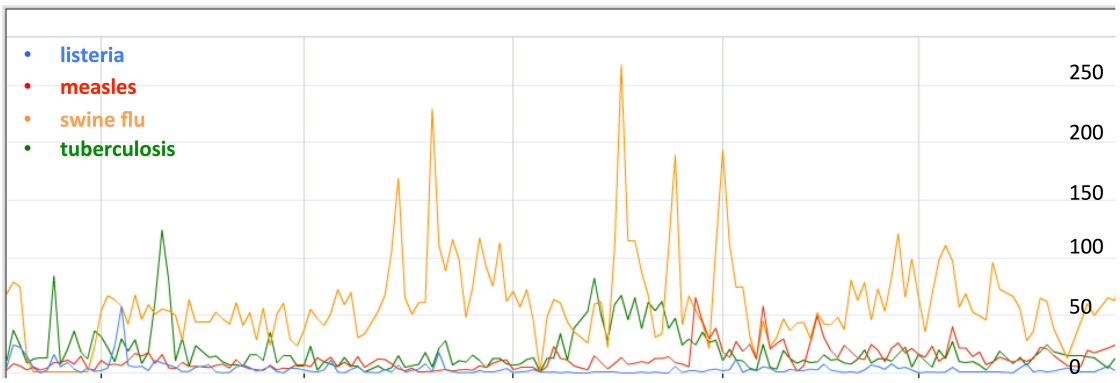


(b)

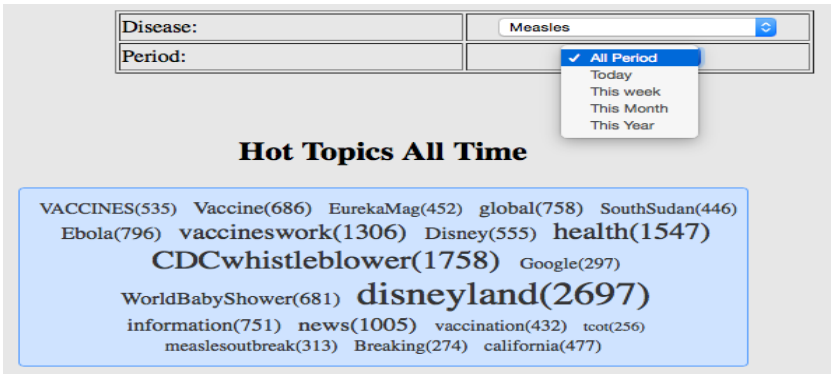


(c)

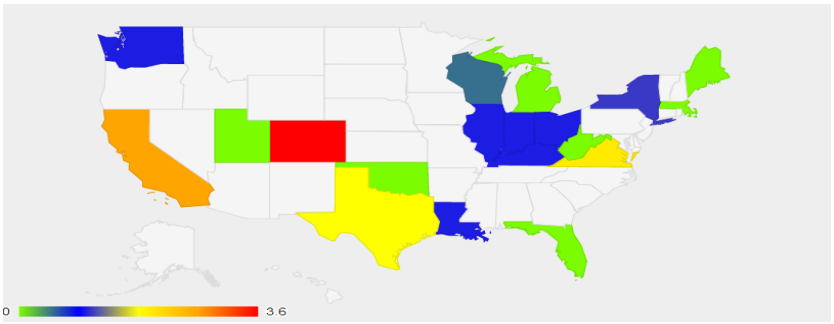
Figure 4.6 Measure of Concern timeline trend (Green) vs. News Timeline Trend (purple): in (a) listeria (b) bipolar disorder (c) air disaster with most frequent topic terms in different peaks.



(a)



(b)



(c)

Figure 4.7 EOSDS visual analytics tools for public concern monitoring (a) sentiment timeline chart (b) topics cloud (c) concern map.

(2) a tag cloud for discovering the popular topics within a certain time period; and (3) a concern map that shows the geographic distribution of concern. The public health specialists can utilize the concern timeline chart, as shown in Figure 4.7(a), to monitor (e.g. identify concern peaks) and compare public concern timeline trends for various

diseases. Then the specialists might be interested in what topics people are discussing on social media during the "unusual situations" discovered with the help of the concern timeline chart. To answer this question, they can use the tag cloud, as shown in Figure 4.7(b) to browse the top topics within a certain time period for different diseases. In addition, the concern map, as shown in Figure 4.7(c), can help public health specialists and government officials to identify parts of the country with different Measures of Concerns towards a particular disease or crisis; thus appropriate preventive actions can be taken in high-concern regions.

4.6 Chapter Summary

We discussed the difficulties of measuring and monitoring public health concerns by traditional public health surveillance systems, due to high expenses, limited coverage, and significant time delays. To address these problems, we used the *Measure of Concern* (MOC), derived from the social network site Twitter, to monitor the public's concern about common health and disaster issues.

To derive the MOC and understand its relationship with the News Count timeline on Twitter, we developed a two-step sentiment classification approach: In the first step, we classify health tweets into Personal tweets versus News tweets. This step separates the tweets that carry the personal opinions of tweeters from those that are third-party factual reports such as News articles. It uses a subjective clue-based lexicon and News stop words to automatically extract training datasets: labeled Personal tweets and labeled News tweets. These auto-generated training datasets are then used to train Machine Learning models to classify whether a tweet is Personal or News. After filtering out News tweets, in the second step, we utilized an emotion-oriented clue-based method to

automatically extract training datasets and generate another classifier to predict whether a Personal tweet is Negative or Non-Negative.

We used the MOC to quantify the health concerns of the tweeting public, and defined a method to both qualitatively categorize and quantitatively measure the correlation between MOC timeline and News Count timeline.

In order to evaluate the two-step classification method, we created a test dataset by human annotation for three domains: epidemic, clinical science, and mental health. The Fleiss's Kappa values between annotators were 0.40, 0.54, and 0.33 for epidemic, clinical science, and mental health, respectively. According to the criteria presented by Landis and Koch [130], the annotators reached a moderate agreement on the epidemic and clinical science datasets, and a fair agreement on the mental health dataset. This result illustrates the complexity of the sentiment classification task, since even humans exhibit relatively low agreement on the labels of tweets.

Experimental results show that (1) In sentiment classification, by combining a clue-based method with a Machine Learning method, our two-step algorithm is able to classify a tweet as Personal Negative, Personal Non-Negative, or News tweet with good accuracy. This overcomes the drawbacks of the clue-based method and the Machine Learning methods when used separately. (2) Quantitatively, the peaks of MOC and the peaks of NN (Non-Negative) (An example of NN peak is that online users express positive emotions when a new vaccine becomes available on the market) correlated with the peaks of News with Jaccard Coefficients of 0.2 to 0.3. Note that this range of Jaccard Coefficient is still too low to make useful predictions. (3) Qualitatively, as we expected, the peaks of News correlated well with the surges of TV, Internet, and newspaper reports

about real-world events. Surprisingly, the surges of MOC also correlated with the surges of News in some cases. This suggests that the general public tends to express negative emotions according to News peaks during these circumstances. (4) As shown in the experiments, our method to derive the MOC is generic and can be applied to topics in other domains, such as mental health monitoring, and clinical science.

CHAPTER 5

PREDICTING INCIDENCE AND TRAJECTORY OF MEDICAL CONDITIONS BY MINING PATIENTS' SOCIAL MEDIA DATA

5.1 Introduction

Through lab tests, certain conditions of patients will be diagnosed, and these diagnosis results will be recorded, e.g., in an electronic health record (EHR). Such a record provides a summary of an individual's medical history and is often made accessible to the patient online. Research has shown that some conditions are correlated with each other to a measureable degree ("comorbidities") [21]. Due to similar molecules, gene structures, and patients' life styles, the appearance of certain conditions leads to a higher likelihood for the occurrence of certain other conditions. These correlation relationships are usually complex. Research has been carried out to predict incidence [24, 131-133] and progression trajectories [134, 135] from her data. However, EHR datasets are usually limited to one medical site or network and have limited coverage of population and time period. Moreover, because of HIPAA law, EHR datasets are rarely accessible to non-affiliated researchers, thus the opportunities for research are often quite limited. To solve these problems, we tapped into self-posted medical histories on a well-known medical social media site, as social media sites are publicly accessible and may cover patients all around the world. The Pew Research Center found that 34% of Internet users have used social media, such as online news groups, websites, and blogs, to read other patients' commentaries and experiences about health issues [136]. There are many patient-oriented social network sites with large user communities. MedHelp [30] has 12 million monthly visitors and claims to be the world's largest health community.

PatientsLikeMe [8], a fast growing social health community, currently has over 187,000 members and covers over 500 health conditions. Compared with EHRs, the data on social media have the advantages of open access and the lack of privacy issues. In patient social media, the patients voluntarily post their health status, with the purpose of letting others view and analyze the data and possibly provide advice.

The objective of this chapter is to predict incidence and progression of medical conditions through modeling comorbidity relationships and trajectories based on self-posted data available on patient-oriented social media. To the best of our knowledge, this is the first non-disease-specific work on modeling comorbidity through patients' social media. Depending on the target of prediction, our objective is divided into two sub-objectives. The first sub-objective is to predict the most probable conditions a patient will develop in the future, given the available medical history posted on his/her social media site. To achieve this goal, we utilized a collaborative filtering technique, which is widely used in applications such as TV programs [137], news [138], books [139], online dating [140], etc. For this problem, patients are viewed as users, diagnosed conditions as items, and presence or absence of a condition as a rating with binary values. We calculated the similarity between a patient's record and other patients' records and derived the risk of a certain medical condition by aggregating similar patients' risks for the same medical condition. The output is a ranked list of medical conditions for a patient, ordered by the likelihood of acquiring this condition. The second sub-objective is to infer medical condition progression trajectories given a certain observed medical condition. The prediction of medical condition incidence and progression trajectory is intended to

help doctors and patients, using patients' social media data, identify potential future conditions more quickly and to practice precision medicine at the earliest possible stage.

5.2 Related Work

One thread of related research is to utilize data mining techniques to predict disease risks for individuals or to rank diseases by their risks. Davis et al. [24, 131] proposed CARE, which is the first well-known system for patient disease prediction using 13+ million elderly patients' hospital visit records. They developed a method to predict the disease risk of one patient based on the disease risks of other similar patients. ICD-9 was used to encode the patients' diseases in their medical records. In addition, similarities of each pair of patients were computed by vector similarity, and then the vector similarity was adjusted by inverse frequency, which gives high weights to rare diseases. Hassan and Syed [133] summarized the reasons why collaborative filtering (CF) can be used to solve the problem of ranking patients along a continuum of risk for adverse outcomes. They incorporated demographics, comorbidity, lab test results, and outcomes into the feature space of their method. They concluded that collaborative filtering is the best method in predicting sudden cardiac death and recurrent myocardial infarction on a real-world dataset containing 4,557 patients' records.

Folino and Pizzuti [132] built a model combining clustering and association analysis to predict the diseases that the patient could likely be affected by in the future. They used a dataset of 1,105 patient records involving 330 distinct diseases collected in a small town of Italy. They utilized the K-Means algorithm to cluster patients and applied association rule analysis to patients in each cluster. Duan et al. [141] proposed to use correlations among nursing diagnoses, outcomes, and interventions to create a

recommender system for constructing nursing care plans. Wiesner and Pfeifer [142] introduced a graph-based data structure of health-related concepts extracted from information in Wikipedia. Based on the health graph, they presented a recommendation procedure that makes use of a similarity measure to compute relevance with regard to users' information needs. Qian et al. [143] investigated the patient risk prediction problem in the context of active learning with relative similarities. They utilized the idea of active learning to predict risks of Congestive Heart Failure and Alzheimer's disease by incorporating relative similarities rather than absolute labels. Different from predicting diseases, Hussein et al. [144] developed the Chronic Disease Recommender System to suggest medical advice and diagnoses to patients.

The above projects predict medical condition incidence but are not able to predict the medical condition progression trajectory (e.g., for Alzheimer's, loss of memory->walking off and becoming lost->difficulty eating and swallowing, etc.). Another thread of research attempts to reveal and infer condition progression trajectories. Jensen et al. [134] investigated the temporal trajectory patterns of all diseases for the entire country of Denmark. On a dataset that contains 6.2 million patients across 14.9 years, they stratified the diagnoses by gender, age, and hospital encounter type, and identified 1,171 significant trajectories. Then they used the Markov Cluster algorithm to identify the five largest clusters of disease trajectories that centered on a small number of key diagnoses (e.g., Diabetes, Cardiovascular Disease). Wang et al. [135] developed a disease progression model based on a Bipartite Bayesian Network; their model was able to identify a few comorbidities and infer the progression trajectory and comorbidity onset of individual patients on a real-world EHR database of over 300,000 Chronic Obstructive

Pulmonary Disease patients over the course of four years. Hainke et al. [145] reviewed a number of disease progression models, which include path models, oncogenetic tree models, distance based trees, directed acyclic graph model, etc.

The existing research suffers from the following limitations. (1) Most of the above methods were developed on a single EHR dataset, which usually has limited coverage in terms of population and time period, and is hard to be integrated with other datasets due to different formats. Our method collected and preprocessed publicly available patients' records, which can cover potentially every patient in the world and across any ongoing time period. (2) Since the graph-based progression trajectory construction process of Jensen et al. [134] is relatively difficult to explain and interpret, we propose a lightweight tree-based model inspired by oncogenetic tree models [145] to help reveal trajectory patterns in an intuitive and efficient manner. Different from oncogentic trees, the actual trajectories and the patients who experience the trajectories were stored in the tree-based model. This was found to be an efficient method for calculating the confidence of future trajectories.

5.3 Predicting Risk of Medical Condition Incidence

In this chapter, we are using PatientsLikeMe. An example of a publicly accessible patient profile on PatientsLikeMe is shown in Figure 5.1. The patient "clairHart" has been diagnosed with 16 medical conditions along with the date of first symptom and diagnosis for each condition. Her diagnosed conditions include Fibromyalgia, Hiatal Hernia (part of the stomach pushes up), Diverticulosis (small and bulging pouches develop in the digestive tract), etc. The objective of this Section is to utilize the collaborative filtering

technique to predict a ranked list of potential conditions for each patient, given the patient's history as posted on PatientsLikeMe.

Figure 5.1 A publicly accessible patient's profile on PatientsLikeMe.

The condition incidence prediction method is shown in Figure 5.2. In the first step, patients' medical histories on their profiles were scraped from the PatientsLikeMe website. After data cleaning and filtering, the preprocessed patients' profiles were fed into the collaborative filtering model, training it to predict comorbidities. When the prediction is applied to an incoming individual patient's record, the collaborative filtering model will compute the similarity between the incoming patient's record and other similar patients' records in the model, and select the neighborhood of patients who are most similar to the specific patient. Finally, the likelihood of each possible medical condition is calculated, and a ranked list of future possible medical conditions is generated for this patient. The details of the condition incidence prediction framework are described as follows.

5.3.1 Collaborative Prediction Model

Patient and condition are represented as a matrix of I by J, where $I = \{all\ patients\}$ and $J = \{all\ the\ possible\ conditions\}$. $J_i = \{C_1, C_2, \ldots, C_n\}$ represents *all the conditions of patient i (ordered by the date of diagnosis)*. Note that $J = \bigcup_{i \in I}(J_i)$. When a new patient is entered, this new patient is regarded as patient *0*, and $J_0 = \{all\ the\ conditions\ of\ patient\ 0\}$. The *head* set H_0 is the new patient's set of conditions that will be compared with other patients to make predictions. In this case $H_0 = J_0$, since all of patient *0*'s diagnosed conditions are used. The set Target of the new patient T_0 is defined as $T_0 = J - H_0$. The goal of the collaborative prediction model is to predict the likelihood and rank each condition in T_0.

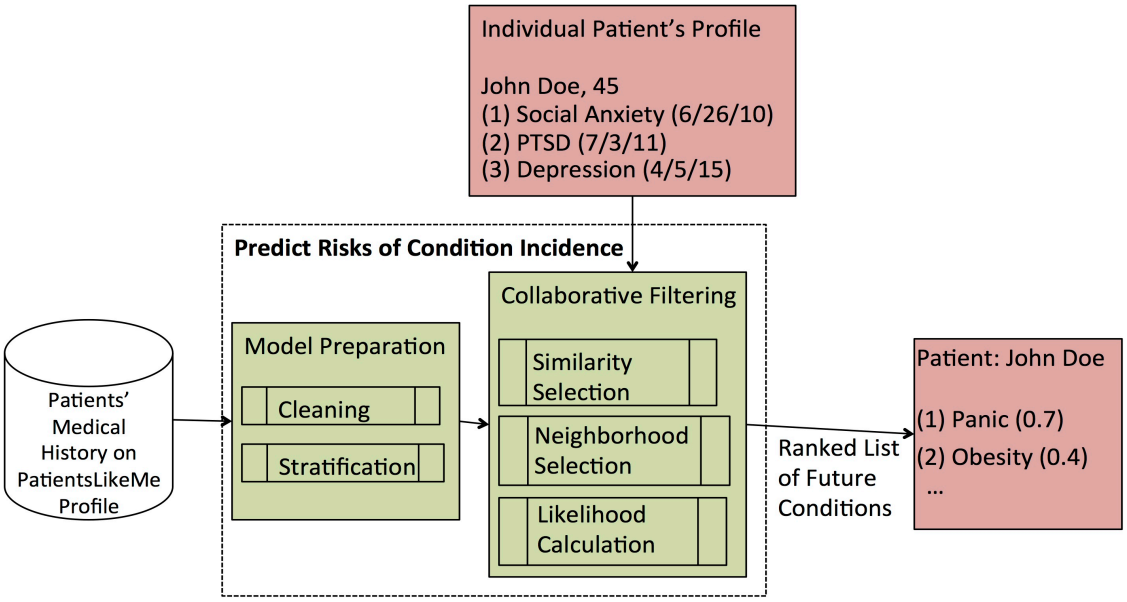


Figure 5.2 Method for predicting risks of medical condition incidence.

For each condition c in T, the neighbors $N_c = \{i \,|\, i \in I \wedge c \in J_i\}$ are all other patients with condition *c*. The probability of patient 0 having condition *c* in the future is calculated by the following equation:

$$P_{0,c} = k \sum_{i \in N_c} w(0,i) \tag{5.1}$$

where k is a normalizing factor, and is defined as the total number of patients in the neighborhood, formally $k = 1/|N_c|$. *w(0,i)* is a measure of the similarity between patient *0* and patient *i,* and is defined as the proportion of conditions of patient *i* to the conditions in *head* set of patient *0*. Formally the similarity of patient *i* and patient *j* is defined in the following equation:

$$w(i,j) = \frac{|\{x \mid x \in head \wedge x \in J_j\}|}{|head|} \tag{5.2}$$

Where *head* contains the conditions of patient *0,* used for comparison with patient *i*. In this new-patient scenario, *head* is the set of all conditions provided by the new patient. Condition *c*'s support S_c is computed by the equation:

$$S_c = \frac{1}{|I|} \sum_{i \in I} f(i) \tag{5.3}$$

where $f(i)$ is an indicator function. $f(i) = 1$ if $c \in J_i$ and $f(i) = 0$ otherwise. The tuple $< 0, c, P_{0,c}, S_c >$ represents the fact that patient *0* has the probability of $P_{0,c}$ of getting condition *c*, and the condition *c*'s support is S_c. After $P_{0,c}$ and S_c have been computed, the list of potential conditions $C*$ is defined as set of tuples $C* = \{< 0, c, P_{0,c}, S_c > \mid c \in T_0\}$ where *c* ranges over every condition in the Target set T_0. The likelihood of patient *0* developing condition *c* in the future is defined by the equation:

$$L_{0,c} = P_{0,c} \times S_c \tag{5.4}$$

where the probability and support are both considered for prediction. In social health

dataset, many high-support conditions (Examples are shown in Table II) are "popular" among patients; Incorporation of condition support measure into the likelihood is able to take into consideration of the frequency of conditions into the prediction.

5.3.2 Collaborative Prediction Example

To illustrate the collaborative prediction model we use a small example dataset shown in Table 5.1. The conditions of each patient are ordered by the patient's diagnosis date for the condition. For example, patient P_2 was first diagnosed with C_1, then diagnosed with C_3, and then diagnosed with C_7 etc. In Table 5.1, the set of all patients is $I = \{P_1, P_2, P_3, P_4, P_5\}$, and the set of all possible conditions is $J = \bigcup_{i \in I}(J_i) = \{C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8\}$. When a new patient P_0 posts diagnosed conditions C_1 and C_3, then $H_0 = \{C_1, C_3\}$. Target $T_0 = J - H_0 = \{C_2, C_4, C_5, C_6, C_7, C_8\}$.

Table 5.1 An Example of Diagnosis Dataset

Patient	Diagnosis
P_1	C_1, C_2, C_3, C_4, C_7
P_2	C_1, C_3, C_7, C_8
P_3	C_2, C_4, C_8, C_7
P_4	C_1, C_5, C_6
P_5	C_5, C_7

Consider the first condition C_2 in T_0, then the following other patients $N_{c2} = \{P_1, P_3\}$ have this condition C_2. $w(P_0, P_1) = 1$ because all conditions are common between P_0 and P_1, and $w(P_0, P_3) = 0$. Then $P_{0,c2} = (1+0)/2 = 0.5$; $S_{c2} = 2/5 = 0.4$. The tuple for condition C_2 is therefore $<0, C_2, 0.5, 0.4>$. Similarly, the tuples for other conditions in T are $<0, C_4, 0.5, 0.4>$, $<0, C_5, 0.25, 0.4>$,$<0, C_6, 0.5, 0.2>$,$<0, C_7, 0.5, 0.8>$, and $<0, C_8,$

0.5, 0.4>. The likelihood of patient 0 developing conditions are as follows: C_2=0.5*0.4=0.2, C_4=0.5*0.4=0.2, C_5=0.25*0.4=0.1, C_6=0.5*0.2=0.1, C_7=0.5*0.8=0.4, and C_8=0.5*0.4=0.2. Therefore the ranked list for patient 0 is $(C_7, C_2, C_4, C_8, C_5, C_6)$. Note that the order is not unique as, e.g., C_2=C_4.

5.4 Constructing Medical Condition Progression Trajectory

In many situations it is more desirable to predict a medical condition progression trajectory, instead of predicting a single medical condition. The trajectories stemming from a medical condition can provide a potential set of paths the patient may end up, as well as explain the likelihood of paths for a final condition for a patient who suffer from one condition. We propose a trajectory model to track the progression and infer the most probable future trajectories given a patient's observed diagnosis history. A trajectory from a condition c is modeled as a tree $T(c) = (N, E)$ where $N=\{C_1, C_2, \ldots C_n\}$ is a set of nodes to represent the conditions and $E=\{e_1, e_2, \ldots, e_3\}$ is a set of edges where each edge $e = (C_i, C_j)$ and represents a progression from condition C_i to condition C_j.

There are three steps to generate and make use of the trajectory tree. The first step is to discover edges of conditions from patients' diagnoses histories as made public in their PatientsLikeMe profiles. The second step is to generate the trajectory model, based on the edges created in the first step. In the last step, the trajectory model is used to infer the confidence value and support of potential progression trajectories given a patient's diagnosis history. More in detail:

Edge Discovery: This step helps identify directional edges of comorbidities, which co-occur for individual patients. A directed edge e_i is defined as: $e_i = \{(C_j, C_k) \mid A$ *patient was diagnosed with conditions* C_j *and* C_k *in temporal order*$\}$. In order to calculate

confidence value and support of a trajectory, the patients with these edges are defined as: $I(e_i) = \{Patients\ who\ have\ the\ edge\ e_i\}$.

Linking: The generated edges are recursively linked to build the condition trajectory tree *T* by recognizing the common node (condition) in two edges. In other words, given $e_1=(C_1,C_2)$, $e_2=(C_2,C_4)$, and $e_3=(C_4,C_5)$ a tree is built with an edge trajectory $e_1 \rightarrow e_2 \rightarrow e_3$ resulting in a condition trajectory ($C_1 \rightarrow C_2 \rightarrow C_4 \rightarrow C_5$). Note that we use edges to represent the trajectory for implementation purpose. For interpretation purposes, the edge trajectory can be converted into condition trajectory by combining overlapping conditions.

The algorithm for building the edge trajectory tree is shown as Algorithm 5.1. In Algorithm 5.1, the current edge $ce = (C_i, C_j)$ and the new edge $ne = (C_k, C_h)$ of conditions are linkable if $C_k = C_j$ are the same condition, and *ne* will not create a cycle in the current path. The trajectory model can be used to infer the confidence value of a medical condition trajectory given a certain observed condition. Suppose a edge trajectory $ti = \{e_1, e_2, e_3, \ldots, e_n\}$, then U_{ti} (the set of patients who have trajectory *ti*) is the intersection of the sets of the patients who have the same chain of linkable edges. Formally:

$$U_{ti} = \cap \{I(e_i)\ where\ e_i\ is\ a\text{n edge}\ in\ trajectory\ ti\} \tag{5.5}$$

Inferring: We are defining the support of trajectory *ti* (slightly differently from the standard definition) as $|U_{ti}|$. The confidence value *C* of edge trajectory ($e_1 \rightarrow e_2 \rightarrow e_3, \ldots, \rightarrow e_n$) given an observed condition *c* is calculated as a conditional probability, where e_1 is the starting edge and $e_1 = (null, c)$.

$$\mathrm{C}(e_1 \rightarrow e_2 \rightarrow e_3, \ldots, \rightarrow e_n \mid \mathrm{c}) = |\mathrm{U}_{ti}|/|I(e_1)| \tag{5.6}$$

The comorbidity index (CI) of trajectory ($e_1 \rightarrow e_2 \rightarrow e_3, \ldots, \rightarrow e_n$) is defined as follows:

$$\mathrm{CI}(e_1 \rightarrow e_2 \rightarrow, \ldots, \rightarrow e_n) = |\mathrm{U}_t| / \sum_{conditions} |PS(c)| \quad (5.7)$$

To better illustrate the above method, let's consider the example dataset presented in Table 5.1. After we applied the pair generation process on this dataset, 20 edges were generated (sorted by number of patients): E_{ex}={*(C1,C3), (C1,C7), (C3,C7), (C2,C4), (C2,C7), (C4,C7), (C5,C7), (C1,C8), (C3,C8), (C7,C8), (C1,C2), (C1,C4), (C2,C3), (C3,C4), (C1,C5), (C1,C6), (C5,C6), (C2,C8), (C4,C8), (C8,C7)*}. By using Algorithm 5.1 and setting the starting condition to be C_2, the trajectory model is generated and shown in Figure 5.3. Algorithm 5.1 is called with these input: E_{ex}, 0, 4, (*null*,C_2), an empty path.

Algorithm 5.1 Build Condition Trajectory Tree

```
Input: set of edges E, current depth cd, maximum depth
md, current edge ce, path pa
Output: trajectory model
begin
   /* limit trajectories to a certain length*/
   if current depth cd is equal to maximum depth md
      return
   end if
   for each edge ne in edge set E
      /*if two edges can be linked, recursively build tree */
      if (ne is linkable with current edge ce)
         add ne as a child of current edge ce
         /*path is used for tracking patients of trajectory*/
         append ne to the tail of path pa
         call Algorithm 1 with input E, cd+1, md, ne, pa
         remove ne from the path pa
      end if
   end for
   return
end
```

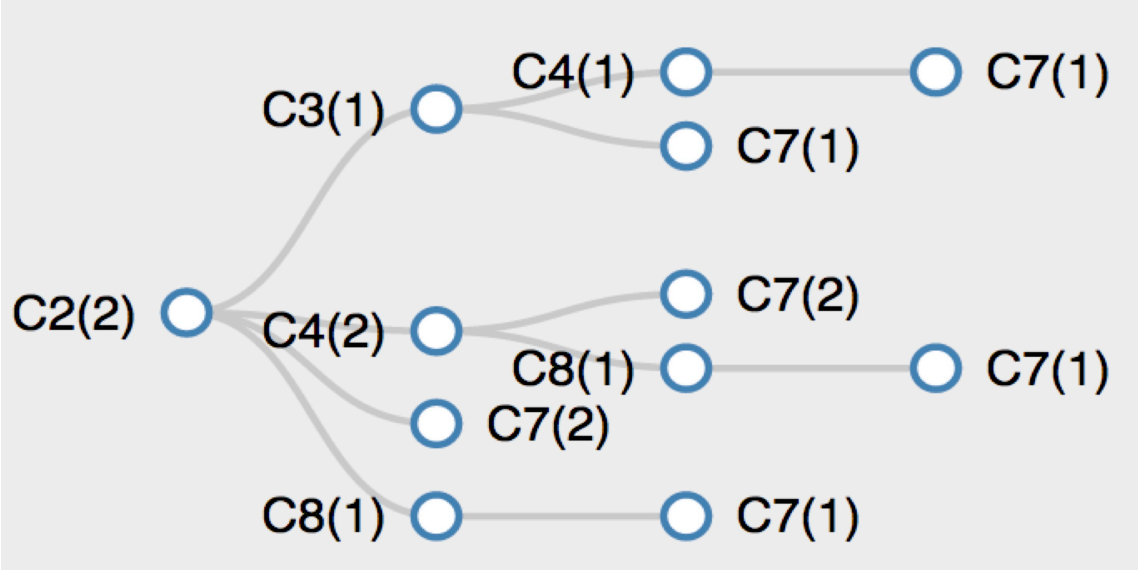


Figure 5.3 The example of condition trajectory starting from condition C2.

5.5 Evaluation Study

5.5.1 Data Description and Analysis

The evaluation dataset was collected by scraping patients' public profiles in PatientsLikeMe. The collected dataset contains 17,418 patients' basic information, including id, username, gender, age and location. Among the patients who have specified their gender, 3,932 are male and 8,023 are female patients.

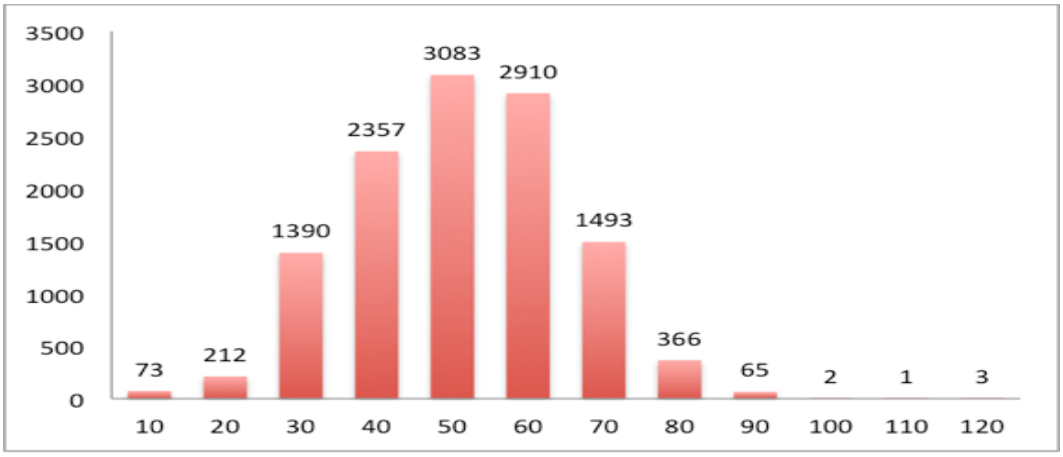


Figure 5.4 The age distribution of collected dataset.

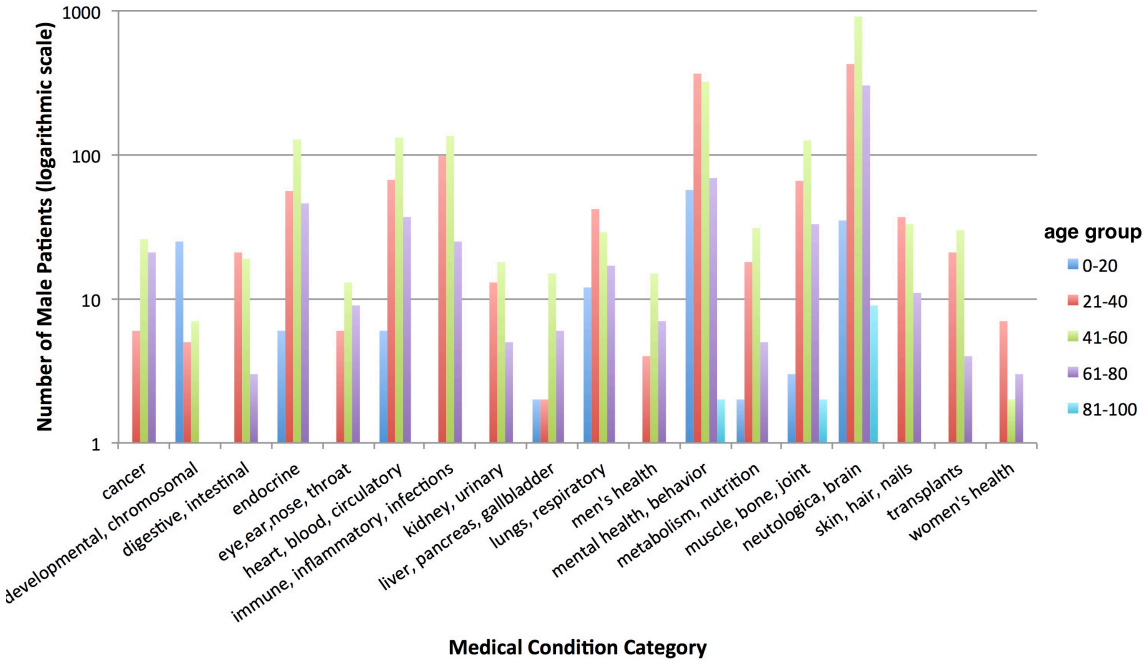
The age distribution is shown in Figure 5.4. The dataset also contains 35,606 diagnoses for these patients. Each diagnosis contains six attributes: PatientId, HasCondition, ConditionId, IsPrimaryCondition, FirstSymptomDate, and DiagnosisDate, for example, "ID: 8, HasCondition: Stroke, ConditionId: 48, IsPrimaryCondition: 0, FirstSymptomDate: May 1998, DiagnosisDate: Sep 1998". This means that the patient (PID=8) had a Stroke (CID=48), which is not his primary condition, and the Stroke's symptoms first showed up in May 1998 and it was diagnosed in September 1998. For each patient, the minimum number of conditions is 0, average is 2, and maximum is 77. The conditions with most patients are shown in Table 5.2.

Table 5.2 The Conditions with Most Patients

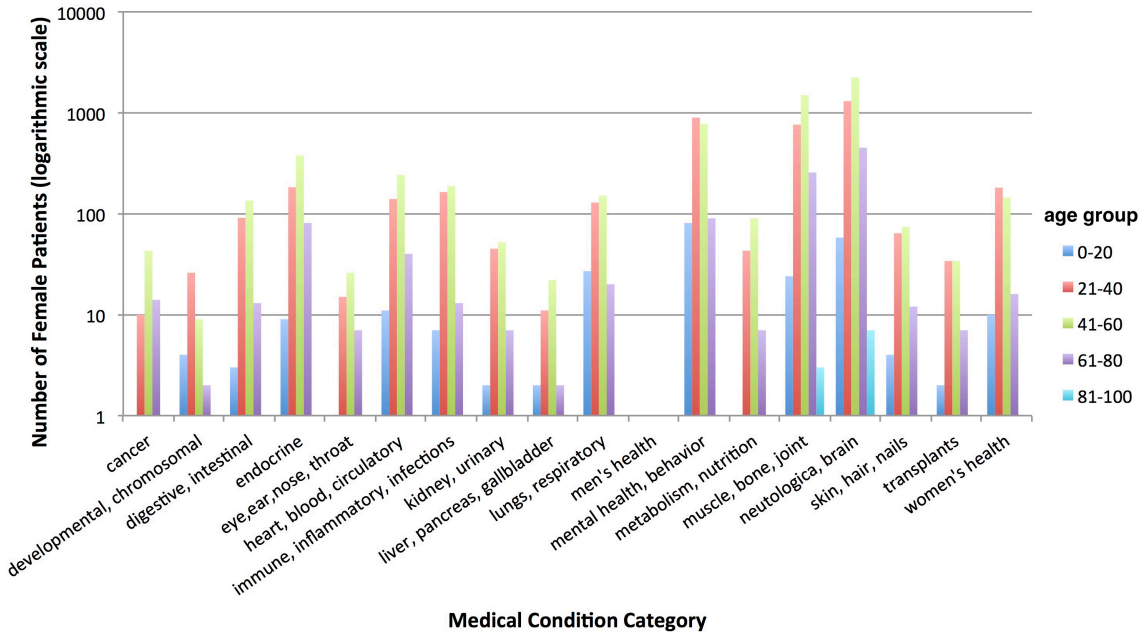
Condition	Number of Patients
MS (Multiple Sclerosis)	3459
Fibromyalgia	3164
Major Depressive Disorder	1624
Generalized Anxiety Disorder	1106
Chronic Fatigue Syndrome	914

Medical conditions strongly correlate with gender and age, thus it is of importance to investigate the distribution of gender and age across different medical condition categories (e.g., mental health, respiratory, etc.). To reveal the effect of gender and age, a stratification analysis was carried out. Figure 5.5 shows the effect of stratification of medical condition categories defined in PatientsLikeMe [146]. There are 18 categories and 174 medical conditions, each of which is classified into one of the categories. Figures 5.5(a)(b) shows the distributions of male and female patients.

In Figure 5.5, we compared the gender difference in each category, and found that most of the categories have similar ratios as the overall ratio (female/male=8023/3932=2.04) except for five categories: "muscle, bone, joint," "digestive, intestinal," "lungs, respiratory," "women's health" and "men's health." These five categories have gender ratios of 11.05, 5.62, 3.27, 27.09 and 0 respectively. For "muscle, bone, joint," the reason is that most of the patients in this category have the condition "Fibromygia"; 90% of Fibromygia patients are female [147], which aligns well with our gender ratio. Most patients in "digestive, intestinal" have IBS (Irritable bowel syndrome) and the ratio of 5.62 is slightly higher than the ratio of 2 reported by Mayo Clinic [148]. The significant gender difference for "women's health" and "men's health" is visible, and we found out that 13 male patients suffer from Postpartum Depression (sic!), a clinical depression after child birth and a "women's health" condition.



(a)



(b)

Figure 5.5 The number of (a) male and (b) female patients in each medical condition category.

We also compared the age distributions. The age range from 40-60 has the most patients, followed by 20-40, and then followed by 60-80. One exception is that "developmental, chromosomal" has significantly more male patients aged 0-20 than ones in other age ranges. This is, because many male patients aged 0-20 suffer from Autism. Another exception is that "neurological, brain" has significantly more male patients aged 81-100.

5.5.2 Evaluation of Predicting Medical Condition Incidence

To evaluate the collaborative prediction approach, we used a leave-one-patient-out validation strategy similar to Davis et al. [24]. Refer to Section III that head of patient i is $|H_i|$ and the head size $|H_i|$ is a parameter in experiments. Only the patients that have

$|H_i|+1$ conditions are used for validation. Among these patients, each time one active patient i is taken out and other patients are used for training. Then the first $|H_i|$ conditions of patient i are fed into the trained model and other $|J_i|$ - $|H_i|$ conditions of patient i are considered as future conditions and used for evaluation. The top-K conditions in the predicted ranked list are considered. We used two metrics: *coverage* and *rank* to evaluate the prediction performance for each patient. The *coverage* is the proportion of correct future conditions in top-K ranked list to the total number of correct future conditions. The rank is the average rank of all correct future conditions in the ranked list for this patient. The process is repeated for each patient, and an average of coverage and rank is computed in the end.

The results are shown in Table 5.3, where K is the size of the predicted ranked list and head size is 2. Collaborative prediction model has a coverage value of 48% and 75% for top-20 and top-100 ranked lists respectively. These results have better coverage (7% and 15% increase) and slightly higher average rank (1.5 and 1.4 increase) when compared with the results reported by Davis et al. [24], which uses the EHR data. Our results show that the collaborative prediction model is able to make good prediction based on patients' social media data.

Table 5.3 Condition Incidence Prediction Results

Top-K	Average Coverage	Average Rank
20	48%	7.25
100	75%	21.59
All	100%	123.29

5.5.3 Evaluation of Progression Trajectories

Next we show the medical condition progression trajectories generated by our tree-based model and compare the trajectory results with the real comorbidities. We selected six medical conditions, namely "Major Depressive Disorder" [149, 150], "Migraine" [151], "IBS" [152-154], "Eating Disorder" [155], "Obesity" [156-158]and "Bipolar I" [159, 160]. Their comorbidities are listed in Table 5.4. We chose each of these six conditions as the "starting root" and generated the tree-based trajectory model (see Algorithm 1) by setting the trajectory's minimum support to 5. The trajectory results are shown in Table 5.5. The trajectories are first ranked in terms of their length. Within the same length, the top-2 trajectories in terms of the comorbidity index are shown.

As shown in Table 5.5, the trajectories cover most of the comorbidities reported in the medical literature. More importantly, different from the previous research [134, 135], which predicts incidence or visualize temporal trajectory patterns, our tree-based model predicts the confidence of the future trajectory and reveals every possible path between any two medical conditions (e.g., IBS->GERD->RLS and IBS->RLS), which can help doctors and patients better understand the medical conditions.

5.5.4 Progression Trajectory Analysis

To illustrate how the trajectories can be used to help doctors reveal the progression paths of medical conditions, we performed a case study on one progression trajectory starting with "Major Depressive Disorder" (MDD, Figure 5.6). The numbers in () indicate the numbers of patients following the trajectory from the root to the current node, e.g., there are 17 patients with (MDD->Fibromyalgia-> IBS).

Table 5.4 Comorbidities of The Selected Conditions from Medical Literature

Major Depressive Disorder (MDD)	Dysthymia, Panic Disorder, Agoraphobia, Social Anxiety, Obsessive–Compulsive Disorder, Generalized Anxiety Disorder, and Post-Traumatic Stress Disorder, Alcohol Dependence, Psychotic Disorder, Antisocial personality, Eating Disorders, Borderline Personality Disorder
Irritable Bowel Syndrome(IBS)	Major Depression, Anxiety, Somatoform Disorders, Fibromyalgia, Chronic Fatigue Syndrome, Gastroesophageal Reflux Disease, Restless Legs Syndrome
Eating Disorder (ED)	Obsessive–Compulsive Disorder, Bipolar Disorder, Substance Abuse, Diabetes, Bone Disease, Cardiac Complications, Gastrointestinal Distress
Obesity	Type 2 Diabetes Mellitus, Hypertension, Dyslipidemia, Cardiovascular Disease, Stroke, Sleep Apnea, Gallbladder Disease, Hyperuricemia And Gout, Osteoarthritis, IBS, Sleep Apnea Disorder
Bipolar I	Substance Abuse, Generalized Anxiety Disorder, Simple Phobia, Social Phobia, Obsessive-Compulsive Disorder, PTSD, Panic Disorder
Migraine	Stroke, Sub-Clinical Vascular Brain Lesions, Coronary Heart Disease, Hypertension, Psychiatric Diseases, RLS, Obesity, Epilepsy, Asthma, Irritable Bowel Disease, Chronic Fatigue Syndrome, Fibromyalgia

Table 5.5 Trajectory Results Starting from The Selected Conditions (Comorbidity Index in Percentage/Confidence in Percentage/Support); * Indicates That The Comorbidity Exists in Medical Literature.

Major Depressive Disorder (MDD)	MDD-> Post-Traumatic Stress Disorder (PTSD)* ->Panic Disorder* -> Social Anxiety Disorder* (0.25/1.3/9) MDD->PD*->SAD*->Phobic Disorder (0.23/1.1/8) MDD->Generalized Anxiety Disorder (GAD)*-> Obsessive-Compulsory Disorder (OCD)* (0.7/3/23) MDD->PD*->OCD* (0.7/2/19) MDD->Bipolar II (1.7/4/21) MDD->Borderline Personality Disorder* (1.2/3/21)
Irritable Bowel Syndrome(IBS)	IBS-> Gastroesophageal Reflux Disease (GERD)*->Restless Legs Syndrome (RLS)* (0.9/3/6) IBS->Fibromyalgia*-> Chronic Fatigue Syndrome (CFS)* (0.3/9/17) IBS->RLS* (6/12/23) IBS->Osteoarthritis (3/10/18)
Eating Disorder (ED)	ED->Tobacco Addiction->Drug Addiction*->PD (0.2/4/5) ED->OCD*->PD->SAD (0.2/4/5) ED->Bipolar II*->Drug Addiction (0.6/5/6) ED->Drug Addiction*->Alcohol Addiction* (0.6/6/7) ED->Postpartum Depression (2/13/15) ED->Alcohol Addiction* (2/13/16)
Obesity	Obesity->Hypertension*->IBS* (0.6/6/5) Obesity->MDD->CFS (0.1/6/5) Obesity->Sleep Apnea Disorder* (4/12/10) Obesity->Plantar Fasciitis (3/6/5)
Bipolar I	Bipolar I->OCD*->Tobacco Addiction* (0.4/5/6) Bipolar I->Tobacco Addiction*->Drug Addiction* (0.4/4/5) Bipolar I->Bipolar II Disorder* (2/6/7) Bipolar I->PD* (1.9/15/17)
Migraine	Migraine->IBS*->Fibromyalgia*->CFS* (0.1/2/7) Migraine->Temporomandibular Joint Disorders (TMJ)->IBS* (0.6/2/5) Migraine->IBS*->CFS* (0.6/4/10) Migraine->Sleep Apnea Disorder (3/7/18) Migraine->Tension Headache (3/6/15)

In Figure 5.6, the most frequent length-2 trajectories are (MDD->GAD) (165 patients) and (MDD->Fibromyalgia) (127 patients). The most frequent length-3 trajectory is (MDD->GAD->PD); (PD=Panic Disorder). In other words, the confidence of (MDD->GAD->Panic Disorder) given the observed condition MDD is 37/680 = 5.4%. The other length-3 trajectories between MDD and PD are (MDD->Dysthymia->PD) (23 patients), MDD->PTSD->PD) (22), MDD->Social Anxiety Disorder->PD) (17).

One possible explanation of this result is that Bouchard et al. [161] found that in young adults with low levels of lead exposure, higher blood lead levels were associated with increased risks of MDD and Panic Disorder, which confirmed the comorbidity of MDD and PD. Our tree-based model reveals the intermediate nodes between these two medical conditions. In this case, MDD can progress to PD via GAD, Dysthymia, PTSD, or Social Anxiety Disorder.

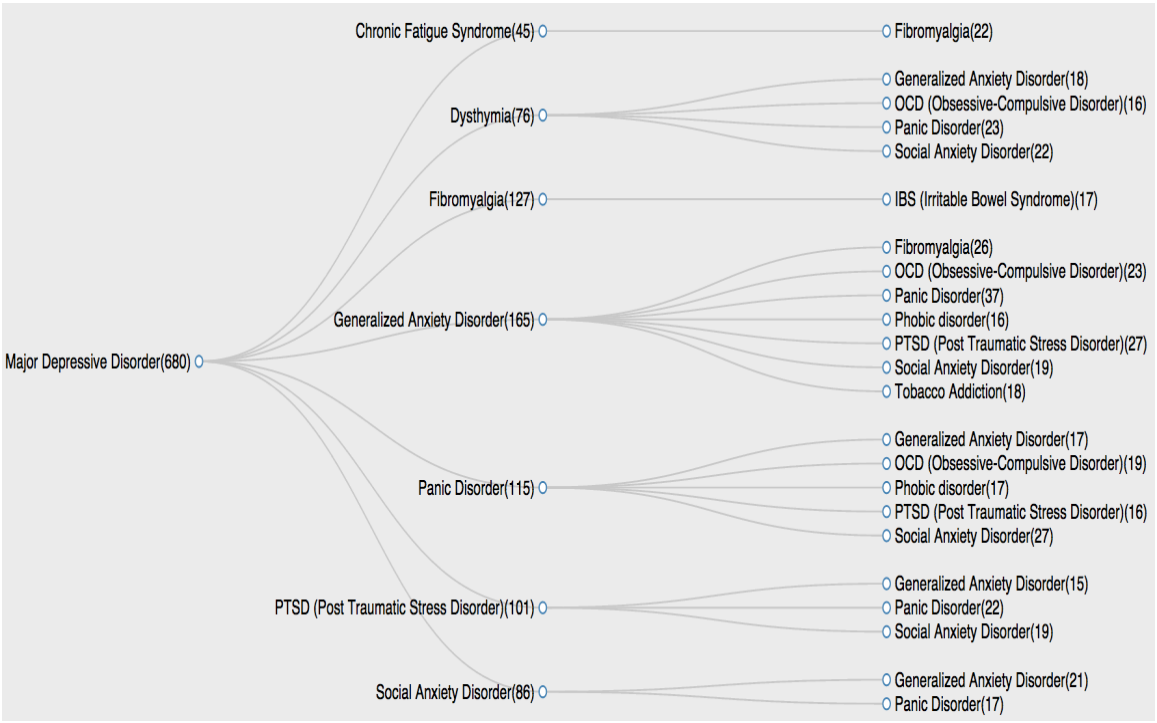
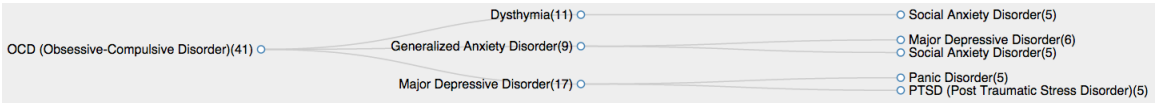


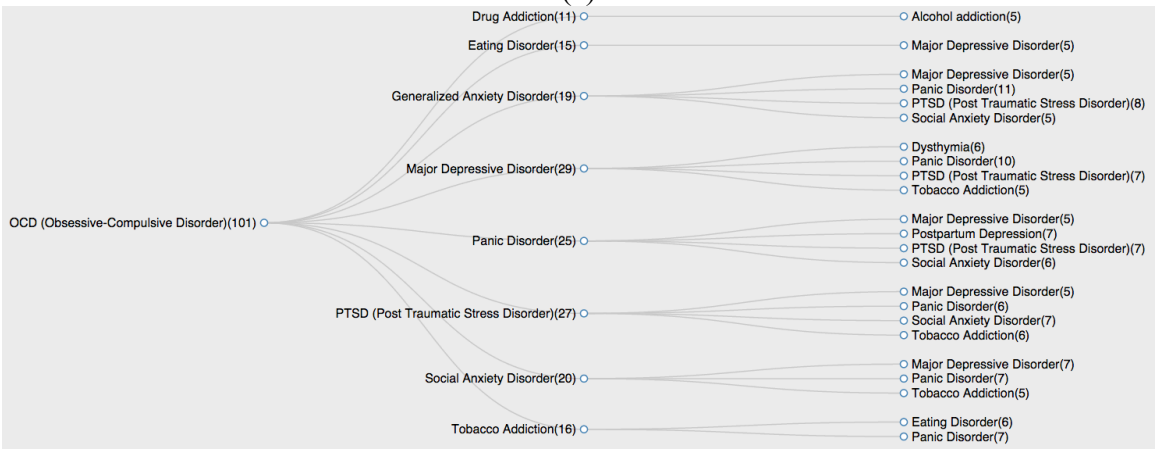
Figure 5.6 The medical progression trajectory starting from "Major Depressive Disorder".

5.5.5 Discussion on Gender-specific Trajectories

In this section, we discuss stratifying the trajectories based on patients' gender to investigate the possible gender-specific disease progression trajectories. We run trajectory model separately on 8,023 female patients and on 3,932 male patients by specifying each condition as the root. The preliminary results show that the trajectory trees show significant difference in terms of the size and progression courses across gender for most of the conditions. The male and female trajectory tree starting from Obsessive-Compulsory Disorder (OCD) is shown in Figure 5.7.



(a)



(b)

Figure 5.7 The conditions trajectories of (a) male and (b) female patients starting with Obsessive-Compulsive Disorder.

Male and female patients show the many identical trajectories (e.g., OCD->Generalized Anxiety Disorder (GAD)->Major Depressive Disorder (MDD), OCD->GAD->Social Anxiety Disorder (SAD), OCD->MDD->Panic Disorder, and

OCD->MDD->PTSD). One exception is that the male patients show the trajectory of OCD->Dysthymia->SAD, which is not found in female patients. To validate this observation, we searched the related medical articles. Assuncao et al. found out that a third of OCD patients has social phobia (SAD), which was significantly associated with male gender, dysthymia, and generalized anxiety disorder (GAD). This study could possibly explain that why only male patients show the trajectory of OCD->Dysthymia->SAD. More experiments need to be carried out to systematically compare the gender-specific condition trajectories in the future.

5.6 Chapter Summary

In this chapter, a framework called Social Data-based Prediction of Incidence and Trajectory (SPIT) was developed to predict risks of medical condition incidence and trajectories using patients' social media data. Different from traditional research, this work only used publicly available patient-reported medical condition data and covers patients around the world. In this framework, a collaborative prediction model based on collaborative filtering (CF) approach is presented to predict a ranked list of future condition incidence. In addition, a trajectory prediction model and algorithm are presented to predict disease progression trajectories given a starting condition. The experimental results show that the collaborative prediction model for a condition incidence predicts future conditions with the coverage of 48% (top-20) and 75% (top-100). The trajectory model reveals each possible progression trajectory for any two conditions. The top-ranked trajectories automatically discovered the comorbidities, which were validated by medical literature. We also discussed the difference of trajectory results across patients' gender.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This dissertation presented a Social Health Analytics framework to better utilize social health data. The user-generated social health data can provide direct experience and opinions on medical conditions, treatments and insights on population health that can benefit clinical doctors, public health officials as well as patients and researchers. We addressed the following research problems in this dissertation.

Chapter 2 addressed the problem of integration of open social health data. There is a large amount of health information available for any patient to address his/her health concerns. The freely available health datasets include community health data at the national, state and community level, readily accessible and downloadable. These datasets can help to assess and improve healthcare performance, as well as help to modify health-related policies. There are also patient-generated datasets, accessible through social media, on the conditions, treatments or side effects that individual patients experience. Clinicians and healthcare providers may benefit from being aware of national health trends and individual healthcare experiences that are relevant to their current patients. The available open health datasets vary from structured to highly unstructured. Due to this variability, an information seeker has to spend time visiting many, possibly irrelevant, websites, and has to select information from each and integrate it into a coherent mental model. We will summarize our solution to this problem in Section 6.1.

Chapter 3 addressed the problem of how to utilize openly available social media data to monitor disease outbreaks with low cost. Search queries have been used to help detect disease outbreaks. However, the research on the detection of epidemics based on

search queries is limited by two factors: First, user input query terms are regarded by search engine corporations as their core assets and are not available to outside researchers. Second, user locations are not explicitly recorded. As the users enter keywords into the search engine, the queries and IP addresses are recorded. However, the IP addresses, which can be converted to actual user locations, are not easily accessible to outsiders; thus, it is difficult to develop applications which use the actual geographic locations of users. We will summarize our solution to this problem in Section 6.2.

Chapter 4 addressed the problem of public health *concern* surveillance using social data. It is critical to monitoring the spread of infectious diseases and deploying rapid responses when there is an indication of an epidemic emerging. Different surveillance strategies have been developed to meet different needs. Besides monitoring the spread of a disease itself, monitoring emotional changes of the general public, brought about by epidemics, is becoming increasingly important for public health specialists. However, for traditional public health surveillance systems, it is hard to detect and monitor health related concerns and changes in public attitudes to health-related issues. Due to their expenses, the existing surveillance methods, such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays. To supplement the current surveillance systems, a novel tool was developed. This tool tracks real-time statistics of emotions related to different health matters, such as epidemics, to provide early warning, and to help the government decision makers prevent or respond to potential social crises that might be the impact of these health-related emergencies. We will summarize our solution to this problem in Section 6.3.

Chapter 5 addressed the problem of how to mine patients' social media data to predict incidence and trajectory of their future medical conditions. Healthcare research has shown that conditions are correlated with each other, for example, in patients with type-2 diabetes, chronic nephatony often results from diabetic nephropathy. This type of correlation is called comorbidity relationship. The comorbidity relationships are often so complex that it is difficult to comprehend them. Existing research utilized electronic health records (EHRs) to predict comorbidity. However, access to EHR data is severely limited by privacy laws and is usually limited to one particular site or health network. We will summarize our solution to this problem in Section 6.4.

6.1 Social Infobuttons: Integrating Open Health Data with Social Data Using Semantic Technology

Chapter 2 discussed an approach to integrating openly available health data sources and presenting them to be easily understandable by physicians, healthcare staff and patients. The described approach enables the integration and analysis of openly available health data sources, with special attention to socially generated data. We first created a health knowledge base where data from multiple open sources is included. Data from these sources is integrated and linked via Semantic Web technology. Then, on top of the knowledge base, we developed a number of analysis tools as part of a system called "Social InfoButtons" that enable end-users to become aware of socially created health information, such as treatments, conditions, experiences, attitudes, and behaviors reported by patients, in contrast with official statistics and other "official" clinical information. We compared ranked lists of treatments and symptoms generated by the top ten conditions from Social InfoButtons against those posted by an authoritative source.

The results show a good correlation between Social InfoButtons and the authoritative source, in which the mean average precision for treatments is 0.84 and mean the average precision for symptoms is 0.72. At the same time, Social InfoButtons also returns treatments and symptoms that are not shown on the authoritative website but are often reported by patients and have been studied by some medical researchers. Case studies on two treatments, Aripiprazole and Cyclobenzaprine were reported to validate this claim.

The contributions of Chapter 2 are summarized as follows. (1) The development of a health data model. This model allows accommodating data features from many different sources. The model is health-centric and focuses on patient-generated data, such as conditions, treatments, and associated information with a focus on integrating health data from social media. At the implementation level, data is stored as RDF (Resource Description Framework) triples, which provide a) great flexibility in describing data with heterogeneous features, b) homogeneous access to data, and c) the opportunity for data linkage and semantic enrichment. (2) The provision of a process for automatic data integration and linkage. Data is automatically collected from multiple sources and transformed into RDF format. Linkage between data is accomplished via a semantic overlay that links terms from different sources that describe the same concept, enabling cross dataset references. (3) The development of an analytic and inference service focusing on medical conditions, treatments, and symptoms. We have developed a set of analytics tools that are embedded in a Web-deployed application referred to as "Social InfoButtons," providing end-users with easy access to the health knowledge base and the capability to explore and to reason with socially distributed health information.

Future work involves exploring measures to evaluate the ranked lists returned by Social InfoButtons. Besides the data sources that are already integrated, health-related information from health professionals' social networks will also be extracted and included into the existing semantic health model. Semantic search operations will be employed to improve or replace the current, embedded SPARQL queries in order to fully utilize the advantages of the Jena triple store. Currently, data collection is automatic but not in real time, so it is desirable to expand the data collection process into a batch procedure or a real-time process.

6.2 Epidemic Outbreak and Spread Detection System Based on Twitter Data

Chapter 3 discussed a social media data ETL (Extract-Transform-Load) method, to provide a user-friendly, dynamic method for analyzing outbreaks and the spread of developing epidemics in space and time. We have developed the Epidemics Outbreak and Spread Detection System (EOSDS) as a prototype system that makes use of the rich information retrievable in real time from Twitter. EOSDS provides four different analytics tool for monitoring spreading epidemics, Instance Map, Distribution Map, Filter Map, and Sentiment Trend to investigate public health threats in the space and time dimensions. (1) Instance Map displays locations of all tweet instances. Instance Map geocodes the geographic information in tweets into (latitude, longitude) coordinates that can be processed by the system. Every location is passed to the Google Geocoding server, and the returned latitude and longitude are mapped by EOSDS to show the estimated location of each tweet. (2) Distribution Map shows the number of tweets for each city, state, and country. The users' profile locations lie at different levels of granularity. The granularity of locations creates a difficulty to identify what state or city a tweet comes

from. To solve this problem, we developed a method called "two-step coding," which first geocodes text-based locations into latitude/longitude and then reversely geocodes latitude/longitude back into standard addresses, which indicate state and country of this location. (3) Filter Map provides users with a dynamic interface to monitor and analyze dynamic trends derivable from health-related tweets. Three filters are incorporated into the filter map: granularity filter, influence filter, and timeline filter. Granularity filter utilizes National Places Gazetteer to match a location to a specific granularity, which enables to select locations with different granularities. Influence filter selects users within a range of follower counts, which is helpful to find how the "influencers" are distributed over the map. Timeline filter provides an additional perspective to gain insights into the temporal distribution and development of an epidemic. (4) The Sentiment Trend contains Concern Timeline Chart and Concern Map. Through sentiment analysis, the Concern Timeline Chart is able to track the public concern trends on the timeline and the Concern Map shows the geographic distribution of concern. In experiments, we compared the results of Instance Map and Distribution Map with CDC reports during the listeria outbreak in September 2011. The Instance Map shows large clusters of tweets on the heavily affected states, such as Colorado and Texas. Among the six states with most tweets on Distribution Map, we observed that four of them also have large numbers of affected patients on CDC reports that appeared three days later. In addition, the Distribution Map made it possible to discover an unusual listeria outbreak situation in Wyoming, which was not reported by the CDC until seven days later.

Future work involves detecting rumors and their sources in tweets, since rumor tweets are able to mislead the EOSDS system. To classify rumor tweets, different

features, such as topics, temporal features, and structural features, can be used. After rumor tweets are identified, node connectivity can be calculated and the more connected a node is, the more likely the node is a rumor source. We also plan to extend the geographic processing by utilizing the above features (e.g, local words, language, and time zone) to further improve the location estimation of diseases in EOSDS system.

6.3 Twitter Sentiment Classification for Measuring Public Health Concerns

Chapter 4 explored the potential of mining social network data, such as tweets, to provide a tool for public health specialists and government decision makers to gauge a Measure of Concern (MOC) expressed by Twitter users under the impact of diseases. To derive the MOC from Twitter, we developed a two-step classification approach to analyze sentiments in disease-related tweets. We first distinguished Personal from News (Non-Personal) tweets. In the second stage, the sentiment analysis was applied only to Personal tweets to distinguish Negative from Non-Negative tweets. In order to evaluate the two-step classification method, we created a test dataset by human annotation for three domains: epidemic, clinical science, and mental health. The Fleiss's Kappa values between annotators were 0.40, 0.54, and 0.33, respectively. These moderate agreements illustrate the complexity of the sentiment classification task, since even humans exhibit relatively low agreement on the labels of tweets.

The contributions of Chapter 4 are summarized as follows. (1) We developed a two-step sentiment classification method by combining clue-based search and Machine Learning (ML) methods by first automatically labeling the training datasets, and then building classifiers for Personal tweets and classifiers for tweet sentiments. The two-step classification method shows 10% and 22% increase of accuracy over the clue-based

method on epidemic and mental health dataset, respectively. (2) We quantified the MOC using the results of sentiment classification, and used it to reveal the timeline trends of sentiments of tweets. The peaks of MOC and the peaks of NN (Non-Negative) correlated with the peaks of News with Jaccard Coefficients of 0.2 to 0.3. (3) We applied our sentiment classification method and the Measure of Concern to other topical domains, such as mental health monitoring and crisis management. The experimental results support the hypothesis that our approach is generalizable to other domains.

Future work involves the following. (1) Measure of Concern (MOC) is currently based on the number of Personal Negative tweets and total number of tweets on the same day. The Measure of Concern was used to define the fraction of tweets that are Personal Negative tweets. We plan to fine-grain this definition to quantify the number of tweets expressing real concern. (2) To improve the performance of classification, we plan to extend the current feature set to include more features specific to micro-blogs, such as slang terms and intensifiers to capture the unique language in micro-blogs. In Personal vs. News classification, we chose to work in the Machine Learning-based paradigm. However, we note that some lightweight knowledge-based approaches could possibly produce competitive results. For example, if the tweet is of the form "TEXT URL" and the TEXT appears on the web page that the URL points to, the tweet is a News Tweet. The intuition behind this approach is that the title of a news article is often pasted into the tweet body followed by the URL to that news article. We would like to perform a comparison of these knowledge-based approaches with our ML approach in the future. (3) Although it is difficult to find the ground truth for sentiment trends, we would like to conduct a systematic experiment on comparing the sentiments derived by our methods

with the epidemic cases reported by other available tools, and with authoritative data sources, such as Health Map and CDC reports. The sentiment trends for topics will also be studied by combining the sentiment analysis algorithms with topic modeling algorithms.

6.4 Predicting Incidence and Trajectory of Medical Conditions by Mining Patients' Social Media Data

Chapter 5 presented a method to predict medical condition incidence as well as its progression trajectory by utilizing publicly available patients' social media data. The experimental results show that our framework is able to predict future conditions for online patients with a coverage value of 40% for a top-20 ranked list. For trajectory risk prediction, our method is able to reveal each possible progression trajectory between any two conditions and infer the confidence of the future trajectory, given any observed condition. The predicted trajectories were validated by comparing them with the comorbidities reported in the medical literature.

Future work includes (1) Improve the tree-based model. Currently the trajectories that have highly "popular" conditions (e.g. Fibromyalgia) tend to be ranked high in terms of confidence because more conditions are paired with "popular" conditions. The presented comorbidity index mitigated this problem since it penalizes "popular" conditions. It will be desirable that the penalization can be done during the tree-model construction process. (2) Evaluation of the trajectory prediction model. The current evaluation of trajectories is based on the observation of the reported comorbidities. More systematic experiments for evaluating the quality of trajectories will be designed and performed in the future to further validate the trajectory prediction performance.

APPENDIX A

DATA COLLECTION KEYWORDS

Table A.1 shows the keywords used to collect tweets for each domain. The keywords extended the condition terms defined by U.S. Department of Health and Human Services [74].

Table A.1 Keywords for Collecting Tweets in Each Dataset

Dataset	Keywords
Epidemic	Listeria, Listeriosis, flu, influenza, h1n1, h5n1, ah1n1, adenovirus, h3n2, h3n8, h7n3, Swine Flu, Swine influenza, pig influenza, hog flu, pig flu, Swine influenza virus, swine-origin influenza virus, measles,measle, rubeola,coryza, morbilli, koplik spots, meningitis, encephalitis, meningococcal, brain infection, meningoencephalitis, meningococcus, neisseria meningitidis, mollarets, tuberculosis, tuberculose, tuberculous, mantoux test, mdr tb, bcg vaccine, phthisis, tdr tb, ebola
Mental Health	Generalized anxiety disorder, Obsessive-compulsive disorder, Obsessive-compulsive neurosis, OCD, Bipolar disorder, Manic depression, Bipolar affective disorder
Clinical Science	skin cancer, melanoma, nivolumab, IMCgp100, PV-10, lambrolizumab, T-Vec, TVEC, imatinib, methotrexate, MPDL3280A
Disaster	aircraft crash, aircraft accident, flood,tornado,earthquake,hurricane,winter storm,blizzard,tsunami,typhoon,tropical storm

APPENDIX B

INSTRUCTIONS FOR HUMAN ANNOTATION

The following are the instructions given to human annotators to label tweets.

1. Task:

Task 1: Label each tweet as Personal or Non-Personal. If the tweet is a Personal tweet, fill 1 into the PERSONAL cell. Otherwise, the tweet is a Non-Personal tweet, fill 1 into the NEWS (NON-PERSONAL) cell.

Task 2: If the tweet is labeled as PERSONAL tweet in task 1, judge whether the tweet is a PERSONAL NEGATIVE or PERSONAL NON-NEGATIVE. Fill 1 into the corresponding cell.

2. Definitions of PERSONAL and NON-PERSONAL:

A Personal tweet is defined to be one that expresses its author's private states. A private state can be a sentiment, opinion, speculation, emotion, or evaluation, and it cannot be verified by objective observation. In addition, if a tweet talks about a fact observed by the Twitter user, such as "The boyfriend is STILL sick from the @fatburger he ate last Thursday. The doctor suspects listeria. :(", this tweet is also defined as Personal. All tweets that are not Personal are defined as Non-Personal tweets.

3. Definitions of PERSONAL NEGATIVE and PERSONAL NON-NEGATIVE:

If a Personal tweet expresses negative emotions or attitude, it is a Personal Negative tweet. Otherwise, it is a Personal Non-Negative tweet. Neutral or positive tweets are both Personal Non-Negative tweets.

4. Examples of PERSONAL NON-NEGATIVE:
(1) RT @sunetrac: Narendra Modi has swine flu- i don't know why but this news is really exciting me
(2) #RememberWhen everyone had the swine flu in 7th grade "
(3) I watched that movie when I had swine flu" - guess who

5. Examples of PERSONAL NEGATIVE:
(1) no more potential skin cancer! huzzah
(2) depression is the worst.
(3) How can you rape a 14 year old tuberculosis patient???? What kinda Konji is that?
(4) @creightonkauss @professor_gram3 meningitis is a bitch

6. Examples of NEWS (NON-PERSONAL):

(1) Metformin shows promise as anti-tuberculosis drug #Pharmacy http://t.co/lUXLx5NA7R
(2) Disneyland says unvaccinated kids not welcome amid measles outbreak http://t.co/eSztH9mIy0
(3) 67 confirmed cases of measles in California-centered outbreak - LA Times http://t.co/mzokIrJdyk #SmartNews

REFERENCES

[1] S. Fox and M. Duggan. *Health online 2013*. Available: http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf (Accessed 07/03/2015)

[2] X. Ji, S. A. Chun, and J. Geller, "Social infobuttons: integrating open health data with social data using semantic technology," in *Proceedings of the Fifth Workshop on Semantic Web Information Management*, New York, NY, 2013, Article No. 6.

[3] X. Ji, S. A. Chun, and J. Geller, "Epidemic outbreak and spread detection system based on twitter data," in *Proceedings of the First international conference on Health Information Science*, Beijing, China, 2012, pp. 152-163.

[4] X. Ji, S. A. Chun, and J. Geller, "Monitoring public health concerns using Twitter sentiment classifications," in *Proceedings of IEEE International Conference on Healthcare Informatics*, Philadelphia, PA, 2013, pp. 335-344.

[5] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," *Social Network Analysis and Mining,* vol. 5, 2015, pp. 1-25.

[6] X. Ji, S. Chun, and J. Geller, "A collaborative filtering approach to assess individual condition risk based on patients' social network data," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, Newport Beach, CA, 2014, pp. 639-640.

[7] J. Sun and C. K. Reddy. *Big Data Analytics for Healthcare*. Available: http://www.siam.org/meetings/sdm13/sun.pdf (Accessed 07/03/2015)

[8] *PatientsLikeMe*. Available: https://http://www.patientslikeme.com (Accessed 07/03/2015)

[9] *Behavioral Risk Factor Surveillance System*. Available: http://www.cdc.gov/brfss/ (Accessed 07/03/2015)

[10] *PubMed*. Available: http://www.ncbi.nlm.nih.gov/pubmed (Accessed 07/03/2015)

[11] *NYC Open Data*. Available: https://nycopendata.socrata.com/ (Accessed 07/03/2015)

[12] *City of Chicago Data Portal*. Available: https://data.cityofchicago.org/ (Accessed 07/03/2015)

[13] N. Collier and S. Doan, "Syndromic classification of Twitter messages," *Electronic Healthcare,* vol. 91, 2012, pp. 186-195.

[14] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, United Kingdom, 2011, pp. 1568-1576.

[15] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic," *PLoS One,* vol. 6, 2011, p. e19467.

[16] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the First Workshop on Social Media Analytics*, Washington D.C., District of Columbia, 2010, pp. 115-122.

[17] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the Social Web," in *Proceedings of 2nd International Workshop on Cognitive Information Processing*, Elba Island, Italy, 2010, pp. 411-416.

[18] *Disease Control Priorities Project*. Available: http://www.dcp-3.org/dcp2 (Accessed 07/03/2015)

[19] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson, "Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic," in *Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, College Park, MD, 2011, pp. 18-25.

[20] M. J. Paul and M. Dredze, "A model for mining public health topics from Twitter," Technical Report, Johns Hopkins University, 2011.

[21] A. Angold, E. J. Costello, and A. Erkanli, "Comorbidity," *Journal of Child Psychology and Psychiatry,* vol. 40, 1999, pp. 57-87.

[22] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, "Defining comorbidity: implications for understanding health and health services," *The Annals of Family Medicine,* vol. 7, 2009, pp. 357-363.

[23] E. G. Willcutt, B. F. Pennington, R. K. Olson, and J. C. DeFries, "Understanding comorbidity: a twin study of reading disability and attention - deficit/hyperactivity disorder," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* vol. 144, 2007, pp. 709-714.

[24] D. A. Davis, N. V. Chawla, N. A. Christakis, and A. L. Barabasi, "Time to CARE: a collaborative engine for practical disease prediction," *Data Mining and Knowledge Discovery,* vol. 20, 2010, pp. 388-415.

[25] L. Fernandez-Luque, R. Karlsen, and J. Bonander, "Review of extracting information from the Social Web for health personalization," *Journal of Medical Internet Research,* vol. 13, 2011, p. e15.

[26] C. A. Smith and P. J. Wicks, "PatientsLikeMe: Consumer health vocabulary as a folksonomy," in *Proceedings of American Medical Informatics Association Annual Symposium*, Washington D.C., 2008, pp. 682-686.

[27] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems,* vol. 5, 2009, pp. 1-22.

[28] C. Bizer. *Evolving the Web into a Global Data Space*. Available: http://www.wiwiss.fu-berlin.de/en/fachbereich/bwl/pwo/bizer/research/publications/Bizer-GlobalDataSpace-Talk-BNCOD2011.pdf (Accessed 07/03/2015)

[29] S. Fox. *The Social Life of Health Information*. Available: http://www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf (Accessed 07/03/2015)

[30] *MedHelp*. Available: http://www.medhelp.org/ (Accessed 07/03/2015)

[31] *Open Government Initiative*. Available: http://www.whitehouse.gov/open (Accessed 07/03/2015)

[32] *WebMD*. Available: http://www.webmd.com (Accessed 07/03/2015)

[33] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems,* vol. 2, 2014, pp. 1-10.

[34] M. Househ, E. Borycki, and A. Kushniruk, "Empowering patients through social media: the benefits and challenges," *Health Informatics Journal,* vol. 20, 2014, pp. 50-58.

[35] A. Sheth and C. Ramakrishnan, "Semantic (Web) Technology in Action: Ontology Driven Information Systems for Search, Integration and Analysis," *IEEE Data Engineering Bulletin,* vol. 26, 2003, pp. 40-48.

[36] A. Harth and Y. Gil, "Geospatial data integration with linked data and provenance tracking," in *W3C/OGC Linking Geospatial Data Workshop*, 2014, pp. 1-5.

[37] L. Specia and E. Motta, "Integrating Folksonomies with the Semantic Web," in *Proceedings of the 4th European conference on The Semantic Web: Research and Applications*, Innsbruck, Austria, 2007, pp. 624-639.

[38] P. Fox, D. L. McGuinness, L. Cinquini, P. West, J. Garcia, J. L. Benedict*, et al.*, "Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience," *Computers & Geosciences,* vol. 35, 2009, pp. 724-738.

[39] S. A. Chun and B. MacKellar, "Social health data integration using semantic Web," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, pp. 392-397.

[40] B. MacKellar, C. Schweikert, and S. A. Chun, "Patient-Centered Clinical Trials Decision Support using Linked Open Data," *International Journal of Software Science and Computational Intelligence,* vol. 6, 2014, pp. 31-48.

[41] J. K. Tofferi, J. L. Jackson, and P. G. O'Malley, "Treatment of fibromyalgia with cyclobenzaprine: A meta-analysis," *Arthritis Rheumatism,* vol. 51, Feb 15 2004, pp. 9-13.

[42] *LinkedLifeData*. Available: http://www.linkedlifedata.com (Accessed 07/03/2015)

[43] J. Kozak, M. Necasky, J. Dedek, J. Klimek, and J. Pokorny, "Linked open data for healthcare professionals," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, Vienna, Austria, 2013, pp. 400-409.

[44] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association,* vol. 88, 2000, pp. 265-266.

[45] *Guidelines for ATC classification and DDD assignment*: World Health Organization, 1996.

[46] S. Nicholas, C. Sherri de, W. H. Margaret, W. H. Frank, S. Wen-Ling, and W. W. Lawrence, "NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information," *Journal of Biomedical Informatics,* vol. 40, 2007, pp. 30-43.

[47] S. A. Collins, L. M. Currie, S. Bakken, and J. J. Cimino, "Information needs, Infobutton Manager use, and satisfaction by clinician type: a case study," *Journal of the American Medical Informatics Association,* vol. 16, 2009, pp. 140-142.

[48] J. J. Cimino, J. Li, S. Bakken, and V. L. Patel, "Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users," in *Proceedings of AMIA Annual Symposium*, 2002, pp. 170-174.

[49] J. J. Cimino, G. Elhanan, and Q. Zeng, "Supporting infobuttons with terminological knowledge," in *Proceedings of AMIA Annual Fall Symposium*, 1997, pp. 528-532.

[50] J. J. Cimino, "Use, usability, usefulness, and impact of an infobutton manager," in *Proceedings of American Medical Informatics Association Annual Symposium*, 2006, pp. 151-155.

[51] J. J. Cimino, J. Li, M. Allen, L. M. Currie, M. Graham, V. Janetzki*, et al.*, "Practical considerations for exploiting the World Wide Web to create infobuttons," *Medinfo,* vol. 11, 2004, pp. 277-81.

[52] S. Attfield, A. Adams, and A. Blandford, "Patient information needs: before and after doctor consultations," *Health Informatics Journal,* vol. 12, 2005, pp. 165-177.

[53] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic acids research,* vol. 32, 2004, pp. D267-270.

[54] D. Rao, P. McNamee, and M. Dredze, "Entity linking: finding extracted entities in a knowledge base," in *Multi-source, Multilingual Information Extraction and Summarization*: Springer Berlin Heidelberg, 2013, pp. 93-115.

[55] A. Hogan, A. Zimmermann, J. Umbrich, A. Polleres, and S. Decker, "Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora," *Web Semantics: Science, Services and Agents on the World Wide Web,* vol. 10, 2012, pp. 76-110.

[56] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang, "LinkedCT: a linked data space for clinical trials," ed. arXiv preprint arXiv:0908.0567, 2009.

[57] *SPARQL Query Language for RDF*. Available: http://www.w3.org/TR/rdf-sparql-query/ (Accessed 07/03/2015)

[58] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical owl-dl reasoner," *Web Semantics: Science, Services and Agents on The World Wide Web,* vol. 5, 2007, pp. 51-53.

[59] X. Ji, P. Cappellari, S. A. Chun, and J. Geller, "Leveraging social data for health care behavior analytics," in *15th International Conference on Web Engineering*, 2015, pp. 667-670.

[60] *PHP Simple HTML DOM Parser*. Available: http://simplehtmldom.sourceforge.net (Accessed 07/03/2015)

[61] B. McBride, "Jena: Implementing the rdf model and syntax specification," in *Second International Workshop on the Semantic Web*, 2001.

[62] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Seattle, WA, 2006, pp. 11-18.

[63] *Mayo Clinic's diseases and conditions information*. Available: http://www.mayoclinic.org/diseases-conditions (Accessed 07/03/2015)

[64] J. C. Nelson, A. Pikalov, and R. M. Berman, "Augmentation treatment in major depressive disorder: focus on aripiprazole," *Neuropsychiatr Disease and Treatment,* vol. 4, Oct 2008, pp. 937-948.

[65] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature,* vol. 457, 2009, pp. 1012-1014.

[66] *Twitter*. Available: http://www.twitter.com (Accessed 07/03/2015)

[67] *Twitter developers documentation*. Available: https://dev.twitter.com/docs (Accessed 07/03/2015)

[68] H. Artman, J. Brynielsson, B. J. E. Johansson, and J. Trnka, "Dialogical emergency management and strategic awareness in emergency communication," in *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Managemen*, Lisbon, Portugal, 2011.

[69] F. Johansson, J. Brynielsson, and M. N. Quijano, "Estimating citizen alertness in crises using social media monitoring and analysis," in *Proceedings of European Intelligence and Security Informatics Conference*, Odense, Denmark, 2012, pp. 189-196.

[70] J. S. Brownstein, C. C. Freifeld, B. Y. Reis, and K. D. Mandl, "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project," *PLoS Med,* vol. 5, 2008, p. e151.

[71] C. Zhiyuan, C. James, and L. Kyumin, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, ON, Canada, 2010, pp. 759-768.

[72] A. Culotta, "Detecting influenza outbreaks by analyzing Twitter messages," in *preprint arXiv:1007.4748*, ed: arXiv, 2010.

[73] *Twitter4J*. Available: http://twitter4j.org/en/ (Accessed 07/03/2015)

[74] *NowTrending Tool by U.S. Department of Health and Human Services*. Available: https://nowtrending.hhs.gov (Accessed 07/03/2015)

[75] K. M. Carley, J. Diesner, and M. De Reno, "AutoMap User's Guide," 2006.

[76] *Google developer's guide*. Available: https://developers.google.com/maps/documentation/geocoding/ (Accessed 07/03/2015)

[77] S. A. Collins, L. M. Currie, S. Bakken, and J. J. Cimino, "Information needs, Infobutton Manager use, and satisfaction by clinician type: a case study," 20081216 DCOM- 20090127 2009,

[78] S. Mazzocchi, S. Garland, and R. Lee. *SIMILE: Practical Metadata for the Semantic Web*. Available: http://www.xml.com/pub/a/2005/01/26/simile.html (Accessed 07/03/2015)

[79] J. Perez, M. Arenas, and C. Gutierrez, "Semantics and complexity of SPARQL," *ACM Trans. Database Syst.,* vol. 34, 2009, pp. 1-45.

[80] S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwich in Glasgow: modeling locations with tweets," in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, Glasgow, UK, 2011, pp. 61-68.

[81] *Americans 'can't give in to hysteria or fear' over Ebola: Obama*. Available: http://www.reuters.com/article/2014/10/18/us-health-ebola-usa-idUSKCN0I61BO20141018 (Accessed 07/03/2015)

[82] X. Zhu, S. Wu, D. Miao, and Y. Li, "Changes in emotion of the Chinese public in regard to the SARS period," *Social Behavior and Personality,* vol. 36, 2008, pp. 447-454.

[83] Guardian. *Chinese panic-buy salt over Japan nuclear threat*. Available: http://www.guardian.co.uk/world/2011/mar/17/chinese-panic-buy-salt-japan

[84] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," 2011,

[85] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining Text Data*, 2012, pp. 415-463.

[86] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *arXiv preprint arXiv:1308.6242*, ed, 2013.

[87] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis," in *Proceedings of 1st Workshop on Emotion and Sentiment in Social and Expressive Media*, Turin, Italy, 2013.

[88] Y. Sha, J. Yan, and G. Cai, "Detecting public sentiment over PM2.5 pollution hazards through analysis of Chinese microblog," in *The 11th International Conference on Information Systems for Crisis Response and Management*, University Park, PA, 2014, pp. 722-726.

[89] D. Liben Nowell and J. Kleinberg, "The link prediction problem for social networks," *Journal of The American Society for Information Science and Technology,* vol. 58, 2007, pp. 1019-1031.

[90] J. Berger, "Arousal increases social transmission of information," *Psychological science,* vol. 22, 2011, pp. 891-893.

[91] C. Heath, "Do people prefer to pass along good or bad news? valence and relevance of news as predictors of transmission propensity," *Organizational Behavior and Human Decision Processes,* vol. 68, 1996, pp. 79-94.

[92] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior," *Journal of Management Information Systems,* vol. 29, 2013, pp. 217-248.

[93] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.,* vol. 2, 2008, pp. 1-135.

[94] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10*, Philadelphia, PA, 2002, pp. 79-86.

[95] G. Mishne, "Experiments with mood classification in blog posts," in *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, Salvador, Brazil, 2005.

[96] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," in *Proceedings of Human Language Technologies Conference on Empirical Methods in Natural Language Processing*, Vancouver, B.C., Canada, 2005, pp. 347-354.

[97] J. Brynielsson, F. Johansson, C. Jonsson, and A. Westling, "Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises," *Security Informatics,* vol. 3, 2014, pp. 1-11.

[98] H. Saif, M. Fernández, and H. Alani, "Automatic stopword generation using contextual semantics for sentiment analysis of Twitter," in *Proceedings of 13th International Semantic Web Conference*, Riva del Garda, Trentino, Italy, 2014.

[99] E. Refaee and V. Rieser, "An Arabic twitter corpus for subjectivity and sentiment analysis," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, 2014, pp. 2268-2273.

[100] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," in *The Semantic Web–ISWC*, 2012, pp. 508-524.

[101] L. Barbosa and J. Feng, "Robust sentiment detection on Twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010, pp. 36-44.

[102] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *Proceedings of the 13th International Conference on Discovery Science*, Canberra, Australia, 2010, pp. 1-15.

[103] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, Valletta, Malta, 2010, pp. 1320-1326.

[104] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon, 2011, pp. 151-160.

[105] Z. Zhou, X. Zhang, and M. Sanderson, "Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion," in *Databases Theory and Applications*, 2014, pp. 98-109.

[106] K. Fukunaga, *Introduction to statistical pattern recognition (2nd edition)*: Academic Press, 1990.

[107] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning,* vol. 20, 1995, pp. 273-297.

[108] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Interactive Intelligent Systems,* vol. 2, 2011, pp. 1-27.

[109] M. Salathe and S. Khandelwal, "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control," *PLoS Computational Biology,* vol. 7, 2011, p. e1002199.

[110] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VAS, 2006, pp. 43-50.

[111] C. Chew and G. Eysenbach, "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak," *PLoS One,* vol. 5, 2010, p. e14118.

[112] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proceedings of International Conference on Weblogs and Social Media*, Washington D.C., 2010, pp. 122-129.

[113] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, 2005, pp. 486-497.

[114] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, 2003, pp. 105-112.

[115] T. Wilson and J. Wiebe, "Annotating opinions in the world press," in *Proceedings of 4th SIGdial Meeting on Discourse and Dialogue*, Sapporo, Japan, 2003, pp. 13-22.

[116] *Profanity Filter Word List*. Available: http://web.njit.edu/~xj25/eosds_beta/files/profanity_list.txt (Accessed 07/03/2015)

[117] *Obscenity, indecency and profanity guide*. Available: http://www.fcc.gov/guides/obscenity-indecency-and-profanity (Accessed 07/03/2015)

[118] *Stopwords*. Available: http://web.njit.edu/~xj25/eosds_beta/files/news_stopwords.txt (Accessed 07/03/2015)

[119] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, Technical Report, 2009.

[120] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Languages in Social Media*, Portland, Oregon, 2011, pp. 30-38.

[121] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Exploration Newsletter,* vol. 11, 2009, pp. 10-18.

[122] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin,* vol. 76, 1971, pp. 378-382.

[123] A. Bruns and S. Stieglitz, "Twitter data: what do they represent?," *Information Technology,* vol. 56, 2014, pp. 240-245.

[124] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose," in *arXiv preprint arXiv:1306.5204*, ed, 2013.

[125] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of Twitter users," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Catalonia, Spain, 2011, pp. 554-557.

[126] *Company Recalls Several Food Products Due to Listeria*. Available: http://fox8.com/2014/03/23/several-nationally-distributed-food-products-recalled-due-to-listeria/ (Accessed 07/03/2015)

[127] *Food Fear: Hummus, Dips from Target, Giant Eagle, Trader Joe's Recalled*. Available:

http://fox8.com/2014/05/20/food-fear-hummus-dips-from-target-giant-eagle-trader-joes-recalled/ (Accessed 07/03/2015)

[128] *Stem cells shed light on treatments for bipolar disorder*. Available: http://blogs.discovermagazine.com/d-brief/2014/03/26/stem-cells-shed-light-on-treatments-for-bipolar-disorder/ - .U-wKD4BdXN8 (Accessed 07/03/2015)

[129] *Malaysia Airlines flight MH17 crash*. Available: http://www.independent.co.uk/news/world/europe/malaysia-airlines-plane-crash-boeing-jet-carrying-295-people-crashes-in-ukraine-9612882.html (Accessed 07/03/2015)

[130] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, 1977, pp. 159-174.

[131] N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: a patient-centered framework," *Journal of General Internal Medicine,* vol. 28, 2013, pp. 660-665.

[132] F. Folino and C. Pizzuti, "A comorbidity-based recommendation engine for disease prediction," in *Proceedings of the IEEE 23rd International Symposium on Computer-Based Medical Systems*, Bentley, Australia, 2010, pp. 6-12.

[133] S. Hassan and Z. Syed, "From netflix to heart attacks: collaborative filtering in medical datasets," in *Proceedings of the 1st ACM International Health Informatics Symposium*, Arlington, Virginia, USA, 2010, pp. 128-134.

[134] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesøe, R. Eriksson, H. Schmock*, et al.*, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients," *Nature communications,* vol. 5, 2014,

[135] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, 2014, pp. 85-94.

[136] *The Social Life of Health Information*. Available: http://www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf (Accessed 07/03/2015)

[137] K. Ali and W. v. Stam, "TiVo: making show recommendations using a distributed collaborative filtering architecture," in *Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle, WA, USA, 2004, pp. 394-401.

[138] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th International Conference on World Wide Web*, Banff, Alberta, Canada, 2007, pp. 271-280.

[139] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *IEEE Internet Computing,* vol. 7, 2003, pp. 76-80.

[140] P. Xia, B. Liu, Y. Sun, and C. Chen, "Reciprocal Recommendation System for Online Dating," ed. arXiv preprint arXiv:1501.06247, 2015.

[141] L. Duan, W. N. Street, and E. Xu, "Healthcare information systems: data mining methods in the creation of a clinical recommender system," *Enterprise Information Systems,* vol. 5, 2011, pp. 169-181.

[142] M. Wiesner and D. Pfeifer, "Adapting recommender systems to the requirements of personal health record systems," in *Proceedings of the 1st ACM International Health Informatics Symposium*, Arlington, VA, 2010, pp. 410-414.

[143] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," *Data Mining and Knowledge Discovery,* vol. 29, 2015, pp. 1070-1093.

[144] A. S. Hussein, W. M. Omar, X. Li, and M. A. Hatem, "Smart collaboration framework for managing chronic disease using recommender system," *Health Systems,* vol. 3, 2014, pp. 12-17.

[145] K. Hainke, J. Rahnenführer, and R. Fried, "Disease progression models: A review and comparison," Dortmund University, Technical Report, 2011.

[146] *PatientsLikeMe medical condition categories*. Available: https://http://www.patientslikeme.com/conditions (Accessed 07/03/2015)

[147] J. H. Juhl, "Fibromyalgia and the serotonin pathway," *Alternative Medicine Review,* vol. 3, 1998, pp. 367-375.

[148] *Irritable Bowel Syndrome by Mayo Clinic*. Available: http://www.mayoclinic.org/diseases-conditions/irritable-bowel-syndrome/basics/causes/con-20024578 (Accessed 07/03/2015)

[149] P. I. Papan Thaipisuttikul, P. Waleeprakhon, P. Wisajun, and S. Jullagate, "Psychiatric comorbidities in patients with major depressive disorder," *Neuropsychiatric disease and treatment,* vol. 10, 2014, pp. 2097-2103.

[150] S. Bellino, L. Patria, E. Paradiso, R. Di Lorenzo, C. Zanon, M. Zizza*, et al.*, "Major depression in patients with borderline personality disorder: a clinical investigation," *Canadian Journal of Psychiatry,* vol. 50, 2005, pp. 234-238.

[151] S.-J. Wang, P.-K. Chen, and J.-L. Fuh, "Comorbidities of migraine," *Frontiers in Neurology,* vol. 1, 2010, pp. 1-9.

[152] W. E. Whitehead, O. Palsson, and K. R. Jones, "Systematic review of the comorbidity of irritable bowel syndrome with other disorders: what are the causes and implications?," *Gastroenterology,* vol. 122, 2002, pp. 1140-1156.

[153] L. B. Weinstock and A. S. Walters, "Restless legs syndrome is associated with irritable bowel syndrome and small intestinal bacterial overgrowth," *Sleep Medicine,* vol. 12, 2011, pp. 610-613.

[154] S. S. Yarandi, S. Nasseri-Moghaddam, P. Mostajabi, and R. Malekzadeh, "Overlapping gastroesophageal reflux disease and irritable bowel syndrome: increased dysfunctional symptoms," *World Journal of Gastroenterology,* vol. 16, 2010, pp. 1232-1238.

[155] *Comorbid diagnoses: when other illnesses occur alongside an eating disorder*. Available: http://www.huffingtonpost.com/kenneth-l-weiner-md-faed-ceds/eating-disorders_b_1761513.html (Accessed 07/03/2015)

[156] W. Ho and B. M. R. Spiegel, "The relationship between obesity and functional gastrointestinal disorders: causation, association, or neither?," *Gastroenterology & Hepatology,* vol. 4, 2008, pp. 572-578.

[157] L. Khaodhiar, K. C. McCowen, and G. L. Blackburn, "Obesity and its comorbid conditions," *Clinical Cornerstone,* vol. 2, 1999, pp. 17-31.

[158] A. Romero-Corral, S. M. Caples, F. Lopez-Jimenez, and V. K. Somers, "Interactions between obesity and obstructive sleep apnea: implications for treatment," *CHEST Journal,* vol. 137, 2010, pp. 711-719.

[159] M. A. Cerullo and S. M. Strakowski, "The prevalence and significance of substance use disorders in bipolar type I and II disorder," *Substance Abuse Treatment, Prevention, and Policy,* vol. 2, 2007, pp. 1-9.

[160] *What is Bipolar with comorbid conditions?* Available: http://www.healthline.com/health-blogs/bipolar-bites/what-bipolar-comorbid-conditions (Accessed 07/03/2015)

[161] M. F. Bouchard, D. C. Bellinger, J. Weuve, J. Matthews-Bellinger, S. E. Gilman, R. O. Wright*, et al.*, "Blood lead levels and major depressive disorder, panic disorder, and generalized anxiety disorder in US young adults," *Archives of General Psychiatry,* vol. 66, 2009, pp. 1313-1319.