Mining Social Media to Understand Consumers' Health Concerns and the Public's Opinion on Controversial Health Topics

by

Yang Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2016

Doctoral Committee:

Associate Professor Kai Zheng, Co-Chair
Associate Professor Qiaozhu Mei, Co-Chair
Associate Professor David A. Hanauer
Associate Professor Joyce M. Lee

ACKNOWLEDGEMENTS

I would like to thank my advisors Kai Zheng and Qiaozhu Mei, who have been a wonderful source of support, inspiration and encouragement during my PhD program. I am greatly indebted to my committee members, David Hanauer and Joyce Lee, for their medical expertise and consistent high standard of research.

There are many other people without whom this dissertation would not have been possible: V.G. Vinod Vydiswaran, whom I have closely collaborated with and learned a lot from; Matthew Davis and Helen Levy, who brought with their public health policy perspective; Maria Woodward and Shreya Prabhu, who have generously given their time and offered ophthalmic expertise; and Jia Liu, Tricia OBrien, Esha Sondhi, and Sonia Zhang, who have helped me with enormous amount of annotation.

I am fortunate to have had many wonderful collaborators while at University of Michigan. Yan Chen, Roy Chen and Wei ai, with whom I worked closely with on a series of economic projects, have provided me with invaluable experience and knowledge of experimental economics. The Health Informatics Innovation group and Foreseer group have been a great source of ideas, feedback and friendship.

Finally, I would like to thank my parents, Aihong Cheng and Xianli Liu, for their love and continuous support.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

ADR Adverse Drug Reaction

ACA Affordable Care Act

API application programming interface

ATAM Ailment Topic Aspect Model

BRFSS Behavioral Risk Factor Surveillance System

CDC The U.S. Centers for Disease Control and Prevention

CHV Consumer Health Vocabulary

CRF conditional random field

ILI Influenza-like Illness

LDA Latent Dirichlet Allocation

LIWC Linguistic Inquiry and Word Count

MMR measles, mumps, and rubella

POMS the Profile of Mood States

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RBF radial basis function

SVM Support Vector Machine

UMLS Unified Medical Language System

ABSTRACT

Mining Social Media to Understand Consumers' Health Concerns and the Public's Opinion on Controversial Health Topics

by

Yang Liu

Co-Chairs: Kai Zheng and Qiaozhu Mei

Social media websites are increasingly used by the general public as a venue to express health concerns and discuss controversial medical and public health issues. This information could be utilized for the purposes of public health surveillance as well as solicitation of public opinions. In this thesis, I developed methods to extract health-related information from multiple sources of social media data, and conducted studies to generate insights from the extracted information using text-mining techniques.

To understand the availability and characteristics of health-related information in social media, I first identified the users who seek health information online and participate in online health community, and analyzed their motivations and behavior by two case studies of user-created groups on MedHelp and a diabetes online community on Twitter. Through a review of tweets mentioning eye-related medical concepts identified by MetaMap, I diagnosed the common reasons of tweets mislabeled by natural language processing tools tuned for biomedical texts, and trained a classifier to exclude non medically-relevant tweets to increase the precision of the extracted data.

Furthermore, I conducted two studies to evaluate the effectiveness of understanding public opinions on controversial medical and public health issues from social media information using text-mining techniques. The first study applied topic modeling and text summarization to automatically distill users' key concerns about the purported link between autism and vaccines. The outputs of two methods cover most of the public concerns of MMR vaccines reported in previous survey studies. In the second study, I estimated the public's view on the Affordable Care Act (ACA) by applying sentiment analysis to four years of Twitter data, and demonstrated that the the rates of positive/negative responses measured by tweet sentiment are in general agreement with the results of Kaiser Family Foundation Poll. Finally, I designed and implemented a system which can automatically collect and analyze online news comments to help researchers, public health workers, and policy makers to better monitor and understand the public's opinion on issues such as controversial health-related topics.

CHAPTER I

Introduction

Social media has revolutionized the way people disclose their personal health concerns and express opinions on controversial public health issues. It provides a unique platform for sharing health-related information without time and location constraints. According to a 2014 Pew Research Center survey, 74% of adults with Internet access use social media sites. (*Pew*, 2014) Another Pew report shows that 11% of social network site users, have posted comments, queries, or information about health or medical matters. (*Fox*, 2011)

In the meanwhile, both the government and individual companies have spent tremendous resources and efforts to track public health conditions,[1] risky health behaviors,[2] and public opinions on controversial public health issues[3] through personal interviews or telephone surveys. Policy makers and public health researchers rely these poll results to monitor population health and develop intervention strategies. Despite the large sample size, the traditional polling methods (*Groves et al.*, 2011) have several disadvantages including their untimeliness, high cost, and respondents' limited availability. Health-related information in social media is a valuable source of information which can be used to overcome these disadvantages. Content analysis of online discussions of controversial public health issues can generate insights about

[1] http://www.cdc.gov/nchs/nhis.htm
[2] http://www.cdc.gov/brfss/about/index.htm
[3] http://kff.org/report-section/kaiser-health-tracking-poll-april-2015-methodology/

public opinions. It can further help us estimate the tendency of public sentiment in real time with very low cost. Collections of personal health concerns expressed in social media can also be translated into effective signals of outbreak of disease epidemics in early stage. (*Ginsberg et al.*, 2009) Finally, statistical analysis of this big data set can help clinical researchers discover new medical knowledge, such as adverse drug events (*White et al.*, 2014) and disease comorbidities.

Despite these opportunities, several challenges to mining social media text have prevented us from effectively utilizing this valuable information. First, the availability and characteristics of medically-relevant data in social media remain unclear. This issue makes it difficult for researchers to determine what questions such social media data can help to answer, and the validity and generalizability of the results generated. Secondly, comparing to other traditional health information sources such as electronic health records, social media data, which could be generated by anybody on the Internet, is inherently noisy due to misspellings, casual language style, and heterogeneous contexts. Extraction of health-related information from this noisy data set can be very challenging. Careless extraction of the data can lead to false alarms of disease outbreaks or biased public opinion estimates. Finally, the lack of efficient and effective methods to analyze and make sense of social media data further impedes the full utilization of this information. Since most existing text-mining and medical natural language processing techniques are designed for processing biomedical text (e.g. clinician notes, published scientific literature), their performance on social media data is questionable without careful evaluations against human-labeled ground truth.

In this thesis, I addressed each of these three challenges respectively. First, I summarized previous work by conducting a systematic literature review of studies on understanding the motivation of online health information sharing and seeking behavior, methods of extracting and analyzing health-related information in social media, and

systems and tools leveraging such methods. I also investigated end user motivation and behaviors in two scenarios, namely user self-initiated groups in a health forum and an online diabetes community on Twitter. Second, to extract health-related information in Twitter, I applied a state-of-the-art medical natural language processing tool, MetaMap, to identify potential mentions of medical concepts. I then evaluated the performance of MetaMap by comparing the eye-related concepts it identified to the results of a manual review of a sample of tweets. Using the manually annotated sample, I trained a classifier to correct the errors introduced by MetaMap to achieve higher accuracy. Third, I applied text-mining and natural language processing techniques to study public opinions using different social media data, and demonstrated the effectiveness of these tools by comparing the machine-generated results to human-annotated data or traditional poll results. Finally, I built a system to incorporate the techniques mentioned above, and to automate the process to facilitate information extraction and insight generation using the framework I developed.

Chapter II presents a literature review of existing techniques and tools for analyzing health-related information from social media discussions. Section 3.1 in Chapter III is based on part of our work published in ICWSM 2014 (*Vydiswaran et al.*, 2014). Section 3.2 is based on unpublished work done in collaboration with Joyce Lee, David Hanauer and Qiaozhu Mei. Section 4.3 in Chapter IV is unpublished work done in collaboration with Vinod Vydiswaran, Kai Zheng, David Hanauer, Qiaozhu Mei, Trishia O'Brien, and Esha Sondhi. Section 5.1 in Chapter V is unpublished work done in collaboration with Vinod Vydiswaran, Kai Zheng, David Hanauer, and Qiaozhu Mei. Section 5.2 is ongoing work in collaboration with Matthew Davis, Kai Zheng, and Helen Levy.

CHAPTER II

Systematic Literature Review

Our goal of this chapter is to summarize prior work in health sciences and computer science pertaining to the following four topics: (1) users' motivations and concerns of sharing health-related data on social media websites, (2) methods of distilling health-related data from social media content including methods of identifying medical concepts expressed in consumer language, (3) both quantitative and qualitative methods of analyzing health-related data, and (4) frameworks and applications using health-related data.

2.1 Methods

A systematic literature review was conducted according to guidelines in the PRISMA statement. (*Moher et al.*, 2009) After consulting other health/computer science interdisciplinary literature reviews, (*Saha et al.*, 2007; *Crutzen et al.*, 2011; *Fry and Neff*, 2009; *Fernandez-Luque et al.*, 2011a), I chose to search four databases in health sciences and computer science: PubMed, WebofScience, Google Scholar, and ACM digital library. The following queries were used to search in the title and abstract fields (full text for Google Scholar) in the literature databases: health AND (twitter or tweets or facebook or myspace or youtube or "social media" or "user generated content"). The publication year must be later than 2005, and the language was limited

to English only. The eligible publications must be analysis of the content from popular social media websites instead of health-specific online communities. Furthermore, studies about the following topics were excluded: health policy research; using social media websites as a communication channel of health promotion or patient education; or health issues caused by using social media. In addition, references of relevant articles were reviewed, leading to 20 more articles being included. The PRISMA diagram is shown in Figure 2.1.

2.2 Results

2.2.1 Benefits and Concerns of Sharing Personal Health Data

Although social media has been widely adopted by all population regardless of gender, education, race, health status, or health care access, (*Chou et al.*, 2009b; *Fisher and Clayton*, 2012; *Shaw and Johnson*, 2011) understanding users' benefits and motivation of sharing their personal health data is still critical to inform future research to improve the design of social media systems and to increase their actual benefits to users.

People often communicate health information online in order to obtain experience-based information about particular treatments or behavior strategies from other patients with similar experience, seek emotional support, engage with others to make progress on their health goals, decrease a sense of isolation, and regain a sense of health by connecting with others. (*Newman et al.*, 2011b; *Ressler et al.*, 2012; *Ziebland and Wyke*, 2012; *Colineau and Paris*, 2010; *Denecke and Stewart*, 2011)

That said, there are several concerns that keep people from sharing health information online — the most prominent one being privacy. (*Newman et al.*, 2011b; *van der Velden and El Emam*, 2013; *Divecha et al.*, 2012) People tend not to share their health information with anyone except their family or close friends. Some users

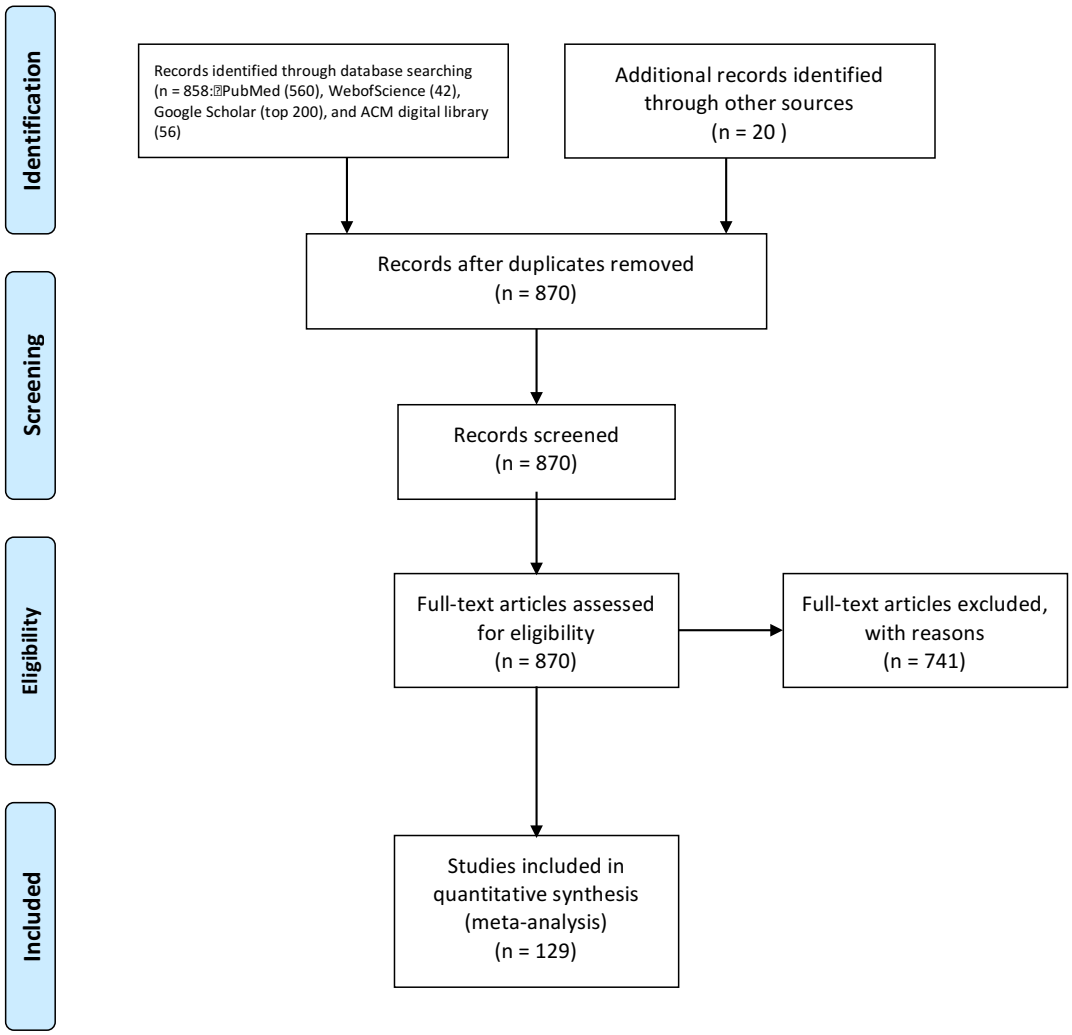


Figure 2.1: PRISMA 2009 Flow Diagram. Figure adapted from *Moher et al.* (2009).

prefer to talk about health issues in person or over the phone. Lastly, impression management is a concern of sharing health information on social media sites because people tend to express a positive and healthy identity on social media sites. (*Newman et al.*, 2011b)

2.2.2 Identifying Health-related Data

To extract health-related content from social media websites, the first step is to understand the language that laypersons use to describe their conditions — known as "consumer language" in the literature. It has been widely recognized that consumer language can be very different from health terminology. (*Smith and Wicks*, 2008; *Smith*, 2011; *Zeng and Tse*, 2006) As a result, searching using keywords in medical vocabulary such as words found in the Unified Medical Language System Metathesaurus may result in low recall of relevant information.

Zeng et al. (2007) and *Doing-Harris and Zeng-Treitler* (2011) explored different methods to develop Consumer Health Vocabulary (CHV). They first extracted candidate ngrams from query log files and medical website pages and excluded those already in the dictionaries. Collaborative human review or automated term recognition methods such as logistic regression were then applied to further filter the candidate terms. *MacLean and Heer* (2013) used a crowdsourcing-based approach to identify medical terms in patient-authored text. They recruited annotators from Amazon's Mechanical Turk to tag words/phrases relating to medical concepts from sentences in MedHelp forums. Then, a conditional random field (CRF) classifier was trained on the labeled data, which outperformed other state-of-the-art medical entity extractors, such as MetaMap and Open Biomedical Annotator.

With the appropriate keyword list in mind, one can easily search for the keywords using the search tools provided by the social media websites of interest. Examples of such search facility include Twitter's search application programming interface

(API) and YouTube's search toolbar. The keyword list is generally compiled either by domain experts or by computational algorithms. A well-developed keyword list is essential to a high recall of retrieved information.

Twitter is one of the most popular social media websites with lots of health information generated by both consumers and providers of healthcare. Various studies, for example, have used keywords such as "influenza" or "H1N1" to extract flu-related tweets. (*Aramaki et al.*, 2011; *Chew and Eysenbach*, 2010; *Culotta*, 2012; *Lampos and Cristianini*, 2010; *Quincey and Kostkova*, 2010; *Signorini et al.*, 2011) Other studies use different keywords to extract tweets about health-related issues, such as insomnia, (*Jamison-Powell et al.*, 2012) cardiac arrest and resuscitation, (*Bosley et al.*, 2013) patient safety, (*Sarah et al.*, 2012) and public health beliefs. (*Bhattacharya et al.*, 2012)

Similar strategies have also been applied to search for health-related information on YouTube. For example, *Ache and Wallace* (2008) searched for "Gardasil", "cervical cancer vaccination", and "HPV vaccination" to analyze human papillomavirus vaccination coverage on YouTube. *Carroll et al.* (2012) searched cigarette- and hookah- related videos on YouTube using four terms: "cigarettes," "smoking cigarettes," "hookah," and "smoking hookah." Many other studies have also performed keywords search to extract relevant videos on YouTube. (*Pandey et al.*, 2010; *Tian*, 2010; *Chou et al.*, 2011; *Richardson et al.*, 2011; *Richardson and Vallone*, 2012; *Seidenberg et al.*, 2012; *Fernandez-Luque et al.*, 2011b)

Carneiro and Mylonakis (2009) discussed the possibility of using Google Trends as a surveillance system to detect epidemics and diseases with high prevalences. They showed that there are strong correlations between Google Trends results of disease-related keywords and data from Centers for Disease Control and Prevention (CDC) for different diseases. Similarly, *Seifter et al.* (2010) queried Google Trends using the key phrase "Lyme disease" to study the exposure of Lyme disease across different

seasons. Without access to Google search log, *Eysenbach* (2006) bid keywords "flu" and "flu symptoms" on Google Ads and used a Google service Adsense to track the number of ad views as a proxy of keywords search volume.

Keyword search has also been applied to identify Facebook user groups. *Ahmed et al.* (2010) used the search term "concussion" to identify Facebook groups related to concussion. *Bender et al.* (2011) used the term "breast cancer" to look for Facebook breast cancer groups. *Freeman and Chapman* (2010) searched for tobacco promotion groups Facebook. *Farmer et al.* (2009) used medical and consumer terms for the most prevalent non-communicable diseases and identified 757 medical-related groups on Facebook.

In addition to compiling keywords manually, many studies formulate search queries using computational algorithms. *Ginsberg et al.* (2009) designed an automatic method of selecting Influenza-like Illness (ILI)-related queries without knowledge about influenza. They explored 50 million candidate queries in Google search log and selected a set of queries with highest correlation with CDC ILI data. Similarly, *Lampos and Cristianini* (2010) tried to find a weighted set of keywords to calculate flu score and maximize its correlation with the official reported flu rates. Consumer Health Vocabulary (*Yang et al.*, 2012) and metaphorical relations (*Neuman et al.*, 2012) are also leveraged to automatically construct lexicons of interest.

Due to its simplicity, extracting health information using keyword searches may include non-relevant examples. For example, tweets mentioning "Bieber fever" will be extracted by the keyword "fever", which are not medically relevant. As a result, surveillance systems based on simple keyword searches can be vulnerable to false alarms due to spurious keywords matches. (*Culotta*, 2012; *Krieck et al.*, 2011) To increase the precision of the extracted information using keyword search, supervised classification methods are commonly applied to further filter out irrelevant information. Standard annotation process is first applied to create training datasets. (*Ara-*

maki et al., 2011; *Sofean et al.*, 2012) Annotators can be either recruited offline or using online crowdsourcing services such as Amazon Mechanical Turk. (*Lamb et al.*, 2012) Supervised classifiers, such as SVM, Naïve Bayes, Maximum Entropy, and decision tree, are then employed to filter health-related information from the keyword search results. (*Aramaki et al.*, 2011; *Culotta*, 2012; *Sofean et al.*, 2012; *Lamb et al.*, 2012; *Collier et al.*, 2011; *Xu et al.*, 2012; *Bian et al.*, 2012) Semi-supervised classifier can also be leveraged when there is not enough training data. (*Sadilek et al.*, 2012b) Features used in the classification process include linguistic features, such as unigram, bigram, and part of speech; (*Xu et al.*, 2012) regular expression; (*Collier et al.*, 2011) as well as medical features such as ontological/semantic features discovered by MetaMap. (*Bian et al.*, 2012)

2.2.3 Analyzing Health-related Data

A variety of methods have been applied to analyze health-related data in social media. In this chapter, I categorize them into two broad categories: manual content analysis mostly performed in health sciences and various computational methods applied in the computer science domain.

To investigate health-related issues using social media data, studies published in health sciences mostly focus on understanding the topics of the content and categorizing the characteristics of the users. Manual content analysis is frequently performed to learn knowledge from the content filtered by keyword search method mentioned in the previous section. The general steps of manual content analysis are described in *Bernard* (2012) as follows: create a coding scheme, apply the codes systematically to a set of documents, test the inter-coder reliability, count the numbers of codes for each document, and finally, analyze the counts with statistical methods.

Various studies have applied manual content analysis to examine text data, such as posts and user profiles on Facebook, (*Ahmed et al.*, 2010; *Bender et al.*, 2011; *De la*

Torre-Diez et al., 2012; *Egan and Moreno*, 2011; *Greene et al.*, 2011; *Moreno et al.*, 2011; *Villiard and Moreno*, 2012) Twitter, (*Chew and Eysenbach*, 2010; *Jamison-Powell et al.*, 2012; *De la Torre-Diez et al.*, 2012; *Dumbrell and Steele*, 2012; *McNeil et al.*, 2012; *Prochaska et al.*, 2012; *Scanfeld et al.*, 2010; *Sullivan et al.*, 2012; *Heaivilin et al.*, 2011; *Kendall et al.*, 2011; *Lyles et al.*, 2013) MySpace, (*Keelan et al.*, 2010; *Moreno et al.*, 2007, 2009) and blogs (*Marcus et al.*, 2012; *Gruzd et al.*, 2012; *Lynch*, 2010; *Simunaniemi et al.*, 2011) about mental health issues, chronic diseases, medications and fitness issues. Standard content analysis processes were also applied to analyze who posted the content, what motivated them to post or create groups, what topics they covered and what the characteristics of their comments were. Researchers have also performed content analysis to interpret video content on YouTube. *Carroll et al.* (2012), for example, compared cigarette- and hookah-related videos on YouTube. *Pandey et al.* (2010) studied YouTube as an information source about H1N1 pandemic. *Briones et al.* (2012) assessed videos related to the human papillomavirus vaccine, *Tian* (2010) analyzed organ donation videos, and *Richardson et al.* (2011) examined smoking cessation videos on YouTube. Similar methods have been applied to analyze videos and comments about tumors, (*Clerici et al.*, 2012) tobacco brands, (*Elkin et al.*, 2010) social support, (*Frohlich and Zmyslinski-Seelig*, 2012) exercises, (*Stephen and Cumming*, 2012) and personal health information (*Fernandez-Luque et al.*, 2009) on YouTube.

Researchers have also performed content analysis to interpret video content on YouTube. *Carroll et al.* (2012) compared cigarette- and hookah-related videos on YouTube. They investigated positive and negative associations between smoking and major content type in the videos. Three trained qualitative researchers used an iterative approach to develop and refine definitions of the coding of variables. Two of them then coded all videos. Finally, they counted the number of videos that contained each code and applied statistical tests to compare the differences between cigarette- and

hookah-related videos on YouTube. Similarly, *Pandey et al.* (2010) studied YouTube as an information source about H1N1 pandemic. *Briones et al.* (2012) assessed videos related to the human papillomavirus vaccine, *Tian* (2010) analyzed organ donation videos, and *Richardson et al.* (2011) examined smoking cessation videos on YouTube. Similar methods have been applied to analyze videos and comments about H1N1 pandemic, (*Pandey et al.*, 2010) vaccine, (*Briones et al.*, 2012) tumors, (*Clerici et al.*, 2012) organ donation, *Tian* (2010) tobacco brands, (*Elkin et al.*, 2010) smoking cessation, (*Richardson et al.*, 2011) social support, (*Frohlich and Zmyslinski-Seelig*, 2012) exercises, (*Stephen and Cumming*, 2012) and personal health information (*Fernandez-Luque et al.*, 2009) on YouTube.

Among various computational methods, regression analysis are commonly used to predict the outbreak of epidemics. Many studies have attempted to leverage big data of user-generated content to accurately detect the outbreak in a timely fashion. (*Chew and Eysenbach*, 2010; *Culotta*, 2012; *Lampos and Cristianini*, 2010; *Signorini et al.*, 2011; *Ginsberg et al.*, 2009; *Culotta*, 2010; *Achrekar et al.*, 2011; *Bilge et al.*, 2012; *Chunara et al.*, 2012; *Corley et al.*, 2010; *Lampos et al.*, 2010; *Ritterman et al.*, 2009; *Sofean and Smith*, 2012) For example, *Ginsberg et al.* (2009) estimated the percentage of ILI-related physician visits by the fraction of ILI-related search query submitted to Google in the same region. They simply fit a linear model using the log-odds of an ILI physician visit and the log-odds of an ILI-related search query. Following Ginsberg et al., *Culotta* (2010) applied a similar linear model using Twitter data. They calculated the fractions of tweets containing different keywords and experimented with the Multiple Linear Regression model using tweet fractions as independent variables. More advanced regression techniques, such as auto regression (*Achrekar et al.*, 2011) and Support Vector Regression, (*Culotta*, 2012; *Signorini et al.*, 2011; *Mogadala and Varma*, 2012) were also used to predict the trend in the data.

Sentiment analysis is a well-established field in natural language processing and

many sophisticated methods have been developed to gauge sentiment from text. (*Pang and Lee*, 2008; *Liu*, 2012) Due to the brevity of social media content and generally the lack of training data, most sentiment analysis studies of social media content applied simple techniques, such as counting sentiment words in a lexicon or off-the-shelf classification methods using unigram and bigram features. *Golder and Macy* (2011) used data from Twitter to identify individual-level diurnal and seasonal mood variations across the globe. For each individual Twitter user in their sample, they measured positive affect and negative effect using Linguistic Inquiry and Word Count (LIWC)) lexicon. Similarly, *Jamison-Powell et al.* (2012) used LIWC lexicon to analyze the positive and negative sentiments in tweets about insomnia. *Tausczik et al.* (2012) accessed the anxiety, health, death, and positive words used in blogs mentioning "swine flu" with LIWC. *Kramer* (2010) looked at "Gross National Happiness" by analyzing the use of emotion words defined in LIWC lexicon in millions of Facebook status updates. *Bollen et al.* (2011) calculated six dimensions of mood in tweets using the extended version of the the Profile of Mood States (POMS). *Mohammad* (2012) extracted a Twitter emotion corpus by searching tweets with emotion hashtags.

Exploratory analysis using topic models can be considered as an automatic way of performing manual thematic analysis. Many studies directly applied one of the most popular topic model Latent Dirichlet Allocation (LDA) and had domain experts interpret meaningful topics to demonstrate the effectiveness of this method. (*Xu et al.*, 2012; *Prier et al.*, 2011) Other studies leveraged LDA output as features to help classification or ranking tasks. (*Diaz-Aviles and Stewart*, 2012; *Diaz-Aviles et al.*, 2012) Unlike most others, *Paul and Dredze* (2012) developed their own topic model called "Ailment Topic Aspect Model (ATAM)" to discover different ailments and learn symptom and treatment associations from tweets. In their model, each topic is a disease, which contains general words as well as specific symptoms and treatments. The evaluation results showed that ATAM produces more unique and accurate ail-

ments than LDA. ATAM also produced more coherent topics than LDA for topics labeled as same ailments. In another paper (*Paul and Dredze*, 2011), they introduced an extension model of ATAM, ATAM+, by incorporating prior knowledge of disease from WebMD.com into the model.

Spatiotemporal analysis is also appropriate to be applied to user-generated content since most of the datasets have both timestamp and location information. For example, tweets have a timestamp indicating when they were posted, with location data of where the person posted the tweet was located. (*Burton et al.*, 2012b) Similarly, search query logs have information about when and where a query was submitted. Most health-related spatiotemporal studies using social media data have focused on disease surveillance. *Ginsberg et al.* (2009) aggregated Google search logs to compute time series of normalized ILI-related search query frequencies in different regions of interest. They found the query frequencies highly correlated with time series of ILI rate reported by CDC. Other studies have performed similar analysis using Twitter data. (*Lampos and Cristianini*, 2010; *Culotta*, 2010; *Paul and Dredze*, 2011) Besides aggregated analysis at the population level, *Sadilek et al.* (2012b,a); *Sadilek and Kautz* (2013) performed spatiotemporal analysis at the individual level. They looked at individuals who frequently posted geo-tagged tweets in New York within a one-month period. Based on different information such as co-location and Twitter friendship, they modeled the health state of individuals on a given day using conditional random field and a regression decision tree.

Association mining is a well-studied method in data mining field to analyze relationships between different items in an itemset. *Nikfarjam and Gonzalez* (2011) and *Yang et al.* (2012) both used association mining to extract associations of drugs and Adverse Drug Reaction (ADR) from social media content. Nikfarmjam and Gonzales applied Apriori algorithm to identify frequent language patterns with a mention of an adverse drug reaction. Yang et al. first extracted potential adverse drug reac-

tions from threads about specific drugs using an ADR lexicon. For each drug and ADR association, they computed lift and leverage as indicators and showed that they were both effective signals to detect ADR.

Finally, social network analysis has also been broadly applied to study health-related information networks, user networks or web link networks in social media sites. *Murthy et al.* (a,b) proposed methods to visualize and analyze cancer-related social media networks in order to understand the information flow and characteristics of widely spread health information. *Burton et al.* (2012a) illustrated methods to visualize and analyze public health communities of videos, authors, subscribers, and commenters on YouTube. *Gruzd et al.* (2012) analyzed the web link network of diabetes blogs. They identified the most influential blogs based on in-degree centrality and observed homophily in the network.

2.2.4 Systems and Applications

Systems and applications using social media content as input are primarily for disease surveillance. Traditional web-based infectious disease surveillance systems, such as HealthMap, (*Brownstein et al.*, 2008) EpiSPIDER, (*Herman Tolentino et al.*, 2007) BioCaster, (*Collier et al.*, 2008) and GPHIN, (*Mawudeku and Blench*, 2006) integrate data from electronic resources including online news aggregators (Google News, Factiva[1]), expert-curated accounts (ProMED-mail), surveillance reports (Eurosurveillance), and official alerts (e.g. from WHO). (*Keller et al.*, 2009; *Lyon et al.*, 2012) They generally aggregate information based on location, venue or other categories and provide relevant articles or send notifications to users of interest. As a different method, many epidemic surveillance systems were recently designed to rely more on social media content. (*Eysenbach*, 2009; *Dreesman and Denecke*, 2011; *Kamel Boulos et al.*, 2010; *Salathe et al.*, 2012) *Chen et al.* (2010) designed a sys-

[1]http://new.dowjones.com/products/factiva/

tem called SNEFT (Social Network Enabled Flu Trends). SNEFT can collect and aggregate online social networks data, extract information from the data, and integrate the information with mathematical models of influenza. *Denecke et al.* (2012) claimed that existing disease surveillance systems relying on certain indicators might fail when confronted with new emerging agents like the agents that caused SARS in 2002. They designed the M-Eco system to complement traditional surveillance systems for early detection of emerging threats. (*Smrz and Otrusina*, 2011) *Kanhabua et al.* (2012) presented an analytics tool for supporting a comparative, temporal analysis of disease outbreaks between Twitter and official resources like WHO. There are also other systems using social media sites to identify emerging trends in recreational drug use (*Deluca et al.*, 2012) and gather real-time topics of interest of a health industry. (*Steele and Min*)

2.3 Summary

The availability of large volumes of social media data provides an opportunity for researchers and healthcare and public health professionals to reconsider answering existing and emerging questions from new perspectives. My systematic literature review shows that using social media as a resource to study health-related problems has been increasingly popular. Many methods have been developed and applied to identify and analyze such datasets. Specifically, extracting health-related social media content can be separated into two steps: searching for information with certain keywords; and filtering irrelevant information using trained classifiers. To achieve accurate and inclusive information, traditional natural language processing and data mining techniques should be applied with domain knowledge. Both qualitative and quantitative methods have been applied to analyze health-related data in social media. Appropriate methods should be adopted depending on the goal of the study and the characteristics of the data. Accurately filtering relevant information and aggre-

gating multiple data sources are critical to enabling accurate and timely analysis. So far, disease surveillance systems are the most successful application leveraging social media data.

CHAPTER III

Health-related User Behavior in Social Media

Understanding who frequently participate in health-related discussions in social media and the reasons of their participation can inform us the characteristics of health-relevant data in social media and the research questions we can answer with such data. In this chapter, I motivation and behaviors in participating in social media discussions in two scenarios – user self-created groups in an online health forum, MedHelp[1], and diabetes conversations on Twitter.

Different from traditional online health forums, MedHelp allows end users to create their own groups in addition to the medical communities provided by the site. These user-initiated groups provide us a unique opportunity to understand online users' health-related informational and emotional needs from their own perspective. In section 3.1, I investigate users' motivations to participate in online health discussions through content analysis of the descriptions of the user-created groups on MedHelp.

Another type of participants participating in health-related discussions on social media are patients and caregivers of chronic disease. They leverage social media to exchange personal health information and foster emotional support. (*Greene et al.*, 2011) In section 3.2, I explore diabetes conversations on Twitter and characterize a community focused on diabetes over a two-year span.

[1] http://www.medhelp.org

3.1 User Self-created Groups on MedHelp

3.1.1 Introduction

Health forums are popular venues frequented by patients and caregivers seeking information and support. For example, in a 2010 poll (*Capstrat*, 2010), 37% of Internet users rated online health forums as somewhat or extremely reliable sources of health information. Research studies have shown that these forums play an instrumental role in facilitating exchanges of health information and/or emotional support (*Wang et al.*, 2012; *Chou et al.*, 2009a) and that their use is associated with higher degrees of patient empowerment (*van Uden-Kraan et al.*, 2009).

In the most prevalent scenario, owners/designers of a health forum provide a predefined list of user communities, often organized around certain medical conditions, ailments, or treatment procedures. Users can join, post questions, and respond to others' questions in these communities. In addition to these site-defined communities, some health forums also allow users to form new communities of their own, on the fly, which creates a whole new paradigm of how users of health forums bond and communicate.

Studying the motivations of such user-initiated communities (hereafter referred to as "user-created groups")[2] can help us better understand the informational and emotional needs from users' perspective. Based on an empirical dataset collected from MedHelp.org, a premier online health forum, we studied the attributes of user-created groups, including reasons leading to their creation.

3.1.2 Data Description

Established in 1994, MedHelp (sometimes referred to as "the Website" in this paper) is one of the earliest and most well-known online forums dedicated to supporting

[2]Note that user communities in online health forums may be called forums, boards, user groups, etc. For simplicity, in this section, we refer to them uniformly as "groups."

user-driven discussions on health or healthcare related topics. The Website had over 12 million registered users when this work was conducted. It has been the subject of study in several prior research endeavors (*Gill and Whisnant*, 2012; *Chuang and Yang*, 2012; *Hagan 3rd and Kutryb*, 2009).

In February 2013, we collected a complete set, i.e. all forum posts and all user profiles, from the Website. The dataset consists of over six million messages (both questions and comments in response to questions) posted by over a million unique users in about 1.4 million threads. From the user profiles collected, we extracted the friendship links among the users as well as their group membership if they had explicitly joined certain user groups.

There are five distinct types of user communities in MedHelp: (1) medical support communities, (2) "ask-a-doctor" forums, (3) forums on pets, (4) international forums, and (5) user groups. In this section, we focus on the medical support communities and the user groups, both of which aim to facilitate interactions among patients and caregivers. The medical support communities are designed and provided by MedHelp (hence "site-defined groups"); whereas the user groups are support communities initiated by end users ("user-created groups").

The dataset analyzed in this study includes all posts in a total of 270 site-defined groups and 747 user-created groups, in addition to the profiles of $1,007,570$ users who had posted at least one message in these groups. Among these users, $9,544$ were members of at least one user-created group and $502,269$ were members of at least one site-defined group. A total of $130,605$ friendship links existed among all users, out of which $113,273$ (86.7%) are between users in our dataset.

3.1.3 Categorizing User Communities by Purpose

The first question we are interested in investigating is why users choose to create so many groups even though the Website has already provided a comprehensive list

of well attended groups. In order to understand the potential motives, we conducted a qualitative content analysis of the data to categorize the user-created groups based on their stated purposes. Two authors first individually analyzed a random sample of a hundred user-created groups to derive a primitive list of categories according to the descriptions of the groups. Then, through a consensus development process, the two lists were reconciled and merged to produce a final categorization scheme. The analysis resulted in ten categories, as listed below.

1. **Specific conditions (Cond)**: Communities related to particular conditions, ailments, or diseases. This includes addictions or rare diseases that may not have an established cure or understanding. Examples include "Arachnoiditis sufferers," "Granulomatous mastitis," and "Vertigo."

2. **Specific treatment (Trmt)**: Communities related to particular treatments and procedures, including conditions that arise from a specific treatment or procedure. Examples include "Mirena IUD side effects support group," "Methadone community," and "Natural health" (a group about alternative medicines).

3. **Recovery (Rcvy)**: Communities related to the process of recovering after a completed treatment regimen, including addiction recovery, smoking cessation, etc. Examples include "Addiction recovery group," "Recovery after vitamin D deficiency," and "Heart surgery recovery."

4. **Family support (Fam)**: Communities related specifically to family members or caregivers of patients or others suffering from specific conditions. Examples include "ADHD parents," "Alzheimer's caregivers," and "Family members of prisoner."

5. **Socializing (Soc)**: Communities created primarily to host social interactions, mostly in a non-medical context, such as chit-chats, specific interests or hobbies,

discussions of current events, etc. Examples include "Prayer group," "All about TV shows and movies," and "Dinner table."

6. **Public policy (Pol)**: Communities related to governmental agencies, the economy, or public policies, including healthcare and insurance. Examples include "Social security disability or SSDI," "Ideas for economic living," and "FDA recalls, US food and drug administration."

7. **Pregnancy (Preg)**: Communities broadly related to pregnancy, including attempting to conceive, conditions and complications during pregnancy, and post-pregnancy care for the mothers and babies. Examples include "Trying to conceive after 40," "Pregnancy after tubal ligation surgery," and "March 2011 babies."

8. **Goal-oriented (Goal)**: Communities related to specific health-related goals, including weight loss, healthy diet, etc. This category also includes communities where members could track each other's health or behavior change progress. Examples include "HCG protocol group," "Diet ideas," "Weight gain," and "The 10% club."

9. **Specific demographics (Dem)**: Communities that target towards specific demographic groups. This category is further classified into the following five subcategories:

   (a) *Gender*: Communities targeted towards users of a specific gender. Examples include "Boy problems" and "Christian women with bipolar disorder."

   (b) *Location*: Communities targeted towards users in a specific location. Examples include "California" and "Problems with children and young people services in the UK."

(c) *Age*: Communities targeted towards users of a specific age group. Examples include "TTC over 40" (TTC is a commonly used abbreviation in health forums for trying to conceive) and "ADHD teens and young adults."

(d) *Profession*: Communities targeted towards users of a specific profession. Examples include "College students" and "The doctors."

(e) *Others*: Communities targeted towards users of a particular demographics not included above, such as marital status (e.g. "Mothers and the balance of a stressful life," a group for single working mothers), race (e.g. "Native American / Canadian circle"), etc.

10. **Miscellaneous (Misc)**: Communities that do not fall in any of the above categories, including the ones with an unclear purpose (missing or vaguely described), or those advocating for a particular business organization (e.g. a specific law firm).

Note that the categories described above are not mutually exclusive and therefore a community could be classified under multiple categories. Overall, 79 (10.6%) user-created groups and 15 (5.6%) site-defined groups were classified into more than one category. Nonetheless, as all pregnancy-related communities were likely gender-specific by nature, we did not include them again in the "Specific demographics / Gender" category. Communities with non-English titles and descriptions were all classified as "Miscellaneous."

Inter-rater agreement The inter-rater agreement is high between the two authors who were involved in classifying the user groups. For user-created groups, the two-rater, ten-category Cohen's kappa coefficient is 0.745. Out of the 747 user-created groups, the two raters agreed on the categorization of 600 of them (80.3%). For site-defined groups, the inter-rater agreement is even higher, with the Cohen's $\kappa = 0.854$.

Category	Site-defined $n = 270$	User-created $n = 747$
1. Specific conditions	173 (64.1%)	260 (34.8%)
2. Specific treatment	15 (5.6%)	53 (7.1%)
5. Socializing	16 (5.9%)	150 (20.1%)
7. Pregnancy	40 (14.8%)	122 (16.3%)
9. Demographics	20 (7.4%)	81 (10.8%)
10. Miscellaneous	3 (1.1%)	77 (10.3%)
Other categories (3,4,6,8)	17 (6.3%)	93 (12.5%)

Table 3.1: Distributions of the categories of site-defined and user-created groups. Note that the columns do not add to 100% since groups may have multiple labels.

3.1.3.1 Category distributions

Table 3.1 summarizes the distributions of the ten categories of site-defined and user-created groups as defined in Sec. 3.1.3. Among site-defined groups, the two most frequent categories are those related to "specific conditions" (class 1, ∼64%) and "pregnancy" (class 7, ∼15%). We also note that only 1.1% of the site-defined groups are categorized as "miscellaneous." This also suggests that the classification scheme works well for site-defined groups, although it is derived based on user-created groups.

The category distribution of user-created groups presents some interesting difference from the site-defined groups. As in site-defined groups, the most salient category of user-created groups continues to be "specific conditions" (class 1), but the percentage drops to about 35%.

The second most popular category is related to "socializing" (class 5), constituting 20% of user-created groups. This indicates a significant use case of user-created groups, i.e., to find "friends" and engage in casual conversations about things that may or may not relate to health, rather than purely seeking information about their medical conditions and treatments. Indeed, we notice that typically, such groups are related to hobbies, discussion around current news, religion (prayer groups), or recreational activities (such as a group on creative writing), which would take one's

mind away from the pain and suffering.

Pregnancy-related groups constitute ∼16% of user-created groups, forming the third largest category. As we see from Table 3.1, the proportion is similar to those related to pregnancy in site-defined groups. Our analysis shows that user-created groups related to pregnancy focus more on particular demographics such as age, specific pregnancy-related complications, or questions on parenting and childcare.

About 11% of the user-created groups in our data set are created for specific demographic classes. Many of these groups (∼69%) are also related to a specific condition, pregnancy, or for socializing. For example, there are groups for patients from a specific age-group and location (e.g. a group intended for diabetic teens in Michigan), or for teenagers who suffer from autism but want to socialize and share their experience with each other. These interesting surface statistics motivate us to dig deeper into the actual reasons why the users create groups.

3.1.4 Why Users Create New Groups?

In general, we find that homophily is one of the driving factors underlying user-created groups. The primary purposes of user-created groups are to socialize and connect with other patients suffering from similar conditions, especially having similar interests or demographic profiles. Based on our analysis, we are able to identify four primary reasons why users initiate new groups given the existing ones. Users tend to create new groups

1. to form communities specific to rare and complicated conditions and new treatments;
2. to communicate with peers with similar demographics;
3. to build or maintain social relationships;
4. without checking if similar groups exist, thus creating duplicated groups.

Reason 1. Users create communities for rare, complicated, more specific, or more general conditions: One of the main reasons users create new groups is to connect with other patients suffering from similar conditions. These conditions are rare or complicated so that they are either not covered by the site-defined groups, covered by a site-defined group that is designed for a much more general condition, or partially covered by multiple groups. We find numerous instances where founders, diagnosed with a rare disease, reached out to others to share their symptoms or experience dealing with a particular diagnosis, and in some cases to educate others who might suffer from similar conditions.

To quantify this behavior, we analyze the text descriptions of the groups to look for motivations behind initiating a new group. We search for phrases such as "I decided to start a group ..." or "I created this group ..." in the group descriptions. We find 85 groups with such statements, among which the founders of 16 groups (19%) explicitly claimed that the absence of a site-defined group that is specific to their conditions was their motivation to initiate a new group.

For instance, the founder of a mental illness group with 321 members, explained the reason to form a new group as:

> "I noticed there are many groups for each illness. So I thought of an idea to have a group where we can talk about many different mental illnesses, share our stories, support each other, and hopefully make friends."[*sic*]

A specific sub-category of such user-created groups that are prevalent on MedHelp are those related to complications in pregnancy. We find multiple groups related to trying to conceive (especially with age-specific focus), and for women who are in similar stages of pregnancy. Further, there are many groups related to complications or specific conditions related to pregnancy, such as multiple births, multiple pregnancies, or pregnancy after contraception procedures, which are not the focus of any of the site-defined groups.

For instance, the founder of the group on low progesterone with 38 members stated:

> "After seeing millions of postings on MedHelp and elsewhere about women having low progesterone, many having one or more miscarriages as a result, I thought there should really be a group focused on low progesterone." [*sic*]

Reason 2. Users create groups to find peers with similar demographics: Looking further at the user-created groups, we observe that even when some conditions have a corresponding site-defined support group, users still initiate groups about these conditions, but dedicate them to particular demographic groups. For instance, the MedHelp data set includes multiple demographic-specific sub-communities of ADHD patients and caregivers, viz. "ADHD teens," "ADHD seniors," and "parents of ADHD children." Peers in the same demographic group may naturally better understand their situations and better communicate with each other. This is related to the principle of homophily in social science.

Analyzing the group descriptions uncovers explicit reasons from founders supporting this reason. For instance, the founder for the group for young mothers (with 45 members) stated:

> "For moms ages 16-25. I was a mom at the age of 16 and I really wanted a group my own age to go to! There isn't one right now so I thought I would start one!"

Reason 3: Users create groups to build or maintain social relationships: Following the principle of homophily, founders tend to form groups to build social relationships with other users similar to them in various aspects, not just in terms of demographics but also including hobbies or interests, religious beliefs, or health goals.

For example, the founder of a group on creative writing (with 58 members) stated:

> "This group is for people to write a journal, or thoughts to be viewed by all of us ... By sharing our writings we will grow closer ..."

We also observe that founders initiated groups to stay connected with members in another group that they are part of. For example, a founder created two groups ten months apart from each other – one on "march 2011 babies" (with 64 members, created in July 2010) and the other on "march 2011 moms" (created in May 2011, with 10 members) to continue in touch with members of the former group and invite other new mothers.

Reason 4. Users create new groups without the knowledge of existing groups with similar objectives: One of the concerns with allowing users to create groups is that there might be many duplicated groups on the same or very similar topics. This leads to fractured communities, since interested members might get split between two similar groups.

3.2 Exploring Diabetes Conversations and Social Media Participation of a Diabetes Community on Twitter

3.2.1 Introduction

A growing community of patients with diabetes, as well as healthcare providers and health professionals and organizations, are participating in discussions and information exchange on online social media platforms. For example, there are specific hashtags on Twitter, such as #dsma, which stands for "diabetes social media advocacy", which are used by communities of patients to find each other, engage in virtual communication and information sharing, and find peer support online. Finally, healthcare professionals and organizations are also starting to participate on social media to increase communication with patients. (*Von Muhlen and Ohno-Machado*, 2012)

The objective of this study was to provide an overall view of diabetes conversations on Twitter by describing the frequency, and geographical origins of diabetes-related tweets, identifying authors relationship to diabetes, and characterizing a specific community of individuals on Twitter over a 2-year span.

3.2.2 Identifying Diabetes Conversations on Twitter

We used a dataset that contains 10% of all tweets in year 2013 and 2014 collected through the Twitter stream application programming interface (API) with Gardenhose access (secured through a formal agreement with the University of Michigan School of Information). We identified tweets with diabetes-related search terms and hashtags based on suggestions from providers and patients in the diabetes community. We used the following query terms and hashtags: "glucose", "blood glucose", "diabetes", "insulin pump", "insulin", "#diabetes", "#t1d", "#type1diabetes", "#type1", "#t2d", "#type2diabetes", "#type2", "#bloodsugar", "#dsma" ("diabetes social

media advocacy", an online advocacy group which holds a weekly "tweetchat"), "#doc" (diabetes online community), "#bgnow" ("blood glucose now", in which individuals share their blood sugars), "#wearenotwaiting" (a phrase coined related to need for rapid access to technology solutions in the diabetes community, "#showmeyourpump" (a tweet campaign that occurred when a Miss American Contestant decided to wear her insulin pump in public), "CWD2014" ("children with diabetes", a diabetes conference for children and families with diabetes), "dblog"(diabetes blog) and "diyps" (a "do it yourself artificial pancreas" project). For each tweet retrieved, we extracted its text content, the username of the tweet, the timestamp, the geo-location information of the tweet if available, and whether the tweet is a retweet. We assessed the frequencies with which the retrieved tweets are used across different terms and hashtags. With the timestamp information, we examined the volume of extracted tweets in each month along with monthly and weekly volume distributions. We conducted an analysis for all users with a diabetes-related tweet, as well as for those who posted at least once with the #dsma hashtag.

3.2.3 Categorizing Users' Relationship to Diabetes

We then examined the identities of two subsets of users. We randomly sampled 500 users from the entire dataset. There were 1,424 individuals who had tweeted at least once with the hashtag #dsma; we chose to focus on a smaller subset, those who had tweeted at least three times with hashtag #dsma, because it would identify more active members of the community and it represented a sample similar in number (n=416). A medical student reviewed each of the Twitter profiles evaluated to identity individuals' relationship to diabetes, which was categorized into one or more of the following 15 categories: physician, nurse, dietitian, diabetes educator, researcher, an individual with type 1 diabetes, an individual with type 2 diabetes, an individual with diabetes not specified, the caregiver/parent/guardian of an individual

with diabetes, a spouse/significant other of an individual with diabetes, a friend of an individual with diabetes, an individual who works with a diabetes related company, healthcare organization, diabetes medical/device company, and other/unknown. A second individual reviewed another 50 randomly selected profiles for both subsets of users. For 50 users in the all user subset, two coders agreed on 44 of them. The Cohens kappa was 0.58. In the subset of DSMA users, they agreed on 40 of them. The Cohens kappa was 0.71.

3.2.4 Results

There were 29,599,683,822 tweets in the entire Twitter dataset, representing 10% of all tweets from 2013-2014. Of these, there were 1,368,575 diabetes-related tweets, based on the selected diabetes terms and hashtags. One third of these tweets (454,261) were retweets and about 2% (26,763) carried geo-location information. Table 3.2 shows the frequencies of tweets and the frequencies of users tweeting with those hashtags/search terms in our extracted dataset.

Figure 3.1 shows the total number of diabetes tweets over the two-year period. The largest number of tweets occurred in November 2013 on the World Diabetes Day. More than 70,000 diabetes-related tweets were posted in our dataset. Figure 3.2 shows the total number of tweets using the #dsma hashtag and figure 3.3 shows the number of unique users tweeting with the dsma hashtag at least once.

Figure 3.4a shows the monthly distributions of diabetes-related tweets, which were most frequent in November, likely attributable to the World Diabetes Day. For tweets using the #dsma hashtag, figure 3.4b shows January had the largest proportion. Based on the manual review of the tweets, new years resolutions for diabetes were a common theme.

Figures 5.1a and 5.1b show the distribution of all diabetes-related tweets and #dsma tweets by day of week. For diabetes-related tweets, the proportion of tweets

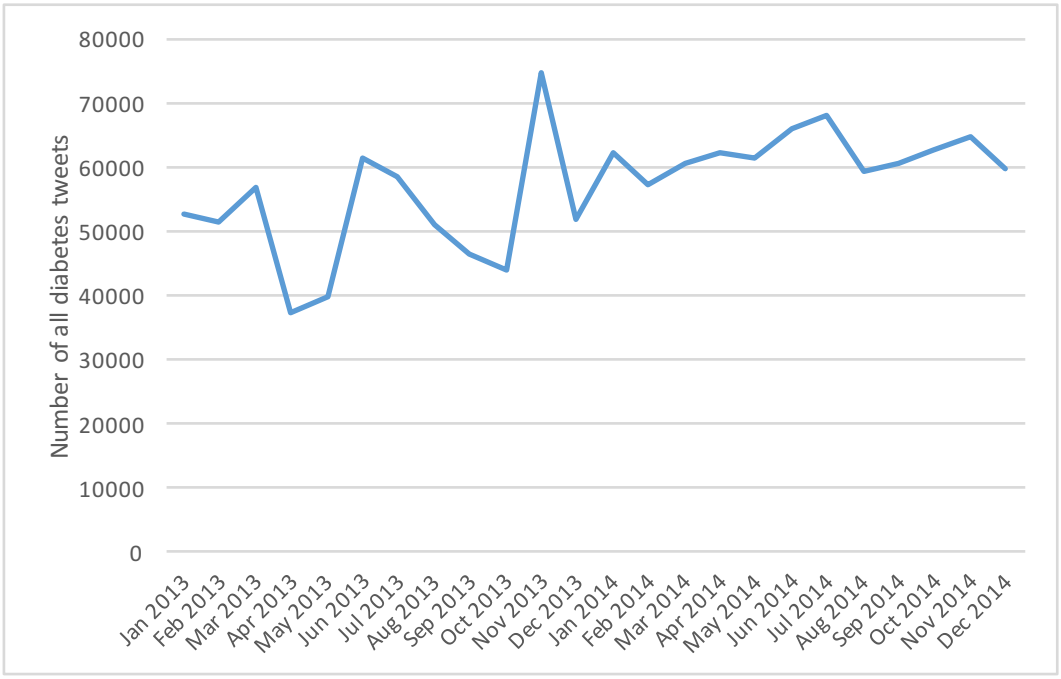


Figure 3.1: Tweet volume for all diabetes-related tweets

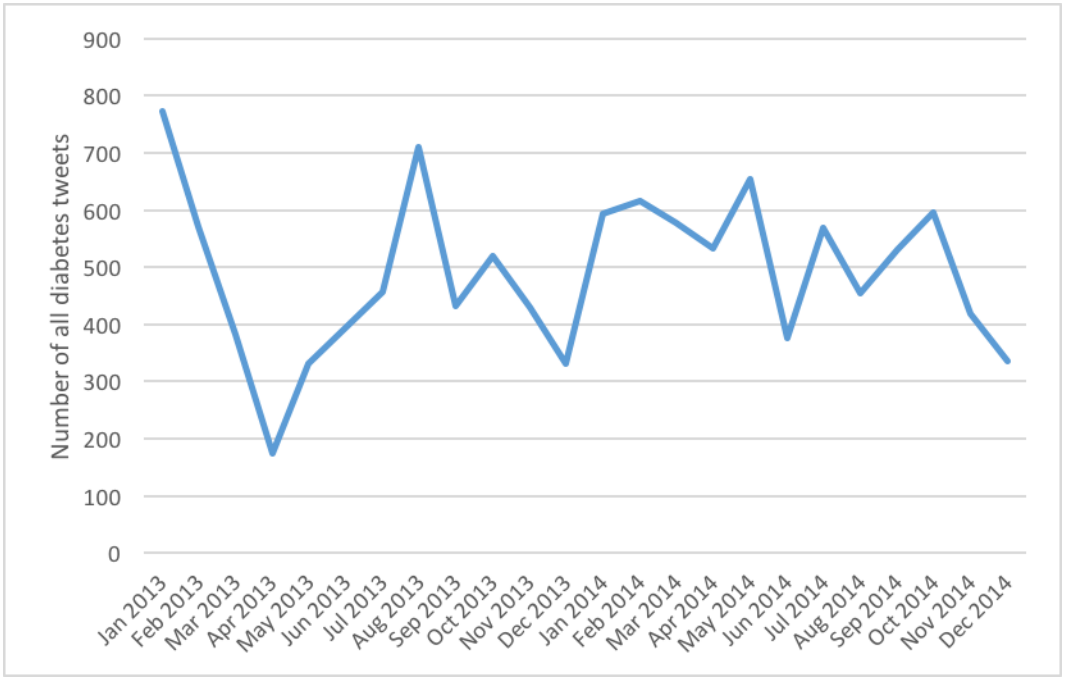


Figure 3.2: Tweet volume for dsma users

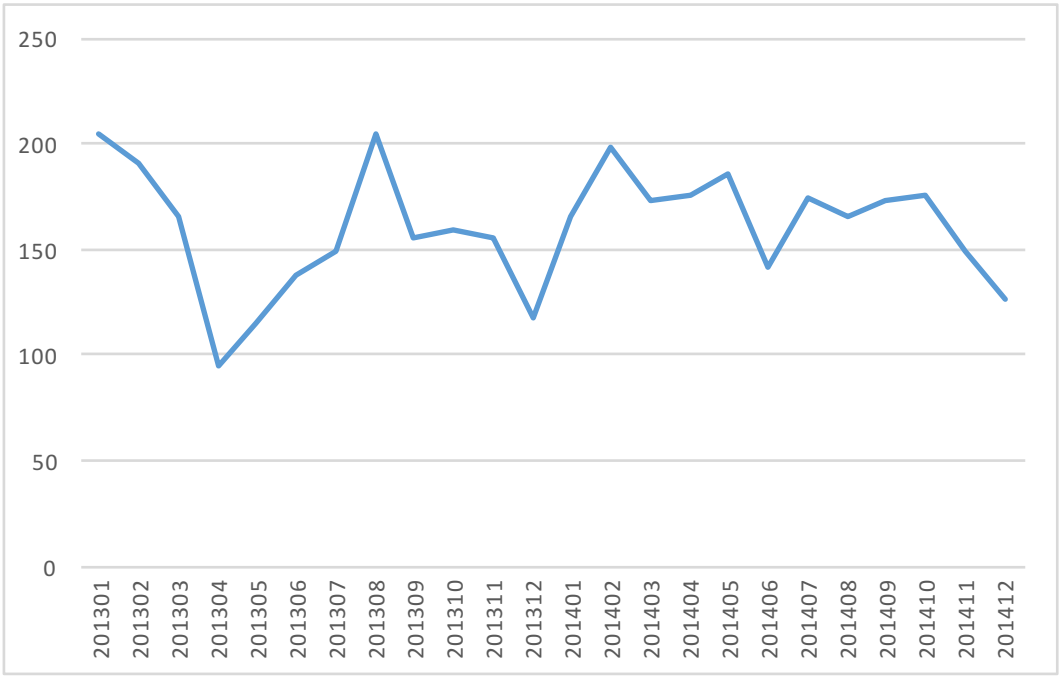
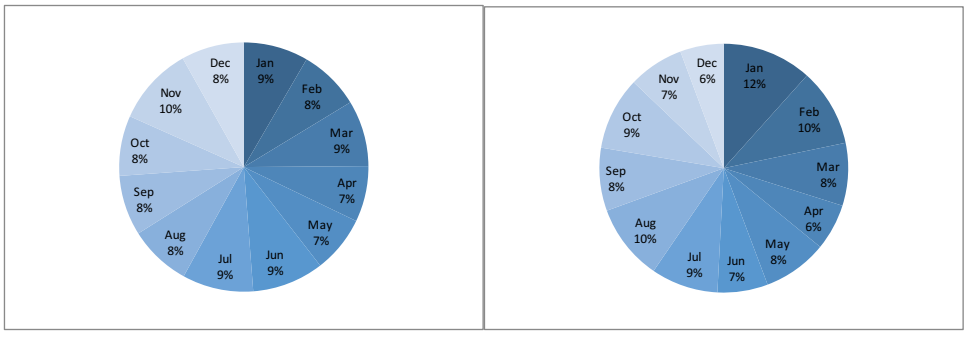


Figure 3.3: Number of unique users who tweet with #dsma each month



(a) All tweets (b) #dsma tweets

Figure 3.4: Monthly tweet volume distribution of all diabetes-related tweets and #dsma tweets

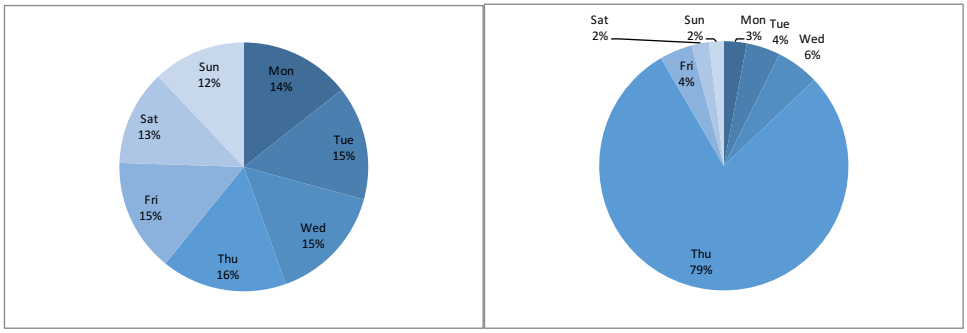
	Number of tweets (%)	Number of users (%)
diabetes	1,200,268 (87.7%)	748,001 (89.6%)
#diabetes	165,868 (12.1%)	67,229 (8.1%)
Insulin	83,820 (6.1%)	59,728 (7.2%)
glucose	60,033 (4.4%)	46,357 (5.6%)
#doc	27,616 (2.0%)	16,457 (2.0%)
#dsma	11,757 (0.9%)	1,424 (0.2%)
blood glucose	10,212 (0.7%)	6,904 (0.8%)
#t1d	9,040 (0.7%)	3,835 (0.5%)
#dblog	5,711 (0.4%)	1,132 (0.1%)
insulin pump	5,179 (0.4%)	4,061 (0.5%)
#type1	3,211 (0.2%)	1,800 (0.2%)
#type2	2,905 (0.2%)	1,468 (0.2%)
#bgnow	2,470 (0.2%)	753 (0.1%)
#type1diabetes	1,812 (0.1%)	1,248 (0.1%)
#bloodsugar	1,718 (0.1%)	1,213 (0.1%)
#type2diabetes	1,388 (0.1%)	1,035 (0.1%)
#t2d	935 (0.1%)	452 (0.1%)
#showmeyourpump	932 (0.1%)	645 (0.1%)
#wearenotwaiting	327 (0.0%)	183 (0.0%)
#diyps	132 (0.0%)	50 (0.0%)
#cwd2014	7 (0.0%)	7 (0.0%)

Table 3.2: Frequency of tweets and users tweeting with those terms/hashtags

was higher during the weekdays compared with the weekend days. On average, there were 2,011 diabetes-related tweets posted on weekdays, compared with 1,684 tweets posted on weekends ($p < 0.001$). In contrast, the majority of #dsma tweets were posted on Thursdays, which is due to the fact that there is an online chat organized by a community of individuals with diabetes and caregivers at 9 p.m. Eastern time every Wednesday night[3], which uses the #dsma hashtag for participating in the conversations.

For the tweets with geo-location information, we plotted the locations of all tweets in Figure 3.6, which will bias toward English-speaking countries because of our query terms. The east coast of the United States, Indonesia, and United Kingdom were frequent locations for geo-tagged tweets. Table 3.3 lists the countries with more than

[3]Date and time values in our dataset were returned by Twitter in UTC time



(a) All Tweets (b) #dsma Tweets

Figure 3.5: Days of the week tweets volume distributions all diabetes-related tweets and #dsma tweets

100 geo-tagged diabetes diabetes-related tweets.

We further counted the frequencies of tweets posted in different states of the United States. We compared the number of geo-tagged diabetes tweets and the number of all geo-tagged tweets in each state. We further performed χ^2 test between the percentage of diabetes geo-tagged tweets in each state and the percentage of diabetes geo-tagged tweets in the US. Table 3.4 shows that users in Missouri and Iowa are much more likely to tweet about diabetes, while people in Louisiana and South Carolina are less likely to tweet this topic.

Of the 500 users randomly selected from the diabetes-related tweets, 471 of them were categorized as other/unknown. Figure 3.2.4 shows the identify distribution of the rest 29 users. There were 12 healthcare organizations, a very small number of healthcare providers, and 7 diabetes patients. On the other hand, only 15.6% of dsma members identities were either not related to diabetes or unknown based on their Twitter profile information. Most of them were diabetic patients and their families and friends. Figure 3.2.4 shows the identity distribution of the sampled dsma users. 52.9% of users in this group were individuals living with type 1 diabetes. The results show that diabetes is a common topic on Twitter that not only people directly related to diabetes will tweet about. However, diabetes community members are more likely to reveal their diabetic identity in their Twitter profiles. They actively

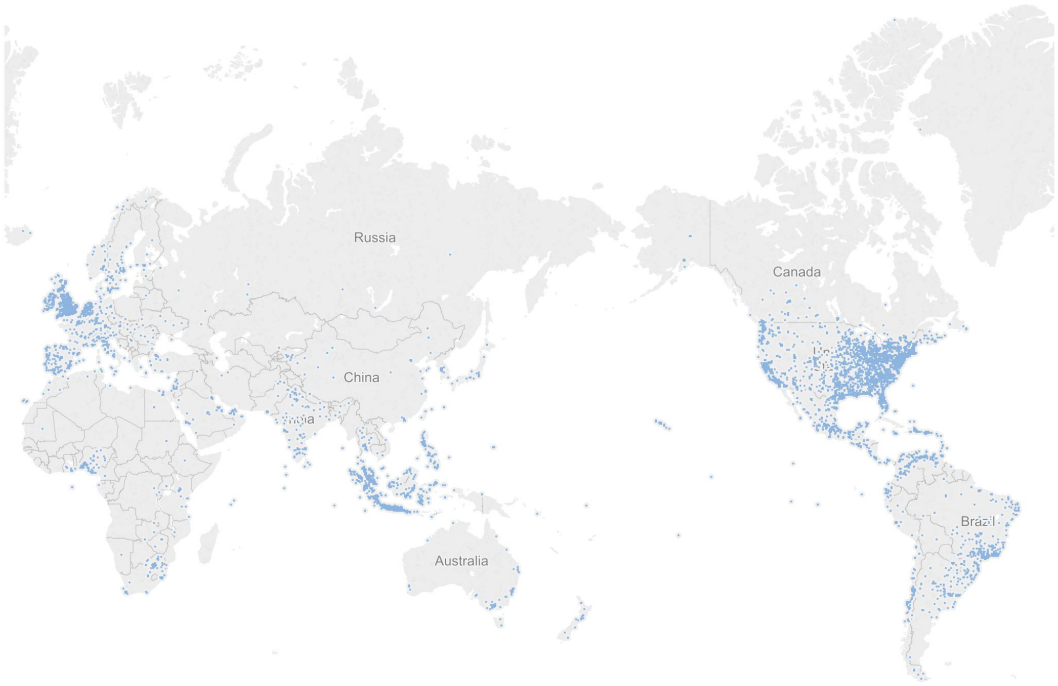


Figure 3.6: Geo-locations of diabetes tweets

use community hashtags to interact with other community members and track their own health updates.

3.2.5 Discussion

Our study shows that diabetes-related conversations happen frequently on popular social media websites such as Twitter. More importantly, Twitter has become a common place for diabetic patients to connect with each other through online communities without physical limitations. With the availability of a 2-year Twitter dataset, our study was able to assess the scale and depict the spatiotemporal distribution of English diabetes conversations on Twitter. Similar amounts of diabetes-related tweets were posted for different months of a year with the exception that the World Diabetes Day attracted most public attention of diabetes on Twitter. It suggests that Twitter can serve as an effective platform to increase publics awareness of diabetes.

In addition to the general public, we specifically investigated a diabetes community

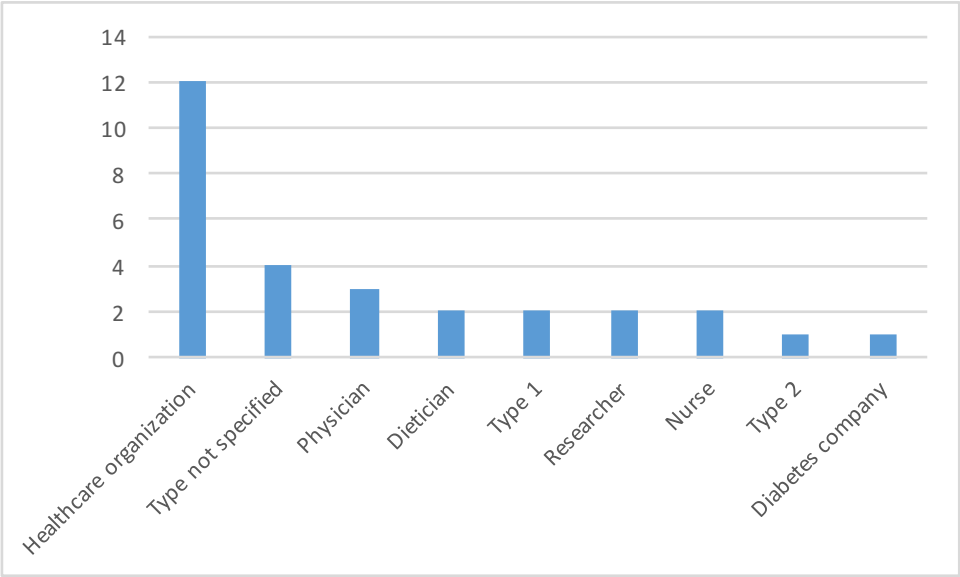


Figure 3.7: Identity distribution of all user group who are not in the "other/unknown" category

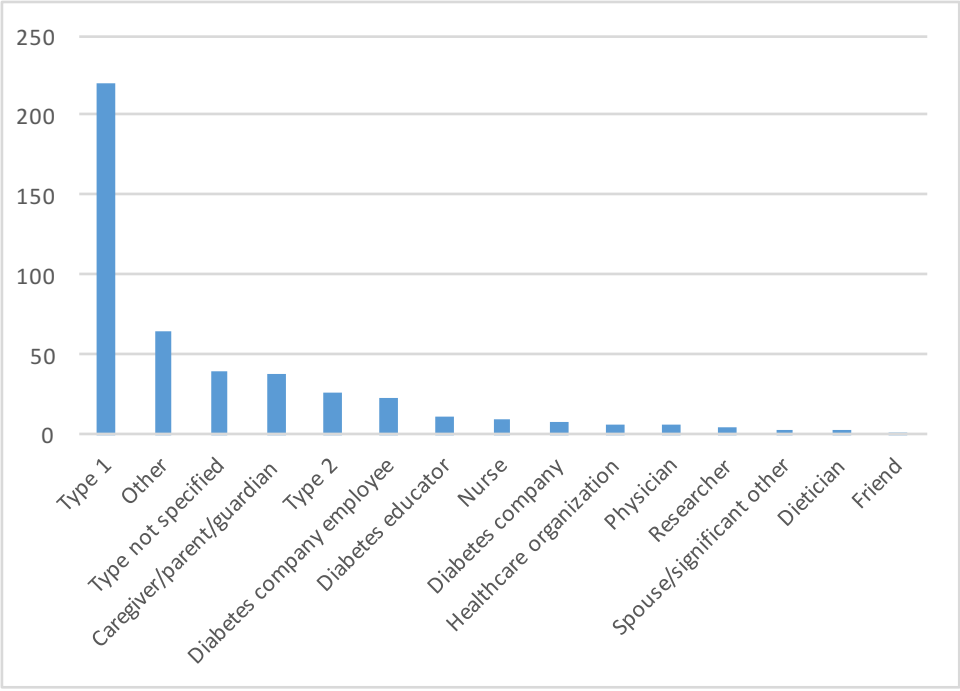


Figure 3.8: Identity distribution of dsma user group

Table 3.3: Frequency of the geo-tagged diabetes tweets in countries with more than 100 appearances

Country	Number of diabetes geo tweets
United States	10047
Indonesia	5355
United Kingdom	1897
Venezula	1172
Mexico	1042
Brazil	816
Mylasia	611
Canada	590
Philipines	439
Ghana	350
Spain	325
Nigeria	299
Argentina	260
Chile	223
India	220
Australia	218
Dominican Republic	199
Netherlands	189
South Africa	185
Colombia	167
Singapore	147
Ireland	107
Sweden	105

DSMA on Twitter. DSMA promotes social media engagement of people affected by diabetes by hosting a one-hour chat each week on Twitter. As a result, most of their communications, which were identified with the hashtag #dsma, happened every Wednesday night during their weekly chat. Unlike the general public who tweeted about diabetes, most of the DSMA community members identified their relationship to diabetes in their Twitter profiles. Diabetes patients and caregivers represented the largest proportions of DSMA members, more than half of which are individuals with type I diabetes.

One limitation of study is the Twitter data we collected were a 10% sample of all tweets available. It restricted us from conducting social network analysis of the

diabetes community on Twitter. In the future, we plan to collect tweets with hashtag #dsma prospectively through Twitter search API, which would allow us to capture all of the communications within the community.

3.3 Summary

In this chapter, we analyzed users' intent to participate in online health-related discussions, characterized their relevant activities and categorized their identities through two case studies. Our results show that health-related information pervades social media. In addition to questions about specific conditions and treatment, users like to socialize with other patients and caregivers in the same demographic group or share similar health goals or interests. People related to chronic diseases such as diabetes actively participate in social media to foster support and education. Participants are mainly patients and caregivers.

State	No. diabetes geo tweets	No. geo tweets	percent	Sig. (χ^2)/Sign
US	*10047*	*178135802*	*5.64E-05*	*NA*
MO	175	2261354	7.74E-05	***/+
IA	115	1525212	7.54E-05	***/+
MD	320	4367820	7.33E-05	***/+
NY	820	11421726	7.18E-05	***/+
AZ	221	3104479	7.12E-05	***/+
MI	438	6290429	6.96E-05	***/+
MA	274	4105766	6.67E-05	***/+
IN	221	3352388	6.59E-05	**/+
KS	106	1627864	6.51E-05	/+
NJ	468	7735544	6.05E-05	/+
NV	123	2038987	6.03E-05	/+
IL	392	6576342	5.96E-05	/+
WI	129	2179750	5.92E-05	/+
CA	1250	21749630	5.75E-05	/+
CO	101	1764414	5.72E-05	/+
TN	175	3071877	5.70E-05	/+
MN	125	2247418	5.56E-05	/-
VA	262	4728438	5.54E-05	/-
CT	112	2049573	5.46E-05	/-
FL	588	10841532	5.42E-05	/-
KY	131	2485082	5.27E-05	/-
OH	428	8155948	5.25E-05	/-
WA	167	3194835	5.23E-05	/-
PA	361	7080794	5.10E-05	*/-
GA	314	6295080	4.99E-05	**/-
TX	854	17933550	4.76E-05	***/-
AL	130	2932046	4.43E-05	***/-
NC	222	5364669	4.14E-05	***/-
SC	103	3086889	3.34E-05	***/-
LA	110	3299788	3.33E-05	***/-

*Note *p<.1, **p<.05, ***p<.01

Table 3.4: Number of geo-tagged diabetes tweets in top states in U.S. with more than 100 appearances sorted by percentage

CHAPTER IV

Identifying Health-related Information on Twitter

In previous chapters, I have shown that health-related conversations prevail in social media sites. This information can be used as effective signals to predict disease outbreak or understand public opinions, yet it is challenging to extract these conversations from heterogeneous and noisy social media data. This chapter describes a framework of identifying health-related information on Twitter using a combination of state-of-the-art medical natural language processing tools and machine learning classifiers. I first explored potential health-related conversations on Twitter by applying MetaMap to extract medical concepts from a year of geo-tagged tweets. A subset of tweets mentioning eye-related signs/symptoms identified by MetaMap were then manually examined to discover common categories of mistakes MetaMap is prone to make in social media context. Finally, I demonstrated the effectiveness of filtering medically irrelevant tweets with high accuracy using a machine classifier.

4.1 Introduction

Social media websites are increasingly used by the general public as a venue to express health concerns and disseminate information during public health crises. (*Steele*, 2011) Numerous prior studies have demonstrated that social media can be utilized as a valuable source of information for the purposes of early detection of disease

epidemics and solicitation of public opinions on controversial medical or ethical issues. (*Dredze*, 2012; *Salath and Khandelwal*, 2011; *Salathe et al.*, 2012; *Zheluk et al.*, 2012) Nevertheless, identifying medically-relevant information is still challenging in the social media context.

Keyword search is one of the most commonly used methods to extract relevant health information from social media. It is commonly applied when the researchers only look for discussions of a specific set of diseases in social media. Developing a comprehensive keyword list however requires extensive relevant professional expertise and knowledge of how the concepts may be expressed in social media context. In addition to keyword search, *MacLean and Heer* (2013) developed a context aware classifier to identify medical terms in social media content. Unfortunately, it lacks the critical ability to map the terms to clinical concepts or determine their semantic category.

On the other hand, extracting medical information/knowledge from biomedical literature or clinical texts is a well-studied problem in the medical natural language processing domain. (*Meystre et al.*, 2008) Various methods and tools have been developed to handle this task with demonstrated performance. (*Savova et al.*, 2010; *Patrick and Li*, 2010) However, social media content is very different from clinical texts in nature. It is not limited to medical context and the language usage is much more casual. Thus it is not clear whether we can directly apply the existing medical information extraction tools to the social media content and achieve similar performance as when they are applied to the clinical text.

We therefore selected one of the most widely used natural language processing tools, MetaMap (*Aronson*, 2006) to process a sample of Twitter messages to explore various health-related discussions in this popular social media platform and evaluate the performance of medical NLP tools in social media context. Although MetaMap was originally designed to map biomedical literature, many studies have applied it to

social media text as well. *Liu and Chen* (2013) applied MetaMap to extract drug and adverse event entities from discussions in patient forums. *Goeuriot et al.* (2011) used MetaMap to tag their corpus of drug reviews with UMLS concepts. More relevant to our study, *Jiang and Zheng* (2013) used Metamap to identify drug effects from drug-related tweets they collected. *Bian et al.* (2012) leveraged MetaMap to extract semantic feature to find possible side effects mentioned in the users' Twitter timeline.

4.2 Extracting Medical Concepts in Twitter Conversations with MetaMap

4.2.1 Materials and Methods

We have a collection of 10% sample of all tweets circulated on Twitter since March 2011, via the streaming application programming interface (API) with Gardenhose access provided by Twitter. Among all tweets in our dataset, around 2% are tagged with GPS coordinates recording the accurate location of where the tweets were sent. The location information is critical if we want to use the tweets to predict disease outbreak or analyze regional risky health behaviors. We therefore selected all geo-tagged, English tweets sent in 2013 from our collection and processed them through MetaMap. MetaMap is a natural language processing tool developed by the National Library of Medicine to extract medical concepts from free text. It is an extension to the traditional dictionary-based methods because of its ability to match not only simple keyword and phrases, but also spelling variants, abbreviations, acronyms, synonyms, partial phrases, or keywords occurring in a different order. (*Aronson*, 2006)

We extracted all the medical concepts from the tweets that were identified as a sign or symptom by MetaMap. In total, 298,767,507 tweets were parsed by MetaMap. 291 unique signs and symptoms appeared in more than a hundred tweets (see Appendix A).

4.2.2 Data Exploration

For each tweet mentioning medical concepts identified by MetaMap, we extracted its text content along with its posted date and time and geo-location information. We further recovered the local time of when a tweet was posted using the time zone inferred by the geo-location information. Then we conducted exploratory analysis of spatiotemporal distributions of tweets for different signs and symptoms. Figure 4.1 shows some examples of spatiotemporal distributions of tweets with different symptoms. We can see that people tend to tweet about sleeplessness around 4 am in the morning. Hangover were more frequently tweeted on Sundays. Pruritus were more likely to be reported during summer. People in Kansas liked to tweet hangover most. We further compared our results of spatial distributions of few concepts to the Behavioral Risk Factor Surveillance System (BRFSS) data provided by The U.S. Centers for Disease Control and Prevention (CDC)[1]. Specifically, we compared the state distribution of hangover tweets with results from different drinking-related survey questions in BRFSS. We computed the Pearson correlation coefficients between normalized number of hangover tweets in a state and percentage of people responding any drinking behavior, binge drinking behavior and heavy drinking behavior in the past 30 days. The correlation coefficients are 0.65, 0.74, and 0.54 respectively.

Although the percentage of medically-relevant tweets identified by MetaMap is high for the aforementioned concepts with less ambiguous terms, we also noticed that MetaMap recognized a large proportion of medically irrelevant tweets especially for concepts with terms having different meanings in non-medical contexts. In the following section, we systematically reviewed the medical relevancy of a random sample of tweets discussing eye-related signs or symptoms and analyzed why MetaMap would fail under different circumstances.

[1] http://www.cdc.gov/brfss/

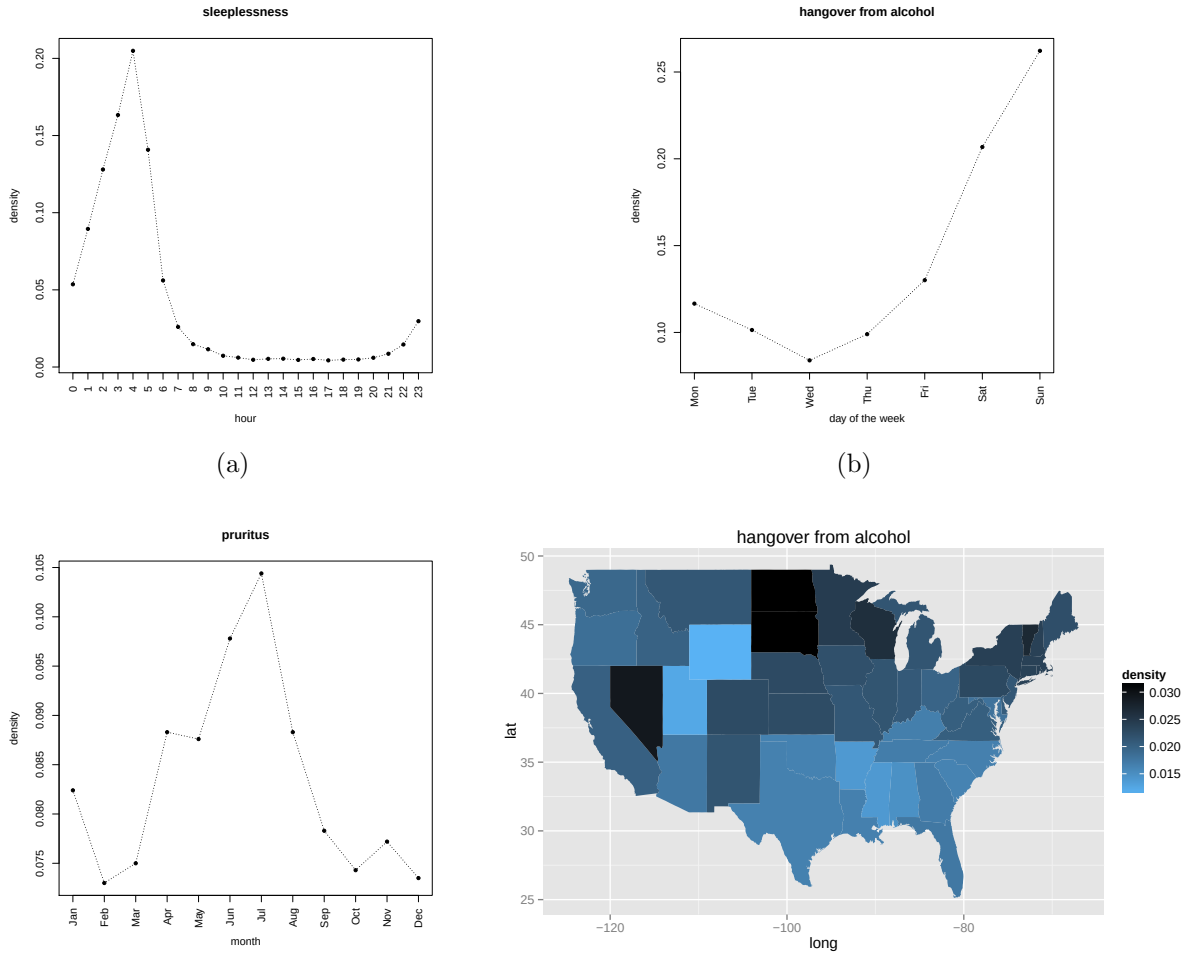


Figure 4.1: Spatiotemporal distributions of tweets of different signs/symptoms. (a) Hourly distribution of tweets of sleeplessness; (b) daily distribution of tweets of hangover from alcohol; (c) monthly distribution of tweets of pruritus; (d) state distribution of tweets of hangover from alcohol.

4.3 Case Study of Tweets with Eye-related Signs and Symptoms Identified by MetaMap

Collaborating with ophthalmologists, we specifically explored the usage of all eye-related signs and symptoms on Twitter. We evaluated the performance of using MetaMap to extract medically-relevant discussions on eye-related diseases from tweets against human-annotated ground truth and trained a machine classifier to filter out medically irrelevant tweets.

Two ophthalmologists went through the list in Appendix A independently, and identified sixteen concepts that appeared to be eye-related signs or symptoms. Table 4.1 shows the list of all sixteen eye-related concepts, along with their frequency of occurrence identified by MetaMap in our Twitter dataset.

Signs and Symptoms	Concept Unique Identifier in UMLS	Frequency of occurrence
Visual halos (disorder)	C0271188	15,209
Redness of eye	C0235267	2,999
Hallucinations, visual	C0233763	1,655
Rolling of eyes	C0522336	1,249
Flasher - visual manifestation	C1705500	951
Eye swelling	C0270996	551
Sore eye	C0578687	510
Watery eyes	C3257803	503
Blurred vision	C0344232	477
Eyes twitching	C0850674	433
Dryness of eye	C0314719	345
Pain in eyes	C0151827	314
Bloodshot eye	C0005858	233
Eye pain	C0151827	219
Feeling of heat in eye	C0234657	130
Circles under eyes	C0686795	105

Table 4.1: Frequency of Eye-related concepts identified by MetaMap

4.3.1 Annotation

Once the tweets that contain words that may be eye-related signs and symptoms were identified, we undertook a manual annotation task to labels the tweets as medically-relevant or not. Two annotators were recruited to determine medically relevant tweets based on the content of the tweet, including text, hashtags, and URLs contained in the tweet (if present). For tweets that were judged as medically-irrelevant, the annotators also categorized the primary cause of exclusion. Both annotators are native English speakers and familiar with language use on social media, especially Twitter.

Figure 4.2 shows the pipeline of our annotation process. In the first round, we randomly sampled twenty tweets for each of the sixteen eye-related concepts. Both annotators independently labeled the tweets as medically-relevant or medically-irrelevant. They also independently noted the reason why they believed a tweet was not medically relevant, based on its content. At the end of the first round, the annotators discussed their disagreements, and jointly proposed a saturated list of categories for medically-irrelevant tweets. I worked with the two annotators to come up with a draft codebook to annotate medically-relevant and medically-irrelevant tweets (with error categories).

Based on the draft codebook, in the second round of annotations, the two annotators labeled a new random sample of twenty tweets for each concept. All new tweets fell into existing categorization scheme. The codebook was then finalized based on the results after the second round of annotations.

Based on the two initial rounds of annotations, we also noted that the medically-relevant tweets could be classified as tweets that described one's own or someone else's signs or symptoms, previously experienced symptoms, hypothetical signs and symptoms that haven't yet developed, or tweets suggesting treatments for specific signs or symptoms.

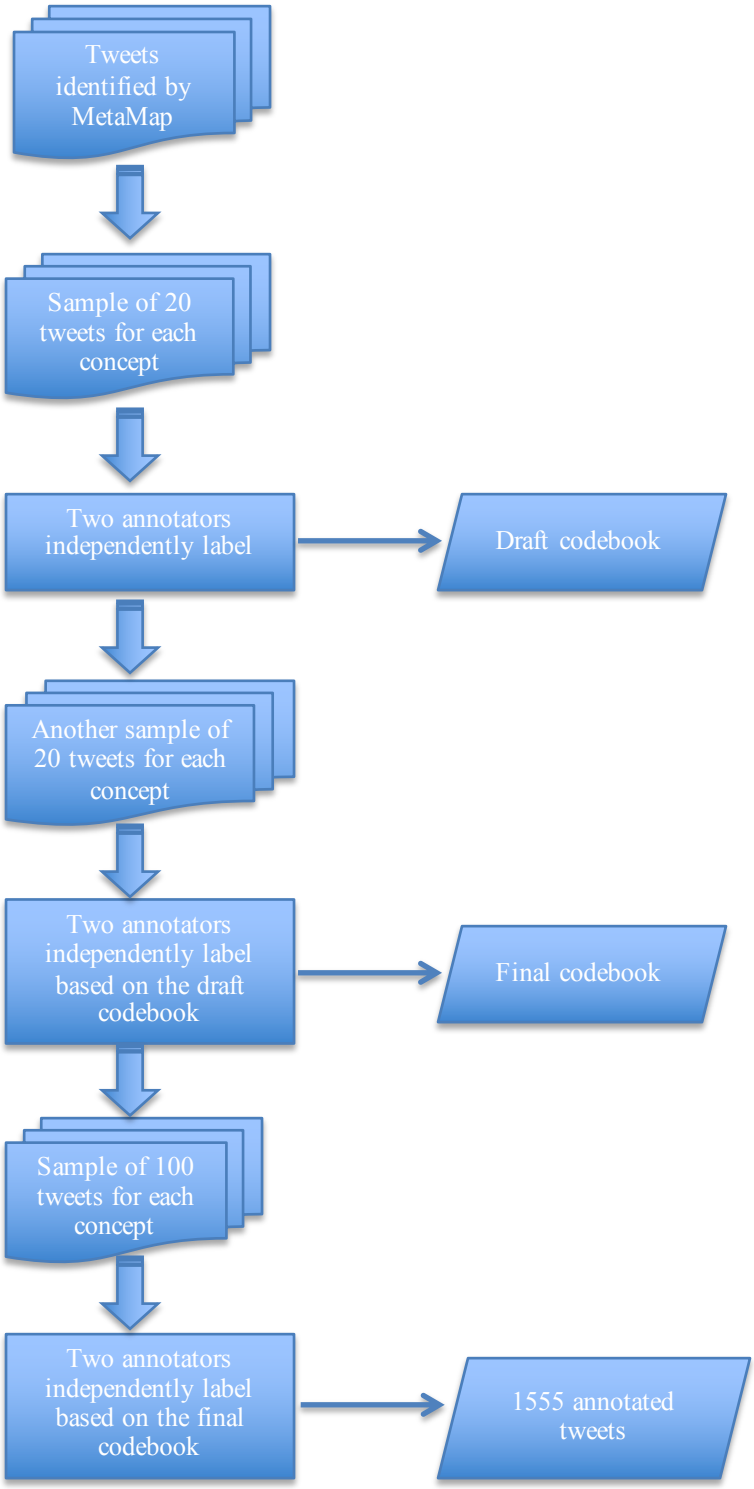


Figure 4.2: Diagram of the annotation process

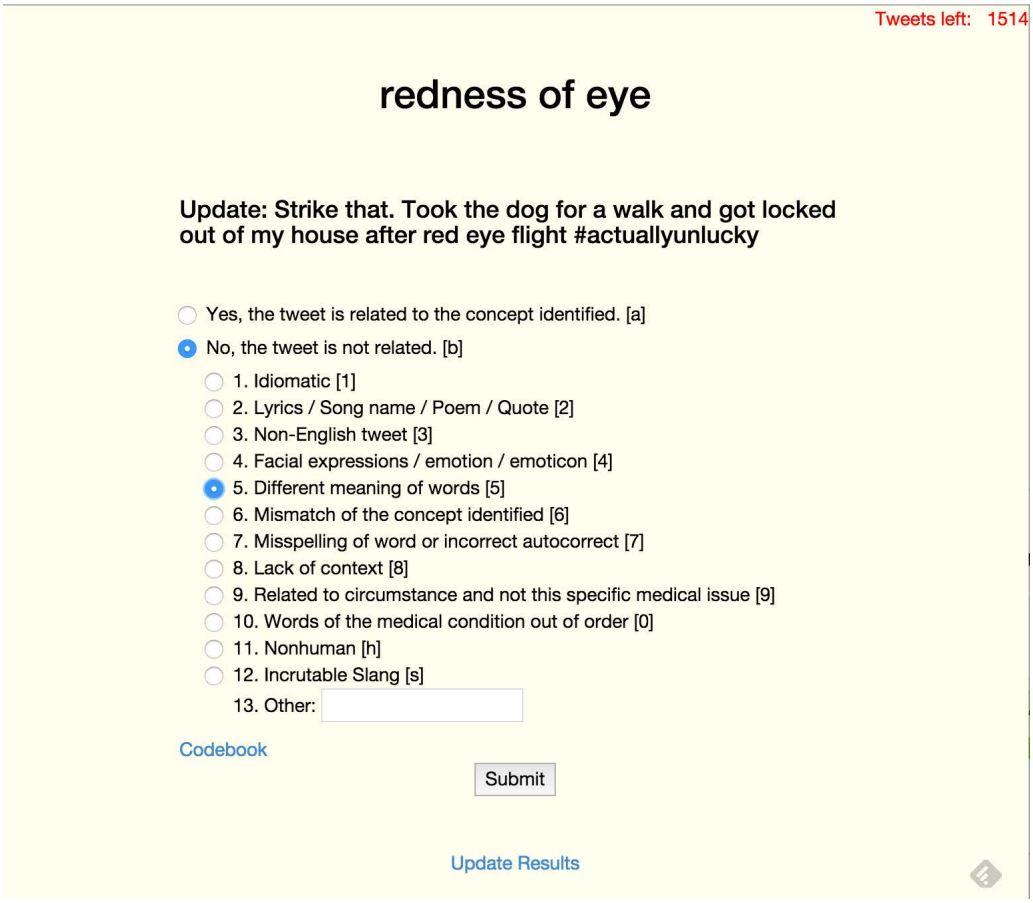


Figure 4.3: Screenshot of the coding interface

After the two rounds of annotations to finalize the codebook, the two annotators coded up to a hundred additional tweets for each concept. The new tweets were randomly selected from the rest of the non-annotated tweets. Since some concepts had less than 100 tweets after the first two rounds of pilot coding, in all 1,555 tweets were annotated in the final round. Over all three rounds, the two annotators labeled a total of 2,195 tweets.

4.3.2 Analysis of Annotation Results

The list below shows the definition of the categories of the medically-irrelevant tweets identified, as well as examples for each category.

- **Idiomatic**: The potential sign or symptom phrase is used in the sense of an

idiom, rather than in its medical meaning. e.g. @*redacted_user* ur a sight for sore eyes shorty) [sore eye]

- **Lyrics of a song, poem, or quote**: The potential sign or symptom phrase is either a song, a quote, the name of an album, or a line in the lyrics of a song. Annotators were asked to search the Web if they were not sure of the lyrics. e.g. You're aching, you're breaking, and I can see the pain in your eyes [pain in eyes]

- **Non-English tweet**: The context of the potential sign or symptom phrase is written in a language other than English. e.g. Aabsent nlg ako whole day! Swollen eyes! Ichy din [eyes swelling]

- **Facial expressions/emoticon**: The potential sign or symptom phrase is used to express an expression, emotion, or an emoticon. e.g. @*redacted_user* at work *rolling eyes* how r u? [rolling of eyes]

- **Different word meanings**: The potential sign or symptom phrase is used in one of its other non-medical meanings, rather than its medical meaning. e.g. Smokin' hot red eye ribs with coffee and chipotle sauce marinating and primed for Saturday eating. #slowandsticky [redness of eye]

- **Mismatch of the sign/symptom identified**: The potential sign or symptom phrase described a medical situation, but not the category being identified. The tweeter used an incorrect term for the condition they are describing. e.g. Trust me ladies, nothing burns more than flinging your intuition razor water into your eye! ... I'm blind! BLIND! [watery eyes]

- **Misspelt word or incorrect autocorrect**: The tweeter meant to write another word, but accidentally wrote the word that is related to a potential sign

or symptom. e.g. Actually just feel like my heats been ripped out [feeling of heat in eye]

- **Lack of context**: Lack of sufficient context makes it difficult to determine the true meaning or intention of the potential sign or symptom. e.g. Eye pain la. [eye pain]

- **Circumstantial**: The sign or symptoms described by the tweeter could potentially match a medical condition, but the context did not describe a medical condition. e.g. Watery eyes when u dice the onions. *hiss [watery eyes]

- **Words out of order**: The words from a potential sign or symptom phrase occur out of order in the tweet, leading MetaMap to falsely identify it as a medical concept. e.g. @*redacted_user* Can't take my eyes off the toilet roll on his desk. Can't investment managers afford tissues? [rolling of eyes]

- **Nonhuman**: The sign or symptom relates to the category being defined, but the symptom is not affecting a human. e.g. Taking my dog to the vet :(she has a strange swollen eye ! [eye swelling]

- **Inscrutable Slang**: The tweeter has used abbreviations, slang, or jargon that makes it difficult to interpret the meaning of the tweet. e.g. Sarap makipag eye to eye contact,lalo n ngaun may sore eyes ako.

In the following sections, we report results over the 1,555 tweets annotated in final round, once the codebook was finalized.

4.3.2.1 Inter-annotator agreement

Out of the 1,555 tweets coded in the final round, the two annotators agreed on 1,387 tweets (89.2%) on whether the tweet was medically relevant or not. Cohen's kappa (*Cohen et al.*, 1960) was 0.76. Among the 1,387 tweets that both annotators

agreed on, 427 were labeled as medically-relevant and 960 were labeled as medically-irrelevant. The two annotators also agreed on the error categories labeling on 606 out of 960 (63.1%) medically-irrelevant tweets. Cohen's kappa was 0.56. We only included tweets that two annotators agreed on for further analysis and analyzed their disagreement in section 4.3.2.4.

4.3.2.2 Analysis of Medically Ambiguous Concepts

The percentage of medically-relevant tweets varies significantly across different concepts. We can see from table 4.2 that ambiguous concepts such as "redness of the eye" are more prone to having non-relevant tweets than less ambiguous concepts such as "circles under the eyes." Further, the terms listed in the Unified Medical Language System (UMLS) Metathesaurus to identify the concepts also play a significant role, since Metamap uses the UMLS Metathesaurus as one of its primary resource of terminologies. For example, MetaMap maps the phrase "seeing things" to the concept "hallucinations, visual." For example, although in biomedical text, most occurrences of the phrase "seeing things" may indicate visual hallucinations, it is just a common phrase on Twitter. Not surprisingly, none of the tweets for this concept were labeled by the annotators as being medically relevant.

4.3.2.3 Analysis of error categories for medically irrelevant tweets

We analyzed the primary reasons for medically-irrelevant tweets being recognized as relevant by MetaMap. The distribution of error categories for the spurious matches is shown in Figure 4.4. The most common reason for the false positives is the difference in word meanings (error category 5). For example, the phrase "red eye" does not have a medical meaning when it is in the context of a "red eye flight" or the "red eye ribs".

The second common reason is that the tweet is a snippet from the lyrics of a song, poem, or quote (error category 2). Nearly 20% of all spurious matches fell in this

Concepts	#relevant	#irrelevant	%irrelevant
Circles under eyes	62	0	0
Eyes twitching	78	11	12.36
Eye swelling	63	23	26.74
Eye pain	45	46	50.55
Sore eye	41	43	51.19
Bloodshot eye	35	37	51.39
Watery eyes	43	47	52.22
Blurred vision	25	64	71.91
Dryness of eye	14	77	84.62
Hallucinations, visual	12	66	84.62
Redness of eye	7	77	91.67
Pain in eyes	2	96	97.96
Feeling of heat in eye	0	90	100
Flasher - visual manifestation	0	87	100
Rolling of eyes	0	96	100
Visual halos (disorder)	0	100	100
Total	427	960	69.21

Table 4.2: Number of medically-relevant and medically-irrelevant tweets on eye-related concepts that both annotators agreed on, sorted by percentage of medical-relevant tweets

category. An example tweet is as follows: "A true friend sees the pain in you [sic] eyes, even when you have a big smile on your face." We note that this category can be very difficult for dictionary-based method to rule out. Even human annotators sometimes needed help from a search engine to tell whether a tweet is a quote or a line from a song or a poem.

The facial expressions / emoticons category (error category 4) was the next significant cause of error, with about 13% of spurious matches attributed to this error category. The large proportion of these may be due to the our selection of eye-related signs and symptoms as the topic of study, specifically due to the concept "rolling of eyes". Rolling eyes are mostly used to indicate annoyance or frustration on Twitter rather than a medical symptom.

The fourth largest category is "words of the medical condition out of order" (error category 10), contributing over 8% of the errors. This is primarily due to MetaMap

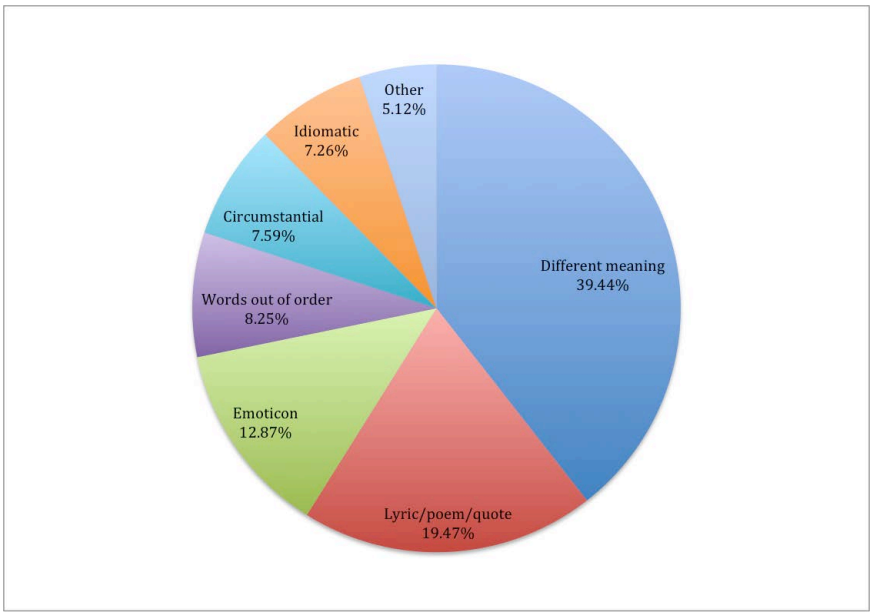


Figure 4.4: Distribution of categories of spurious matches

reduced emphasis on word order when searching for matched items. Although forcing the exact order while matching a medical phrase would eliminate this issue, it can cause a significant reduction in the recall of medically-relevant information at the same time.

The circumstantial category (error category 9) accounts for about 7.6% of the spurious matches. This includes situations where the symptom described is not due to a medical issue. For example, tweets complaining watery eyes while slicing onions are not considered medically relevant. Idiomatic usages (error category 1) are another cause of spurious matches. Idioms such as"a sight for sore eyes" and "see pain in someone's eyes" were challenging to differentiate even for human annotators.

4.3.2.4 Analysis of Disagreements during Annotation

We further analyzed the frequency and nature of disagreement between the two annotators, both on the judgment on whether a tweet is medically relevant or not, and if judged irrelevant, the assignment of the error category. Table 4.3 lists the most common disagreements between annotators on judging a tweet as medically relevant

or irrelevant. The most common case was the disagreement of whether the mention of a sign/symptom is circumstantial (error category 9) or a true medical condition. Circumstantial mentions are difficult to judge with little context present in tweets. One example of such a case is: "My right eye twitching; who tf' talkin bout me"[sic].

The annotators also expressed difficulty in agreeing on the meaning of some tweets given their terseness. In our annotation study, one of the coders tended to take the words more literally, while the other was more familiar with metaphors and colloquial styles on Twitter, and could better interpret different nuances of phrases. For example, when classifying the tweet, "waking up round two.. *swollen eyes*", one coder interpreted it as describing the tweeter's swollen eyes when they woke up. The other coder interpreted the asterisk as an implication of an exaggeration or an emoji, thus classifying this tweet as non-relevant.

Lastly, annotators reported that they do not have a perfect knowledge of song lyrics, poems, or quotes. They mainly judged tweets as such by looking for rhyme or unusually poetic language in the tweets, and confirming the same via a Web search.

Disagreement	Frequency (N>=10)
Y, N9	74
Y, N5	24
Y, N2	16
Y, N8	15
Y, N1	14

Table 4.3: Top annotation disagreements on judging medical relevance

Table 4.4 lists the most common disagreements between annotators on the error category, when they both agree that the tweet is medically irrelevant. We can see that different word meanings category (error category 5) seems to be the most confusing category for annotators. The category was mostly confused with the idiomatic use (error category 1).

Disagreement	Frequency (N>=20)
N1, N5	118
N2, N5	37
N5, N10	28
N5, N9	24
N1, N2	22

Table 4.4: Top annotation disagreements between two error categories

4.3.2.5 Analysis of MetaMap Matching Scores

For every phrase that MetaMap identifies as a medical concept in a given text snippet, it assigns a matching score from 0 to 1000 based on the quality of the match between the phrase and a potential candidate in the UMLS Metathesaurus. (*Aronson*, 2006) The evaluation is based on four measures that examine whether it is a perfect match, a partial match, a match with gap, or a match with variants. In our study, we included all concepts identified by MetaMap as candidates, irrespective of their matching score. This allows us to include as many potential candidate phrases as possible, but it may introduce unwanted errors when the matching score are low.

Figure 4.5 shows the ROC curves for MetaMap with respect to the MetaMap score. We also computed F1 scores for each MetaMap score threshold to look for best trade-off between precision and recall. The MetaMap matching score threshold of 820 had the highest F1 score of 0.539. We further conducted additional qualitative analysis of tweets with different matching scores to further understand the performance of MetaMap. An example of a true positive tweet with a low matching score is "My vision getting blurry". Although it has the same meaning as the matching string "blurry vision", it received a score of 783 because the words appear out of order with an additional non-concept word in the middle. While the exact keyword search method failed, the flexible matching algorithm of MetaMap was able to identify this tweet as a positive instance.

On the other hand, such flexibility also occasionally introduced errors. Figure 4.6

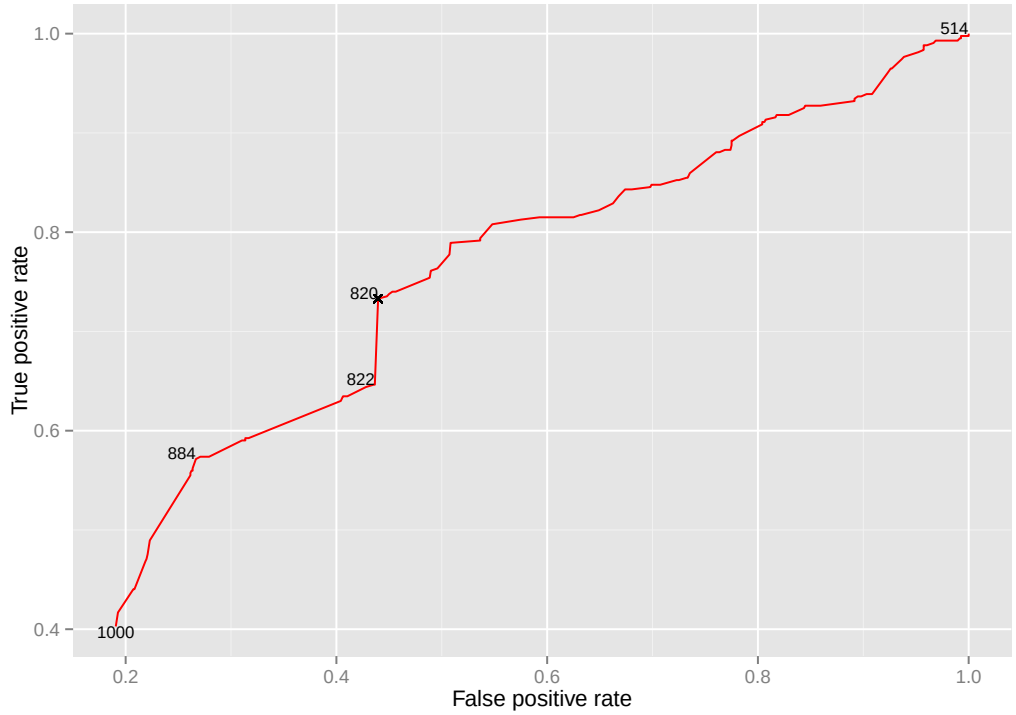


Figure 4.5: ROC curve with MetaMap score

shows that most of the mistakes in the "word out of order" category were introduced by concepts with imperfect matching score. One example of such an error is the false positive mapping of the concept "watery eyes" to the tweet "aww quick wash ur eyes out with water."

4.3.3 Automatic Classifying Medical Relevancy of Tweets

To further exclude medically-irrelevant tweets with eye-related signs and symptoms identified by MetaMap, we built a Support Vector Machine (SVM) (*Cortes and Vapnik*, 1995) classifier using 1,387 unanimously coded tweets as training corpus. Using unigram and bigram features with binary weighting scheme, we were able to train a classifier using radial basis function (RBF) kernel with a high accuracy of 90.1% with 5-fold cross validation.

The result informs us that classifying medical-relevant tweets is a relatively easy

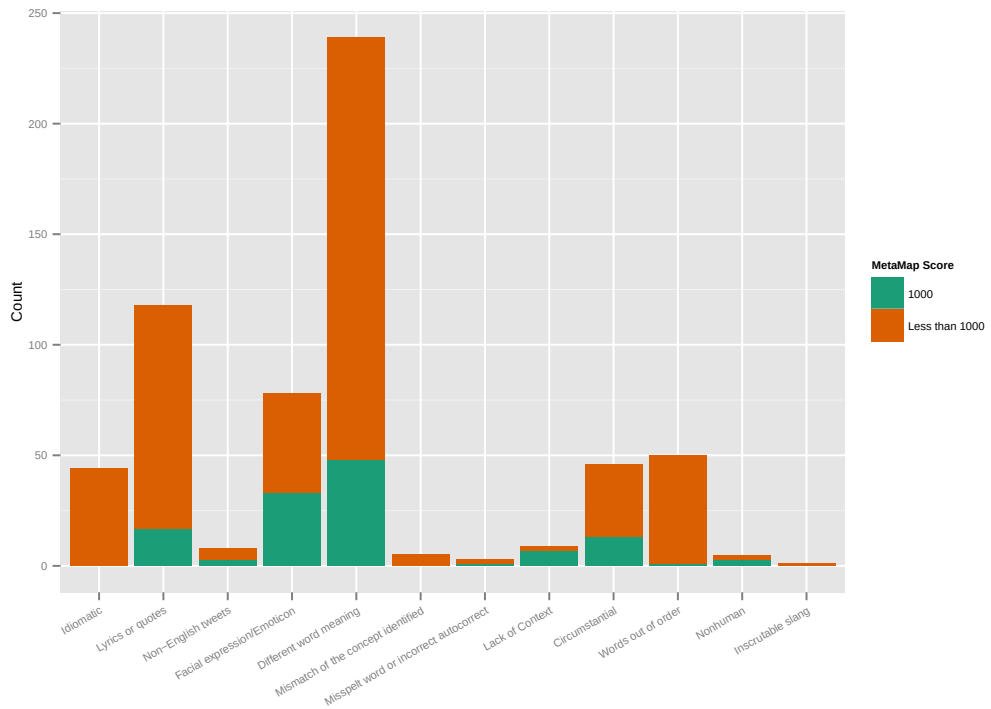


Figure 4.6: Number of false positives in different categories by MetaMap score

text classification task. With adequate manual-annotated training data, a classifier can reach very high accuracy with just unigram and bigram feature. Given the brevity of tweet text, the binary weighting scheme outperformed the TF-IDF weighting.

I further conducted feature analysis using the feature selection tool in LIBSVM. (*Chen and Lin*, 2006) The word "eye" was the most useful feature as it appeared in most positive examples and were absent from many negative examples such as tweets mentioning "seeing things". Phrases from concepts with very high medical relevancy or very low medical relevancy were also very useful. As I discussed before, medical relevancy varies significantly among different concepts. Instead of annotating same amount of documents for all concepts, focusing on concepts with highly ambiguous terms can save human efforts significantly. Another category of helpful features were symptoms of other diseases (e.g. "sore throat", "runny nose"). It makes sense since users often complain different kinds of symptoms in the same message. Finally, some

Twitter specific features such as retweet or at another users were also among most discriminative features. It indicates that metadata is also an important feature when filtering out medically-irrelevant tweets.

4.3.4 Discussion

As a case study to evaluate the effectiveness of applying medical information extraction tool to distillate medical concepts in social media, we applied the popular tool, MetaMap, to identify eye-related clinical concepts in geo-tagged tweets. Based on our manual review results, only 30% of concepts recognized by MetaMap were medically relevant. The percentage of medically-relevant tweets varied significantly across different eye-related concepts. It depended heavily on the clinical or consumer terms a concept maps to in external knowledge resources such as the UMLS, and how those terms are commonly used on Twitter.

Given a clear definition of medically-relevant tweets, human annotators with little medical background knowledge were able to identify erroneous matches with a reasonably good inter-annotator agreement. Most of the disagreements were due to different interpretations of the tweet content, especially because of the brevity of tweets and the lack of sufficient context. Dictionary-based methods had the most trouble in disambiguating different context-dependent meanings of potential medical phrases. Medical terms in song lyrics, poems, or quotes were another major source of non-relevant tweets. In the future, our technique could be enhanced with the use of idiomatic expression and quotation databases. Another source of error was related to the nature of Twitter usage, in describing day-to-day activities and common symptoms much more likely than serious medical conditions.

Finally, we demonstrated the effectiveness of applying machine classifier to filter medically irrelevant information with high accuracy.

This study focused on identifying eye-related signs and symptoms in a sample

of Twitter stream. Our choice of eye-related signs and symptoms may also bias the distribution of erroneous matches towards certain error categories. However, the error categories we identified and the effectiveness of further filtering with machine classifier should still generally apply to other medical concepts studies on Twitter.

4.4 Summary

In this chapter, we discussed a framework of identifying health-related information in Twitter. Leveraging state-of-the-art medical NLP techniques, we first identified potential tweets that are medically relevant and mapped them to UMLS. Through a careful manual review of tweets with eye-related concepts, we demonstrated that using traditional medical NLP tools to extract medical concepts from social media is prone to generate various kinds of spurious matches. Different meanings of English phrases, and famous quotes containing medical concepts are common sources of errors. We finally showed that a machine classifier can be used to exclude such spurious matches effectively.

CHAPTER V

Public Opinions Analysis using Social Media Data

Social media are increasingly used by the general public to express their opinions. In this chapter, I demonstrate the power of using text-mining and natural language processing techniques to solicit public opinions from different sources of social media data. Section 5.1 describes a study of analyzing public concerns towards vaccination-autism linkage using online news comments. Section 5.2 describes another study that accesses the public sentiment of the Affordable Care Act through sentiment analysis of relevant discussions on Twitter.

5.1 Mining Online News Comments for Public Opinions Regarding Vaccination-Autism Linkage

5.1.1 Introduction

Childhood vaccination is arguably one of the most important public health interventions ever developed, yet parental concerns continue to prevent many children from receiving all recommended immunizations (*Gust et al.*, 2004). Many parents and caregivers have misperceptions and concerns about the safety of vaccination (*Gellin et al.*, 2000; *Kennedy et al.*, 2011) despite the fact that the medical community has confirmed its safety through rigorous clinical studies (*Marshall and Baylor*, 2011; *Katz*

et al., 2012). As a result, understanding the concerns about vaccines that lead parents to decline childhood immunizations remains an important public health issue (*Larson et al.*, 2011).

The controversy surrounding the measles, mumps, and rubella (MMR) vaccine epitomizes the ongoing discord between parents and healthcare providers. In 1998, Dr. Andrew Wakefield published a study in the Lancet suggesting a correlation between the receipt of the MMR vaccine and the subsequent development of pervasive developmental disorders such as autism (*Wakefield et al.*, 1998). This study, of only twelve patient subjects, generated widespread fear and suspicion among parents around the world and was partially responsible for declined vaccination rates. Whereas autism previously had no known causes, families could now point to vaccination as the culprit.

Over time, researchers were able to rule out the MMR vaccine as a cause for autism (*Halsey and Hyman*, 2001; *DeStefano et al.*, 2004; *Mrozek-Budzyn et al.*, 2010; *Uno et al.*, 2012; *Madsen et al.*, 2002; *Dales et al.*, 2001), including the mercury-based thimerosal preservative that it used to contain (*Parker et al.*, 2004; *Hviid et al.*, 2003; *DeStefano*, 2007), and research into autism spectrum disorders has demonstrated the existence of a strong genetic component (*Kumar and Christian*, 2009; *Hallmayer et al.*, 2011). In 2010, after an investigative journalist looked into the background of the original paper (*Deer*, 2011b,a), the Lancet publicly retracted the original article (*Caplan*, 2009) and in 2011 the British Medical Journal denounced the article to be an "elaborate fraud" (*Godlee et al.*, 2011). Despite convincing evidence to the contrary, many people remain concerned about the potential link between autism and vaccines in general, and MMR in particular (*Bazzano et al.*, 2012; *Smith et al.*, 2007; *Casiday et al.*, 2006). In addition to exposing the fraudulent work by Dr. Wakefield, the news media have long been implicated in fomenting the concerns about vaccines (*Holton et al.*, 2012; *Guillaume and Bath*, 2008; *Speers and Lewis*, 2004;

Dobson, 2003; *Goldacre*, 2007; *Hackett*, 2008). In fact one recent study found that more than a third of news articles about vaccines, over a 10-year period of time portrayed vaccines in a negative light (*Hussain et al.*, 2011).

Many news websites now incorporate a social media component in which readers can leave comments related to certain articles. These comments, presumably left by peers, have the potential to influence readers beyond what was reported in the original article (*Kushin and Yamamoto*, 2010). Because the news media can play such an important role in the debate about vaccine safety, both to expose fraud and to perpetuate fears related to vaccinations, I sought to explore the sentiments left in the reader comments of news articles related to highly publicized announcement that the work by Wakefield's study was fraudulent. The results would provide a unique opportunity to explore the public's concern about the MMR vaccine as it related to the media coverage.

In this paper, I used automated machine learning techniques to analyze user-generated online news comments. Specifically, I employed two state-of-the-art text-mining approaches, text summarization and topic modeling, to distill key opinions from user comments. Such an approach could provide an alternative way to review and understand the public concern about the MMR vaccine and help to determine the feasibility of applying automated machine learning techniques to understand other public health issues in social media.

5.1.2 Materials and Methods

5.1.2.1 Collection of News Articles and Comments

The data collection for this study was conducted in June 2011. I retrieved major news articles written in reference to the January 2011 British Medical Journal publication that announced the retraction of Dr. Andrew Wakefield's paper and described it as fraud. (*Deer*, 2011b) I used the Google advanced search function

Name	Website
ABC News	abcnews.go.com
CBS News	cbsnews.com
CNN	cnn.com
Digg	digg.com
The Huffington Post	huffingtonpost.com
MSNBC	msnbc.com
National Public Radio	www.npr.org
New York Times	nyt.com
Reddit	reddit.com
USA Today	usatoday.com
The Washington Post	washingtonpost.com
The Wall Street Journal	www.npr.org

Table 5.1: List of online news websites

to identify news articles written between January 1, 2011 and April 1, 2011 that contained the keywords "vaccine" (or in variant forms such as "vaccination", "immunization") and autism. Article selection was limited to those published by the most popular news websites in the U.S. based on the Alexa Global Traffic Rank (http://www.alexa.com/topsites) and U.S. Traffic Rank by Compete and Quantcast (http://www.quantcast.com/top-sites), respectively. I also used the search function embedded within each site to search for relevant new coverage.

Further, for a news website to be included in this study it must meet the following two additional requirements: (1) the site must provide a commenting feature for readers to express their views with regard to the news story covered; and (2) the site must offer open access to the content of the news article as well as all comments that readers provided.

The news Websites analyzed in this study are listed in Table 5.1. In addition to the widely known newspapers and broadcast networks, I also included the Huffington Post, which is a hybrid online-only site that provides a mix of aggregation from other sites and original reporting; and Digg and Reddit which are social news websites that allow users to vote and comment on stories that were originally published on other sites.

I ultimately identified 30 online news articles that satisfied our selection criteria. Each of these articles had between 8 and 3,281 user comments, for a total of 13,698 comments. Each comment was segmented into sentences using LingPipe 4.1.0, (*Alias-i*, 2008) yielding a total of 54,575 sentences. 417 stop words such as "the", "of", and "is" were then removed. I removed all sentences that contained fewer than five words after the removal of stop words. The processed dataset contains 11,556 comments and 35,160 sentences.

Once the data were processed, the next step is to summarize the common concerns about the news event. A desirable succinct summary should convey the most important content from the wide range of user's comments, yet should cover all aspects of the comments. In our scenario, it is desirable to extract the most representative sentences from the comments of the news articles, and also convey as much diversity or as little redundancy as possible among the selected sentences. Below we present two classical machine-learning approaches to achieve this goal.

5.1.2.2 Text Summarization

Text summarization is a classical natural language processing task that automatically creates a compressed version of one or more documents that should still cover the content of the original. The most robust summarization methods generate a summary by extracting and concatenating the most representative yet non-redundant sentences from the original documents. Such an approach generally performs better than abstractive techniques that require paraphrasing the original text. (*Erkan and Radev*, 2004)

In this paper, I employ a novel summarization algorithm, DivRank, to extract the most representative sentences that cover diverse aspects of the original forum messages (*Mei et al.*, 2010). DivRank is a network-based ranking algorithm similar to PageRank (*Page et al.*, 1999), employed by the Google search engine. While

PageRank assumes the importance of a website is determined by the number of links it receives, DivRank orders the representativeness of a sentence by its semantic similarity to other sentences. Specifically, DivRank constructs a language network in which each vertex corresponds to a sentence in the original corpus. Two vertices are connected only if the semantic similarity between two sentences is above a predefined threshold. Different from PageRank, DivRank uses a reinforced random walk to avoid redundancy in top-ranked sentences (*Mei et al.*, 2010).

In this study, I first represented each sentence as a vector of unigrams, quantified using a binary value based the appearance of the word. To generate a better representation of sentences, I also manually reviewed the 500 most frequently appearing words in the corpus after stop word removal (ranging from 159 to 8,953 instances each) and mapped semantic related words so that they would be treated as the same concept for the following analysis. Examples include mapping 'money' to 'dollar', 'outbreak' to 'epidemic', and 'whoop' to 'pertussis'. This concept-matching step reduced the size of vocabulary into 18,196. I then create the network of sentences by connecting sentence pairs with a cosine similarity greater than 0.4. The resultant network has 28,507 nodes and 1,496,632 edges. Two parameters of DivRank, damping factor and self transition probability, are set to 0.85 and 0.3 respectively. Top 30 sentences with highest DivRank scores were then extracted.

5.1.2.3 Topic Modeling

An alternative way to summarize a corpus is through topic modeling. (*Blei et al.*, 2003; *Hofmann*, 1999). Topic modeling assumes that the corpus consists of documents that are generated from a mixture of abstract topics. Every topic is represented as a probability distribution over all distinct words in the corpus and it corresponds to a distinct aspect of the corpus content.

There are many topic models proposed in the literature that mainly differ with

respect to the assumptions of how a document is generated. (*Steyvers and Griffiths*, 2007) The Latent Dirichlet Allocation (LDA) is one of the most popular models and has been widely used in previous literature. (*Blei et al.*, 2003) Although it generally works well for long documents that exhibit a mixture of multiple topics (e.g., news articles or scientific papers), its effectiveness might be reduced when documents are short and on a single topic. Since our corpus consists of sentences as "documents", we also employed a simpler topic model, the mixture of unigrams model (*Nigam et al.*, 2000) with a simplified assumption that each sentence only contains one topic.

As in the text summarization approach described above, I treated every sentence in the corpus as a separate document. For both the LDA and the mixture of unigrams models, I set the number of topics as thirty, equivalent to the number of sentences extracted by DivRank. Once the topics were learned, I assigned all sentences to the closest topic based on the document-topic distributions. Finally, I extracted the top sentence for each topic ranked by the weighted percentage of words assigned to the topic.

To quantitatively evaluate and compare the topics generated by LDA and the mixture of unigram model, I calculated the average point-wise mutual information (PMI) across all topics. PMI is a frequently used metric to measure the semantic coherence of topic models. (*Newman et al.*, 2011a) Specifically, I first computed the PMI between each pair of words from the 20 words that had the largest probabilities in each topic, then calculated the PMI of the topic by taking the average of these word pairs, and then averaged the scores for 30 topics. Specifically,

$$PMI(\Phi) = \frac{1}{30}\sum_{\theta\in\Phi}\frac{1}{190}\sum_{1\le i\le j\le 20} log\frac{p(w_{i,\theta}, w_{j,\theta})}{p(w_{i,\theta})p(w_{j,\theta})}, \tag{5.1}$$

where $w_{i,\theta}, w_{j,\theta}$ are the words ranked at the i^{th} and j^{th} position in topic θ. $p(w_{i,\theta}, w_{j,\theta})$ is the probability that both $p(w_{i,\theta})$ and $p(w_{j,\theta})$ appear in one sentence, and $p(w_{i,\theta})$ is the probability that $w_{i,\theta}$ appears in a sentence.

The results show that topics extracted by LDA yielded a much higher averaged PMI score, 0.688, than those extracted by the mixture of unigram model (averaged PMI score = 0.154). A manual review of two sets of topics also confirmed that LDA extracted more coherent and representative topics. In the following analysis, we therefore only report results from the LDA model.

5.1.2.4 Evaluation Methodology

To evaluate the performance of the machine learning algorithms on distilling key aspects of the vaccination-autism debate, I compared their outputs against factors identified by qualitative studies on this subject in literature. Specifically, I focused on the set of concerns identified by previous studies that solicited parents' perceptions of MMR vaccines using qualitative methods such as focus group interviews and questionnaires. The evaluation was then done based on what fraction of such concerns was also revealed by the machine learning approaches.

To collect the evaluation data set, we searched the title and abstract fields of published studies indexed by PubMed using the following query: (parents OR parental) AND (survey OR interview OR "focus group" OR questionnaire) AND MMR AND (vaccine OR vaccination). In all, 58 articles were returned and 13 of them were found to be relevant after a manual review. The sentences describing the opinions/concerns of individuals regarding the MMR vaccine were extracted from these articles and separated into two groups: those that were collected from responses in focus group interviews (usually more open-ended responses), and those that were responses to survey questionnaires (more narrow, specific responses). Two coders then used the card sorting method (*Cairns and Cox*, 2008) to group up sentences into topical aspects and labeled them. Then, the two coders also matched the top sentences generated by DivRank and the topics extracted by LDA to the labeled topical aspects. If new topical aspects were found, they were added and labeled accordingly. The coders

Rank	Sentence	Score
1	"vaccines cause autism by the mercury in the shots . . . "	0.0440
2	"What else is needed to get parents to vaccinate their children against measles."	0.0315
3	"I do believe that for most children vaccines are safe and I do vaccinate my children."	0.0301
4	"Publishing a paper doesn't require peer review in some journals as they are likely there to get the research paper out so others CAN peer review it."	0.0247
5	"If vaccines have nothing to do with autism, then never vaccinated kids should also have a one percent autism rate."	0.0226
6	"No such studies exist in any peer reviewed journal."	0.0222
7	"I did vaccinate my child, one vaccine at a time."	0.0210
8	"All parents of autistic kids would like to know why their children are autistic."	0.0164
9	"It is similar with autism and autism spectrum disorders such as Asperger's Syndrome."	0.0160
10	"I am a parent of an autistic child and I have proof that the Vaccines caused him Autism."	0.0137
11	"People will say vaccines have been studied' but what they really mean is two vaccines, not ALL vaccines?"	0.0135
12	"How do we know that Autism Is not caused by the childs parents having vaccinations and when the child gets his vaccine the Autism shows up."	0.0130
13	"With all of the vaccinations people get, of course people with autism will have had a vaccine!"	0.0126
14	"Parents who have an autistic child or children go through a lot."	0.0126
15	"Was it because you studied scientific research and reached the conclusion on your own?"	0.0116
16	"I am the parent of a child with Autism and my child received all of her vaccines on schedule."	0.0111
17	"People will believe what they want to believe, parents most of all."	0.0110
18	"which there is as more people get vaccinated more people are getting autism . . . "	0.0106
19	"My kid received a shot then was diagnosed with Autism; therefore the shot caused the Autism."	0.0103
20	"As far as drug companies go, vaccines are not a largely profitable drug."	0.0092
21	"Statistical science isn't science, but insurance."	0.0090
22	"The problem is that there is little or no money to be made making vaccines."	0.0089
23	"As the parent of a child with Autism, I never believed that vaccines caused Autism."	0.0082
24	"Many people are too selfish to think about other people's children."	0.0077
25	"People and children are dead because of this man's actions."	0.0074
26	"These are not researchers funded by pharmaceutical companiesthere is no big pharma conspiracy here."	0.0069
27	"And if you don't think people take vaccines every day then don't say so."	0.0069
28	"The vaccine was a cause of disease, not a preventive of disease!"	0.0068
29	"There is this study which is fraudulent and then there is every other study which says there is no link."	0.0066
30	"Yet we give shots for immunity to these diseases when they don't even have an immune system?"	0.0064

Table 5.2: The top 30 sentences derived from the corpus based on DivRank score. The higher the score, the more a sentence represents other sentences in the corpus.

discussed their disagreement and reached consensus on the final assignments.

5.1.3 Results

5.1.3.1 Text Summarization

The 30 sentences with the highest DivRank scores are shown in Table 5.2. These range from sentences that are supportive of the routine childhood vaccination recommendations, to those that are against the recommendations or raise questions about their safety.

5.1.3.2 Topic Modeling

Table 5.3 shows the top 20 words most likely to be generated from each topic. The most popular topics (those with the highest likelihood) include parents' decision making about vaccinating their children and concerns of vaccine's side effects. For each topic we also extracted sentences with the largest proportion of words assigned to that topic to provide context for the topic.

5.1.3.3 Evaluation

Two coders extracted seventeen topical aspects from results of previous qualitative studies soliciting parents opinions towards vaccination, which are summarized below:

1. MMR delivery schedule
2. vaccine is a causal factor of autism
3. mistrust of advice and/or information from external entity/stakeholders
4. social responsibility/norm of vaccinating child
5. seek/want more research/explanations concerning MMR risk to children
6. media about mmr vaccine/MMR controversy
7. perceive vaccine-preventable disease severity/symptoms as mild
8. vaccine effectiveness and safety
9. side effects such as allergies and asthma, as potential risks
10. trust in external entity/providers
11. natural immunity of kids
12. subgroups of kids with weak/sensitive immune systems

1	2	3	4	5	6	7	8	9	10
drug	health	medical	mercury	mmr	disease	child	effect	autism	wakefield
company	public	people	vaccine	vaccine	vaccine	vaccine	vaccine	cause	andrew
money	medical	govern	contain	pertussi	risk	autism	reaction	vaccine	study
make	state	industry	preserve	die	child	age	side	link	fraud
profit	care	world	dose	polio	prevent	month	adverse	don	report
big	vaccine	doctor	amount	child	populate	parent	cause	say	medical
vaccine	dr	line	remove	baby	death	mmr	damage	believe	article
industry	school	trust	schedule	death	reduce	develop	brain	know	read
fda	govern	truth	poison	disease	number	old	long	evidence	publish
govern	america	think	safety	0	die	son	fever	think	deer
pharma	medicine	big	influenza	case	benefit	receive	mmr	prove	paper
manufacture	national	lie	inject	kid	childhood	given	death	doesn	journal
doctor	fund	research	childhood	parent	people	start	seizure	proof	news
dollar	university	money	safe	outbreak	infectious	normal	bad	claim	did
pay	cdc	pharmaceutical	single	old	chance	baby	list	child	britain
market	center	fact	baby	small	protect	time	severe	possible	brian
business	injury	media	recommend	kill	rate	sign	suffer	connect	media
sell	institute	make	time	pox	effect	two	serious	fact	story
billion	court	look	test	varicella	potential	symptom	high	case	lancet
cost	concern	keep	fda	got	unvaccined	change	child	study	interest
11	12	13	14	15	16	17	18	19	20
autism	http	autism	kid	science	medicine	food	school	virus	child
rate	www	disorder	don	evidence	science	water	child	vaccine	parent
increase	com	genetic	vaccine	theory	good	eat	kid	influenza	vaccine
vaccine	article	spectrum	know	conspiracy	thing	drink	parent	live	make
child	org	cause	sick	fact	modern	car	home	polio	decision
number	search	diagnose	child	scientific	food	put	class	infect	choice
study	gov	factor	take	believe	hat	hand	high	disease	risk
high	cdc	mental	think	global	practice	sure	student	smallpox	kid
rise	link	symptom	parent	warm	world	clean	work	kill	health
case	google	condition	people	people	medical	good	teach	h1n1	inform
amish	2010	environmental	blame	base	bad	healthy	autism	new	doctor
incidence	html	asd	tell	change	knowledge	don	educate	mmr	fear
show	watch	child	oh	data	natural	make	require	body	right
diagnose	read	diagnosis	did	support	human	kid	time	strain	put
correlate	video	syndrome	play	real	society	feed	treat	swine	believe
diagnosis	wikipedia	neurology	got	claim	advance	air	mitochondrial	case	refuse
drop	website	increase	away	belief	know	chemical	likely	ininfluenzaenza	responsible
kid	site	gene	time	study	least	milk	take	cause	live
populate	2005	adhd	won	scientist	isn	night	result	antibiotic	harm
age	2011	mitochondrial	stop	matter	think	baby	grade	dead	personal
21	22	23	24	25	26	27	28	29	30
anti	gate	child	study	study	mercury	immune	doctor	read	don
jenny	world	autism	link	medical	brain	system	know	post	know
mccarthy	bill	parent	autism	scientific	body	vaccine	child	question	say
people	cancer	son	wakefield	research	poison	disease	did	comment	thing
don	people	feel	vaccine	journal	blood	child	son	article	think
think	live	life	mmr	data	cell	herd	office	don	mean
re	money	blame	research	result	damage	response	ago	call	people
try	populate	family	done	publish	chemical	body	eye	answer	re
ignorant	time	help	find	scientist	level	protect	mother	side	doesn
read	country	live	show	find	human	virus	patient	people	right
say	think	love	found	commune	inject	natural	back	argument	make
talk	billion	problem	claim	conclusion	effect	effect	told	believe	sure
believe	research	answer	did	science	high	person	play	good	happen
stop	waste	normal	researcher	read	toxin	given	come	debate	wrong
post	smoke	understand	lawyer	done	cause	develop	two	ask	talk
listen	spent	way	result	base	found	healthy	right	write	bad
comment	dollar	know	prove	researcher	metal	different	give	name	understand
know	spend	sorry	causal	review	develop	young	old	actual	safe
stupid	long	find	data	paper	dna	time	ask	fact	doctor
guy	life	cure	paid	evidence	heavy	individual	went	time	point

Table 5.3: Topic model results

13. personal experience of vaccination

14. self guilt

15. adjuvant and preservatives

16. consideration of alternative treatment

17. parents' freedom of choice

Table 5.4 and Table 5.5 show the coverage of different topical aspects reported by interview studies and questionnaire studies respectively. The columns in the table represent the thirteen studies published between 2010-2012 on the presumed link between vaccination and autism. The columns are sorted by author names. The rows correspond to the seventeen aspects identified by the coders using the card sorting method described earlier. Each aspect was given a prevalence score calculated as the number of studies mentioning it. Every individual qualitative study as well as every automated algorithm used in this paper received two coverage scores calculated as the sum of the prevalence scores (based on either interview studies or questionnaire studies) of all topical aspects they covered. Note that each algorithm gets two coverage scores. Besides the 17 topical aspects extracted from previous qualitative studies, DivRank and LDA both discovered three additional aspects: "perception of conspiracy theory;" "validity of Wakefield's study"; and "additional resources needed from parents". Table 5.6 shows the coverage of these 20 topical aspects by DivRank and LDA.

The results confirmed that the set of thirty top ranked sentences from both DivRank and the topic modeling approaches could successfully cover most of the aspects extracted from previous qualitative studies. The coverage scores of DivRank and LDA models are higher than the highest coverage scores obtained by individual interview and questionnaire studies. Between the two machine learning algorithms, LDA yielded higher coverage scores and also covered more unique topical aspects.

Theme	Brown et al. 2012	Casiday 2007	Evans et al. 2001	Hilton et al. 2007	Tickner et al. 2010	Prevalence score
1	X	X	X		X	4
2	X	X	X	X		4
3	X	X	X	X		4
4		X	X		X	3
5	X	X			X	3
6	X			X	X	3
7	X	X	X			3
8	X	X	X			3
9	X	X				2
10		X			X	2
11	X		X			2
12		X		X		2
13	X				X	2
14				X	X	2
15	X					1
Coverage score	31	30	23	15	19	40

Table 5.4: Coverage of different themes of interview studies. "X" represents the article covers that aspect. The last column is the prevalence score of different themes, which is simply the number of articles covering the theme. The last row is the coverage score of each article measured by the sum of prevalence scores of all themes it covers.

The results confirm that both DivRank and topic modeling cover most of the themes extracted from previous qualitative studies. Their coverage scores are higher than or close to the highest scores of interview and questionnaire studies. Furthermore, DivRank can get high coverage score with few top-ranking sentences.

5.1.4 Discussion

Our results demonstrate that online news site readers have various concerns about vaccinations and that text summarization is an effective approach to understanding public health issues from social media text.

Based on the summarization results and topics extracted from the text, it is easy to identify readers' attitudes to vaccines which are similar to those reported in the literature, predominantly composed of survey- or interview-based studies. Some readers

Theme	Alfredsson et al. 2004	Borra et a.l 2009	Casiday et al. 2006	Cassell et a.l 2006	Dannetun et al. 2005	Flynn and Ogden 2004	Gellatly et al. 2005	Smailbegovic et al. 2003	Prevalence score
1			X	X					2
2		X	X				X	X	4
3		X	X	X					3
4			X	X					2
5			X	X					2
6				X					1
7	X				X	X		X	4
8	X	X	X				X		4
9	X	X	X		X		X		5
10			X	X		X	X		4
11	X	X		X	X				4
12	X			X	X				3
13	X			X					2
14	X	X		X		X			4
15					X				1
16								X	1
17			X	X					2
Coverage score	26	24	28	29	17	12	17	9	48

Table 5.5: Coverage of different aspects of questionnaire studies

	DivRank	LDA
1	X (7,16,30)	X (4)
2	X (1,5,10,12,13,16,18,19,23)	X (2,11)
3	X (4)	X (1,3,22,24)
4	X (24)	X (14)
5	X (8,11,15,21,29)	X (15,16,25)
6	X (25)	X (21)
7		
8	X (3,5,27,28)	X (5,6,19)
9		
10	X (20,22,26)	X (2)
11		X (27)
12	X (5)	X (27)
13	X (16,19)	X (7)
14	X (17)	X (23)
15	X (1)	X (4,26)
16		
17		X (14,20)
Wakefield's study validity	X(25,29)	X (10,24)
Need more resources	X(11)	X (15)
Scientific research validity	X(4,6,15)	X (25)
Coverage score	33(I)/32(Q)	37(I)/43(Q)

Table 5.6: Coverage of topical aspects by DivRank and LDA. Numbers in parentheses are the sentence or topic IDs that cover that aspect. The last row is the coverage score based on either interview studies (I) or questionnaire studies (Q)

believed in the conspiracy theory that vaccination is a lie told by the government and large pharmatheutical companies to earn huge profit. Even doctors, researchers, and public media are corrupted in their opinion. People also have various concerns about the MMR vaccinations such as toxicity concern, medical concern, and schedule concern. On the other hand, some readers argued about the importance of vaccination as an effective way to prevent the outbreak of epidemics. It is also interesting to notice that people claimed that they had done research and quoted various references to support their opinion. With respect to online references, while authoritative resources such as NLM, CDC and BMJ were frequented cited, pseudo-scientific anti-vaccine websites such as whale.to and ageofautism were still among most popular resources. Social media websites such as YouTube and Facebook were also mentioned within some comments. Our results show that although the MMR vaccine rate has become much higher in recent years, people still have similar concerns and misperceptions as previously held.

The consistence between the results of this study and traditional survey conclusions also proves the effectiveness of leveraging social media data and machine-learning techniques to understand public health issues. As more user-generated online content becomes available, many researchers are leveraging it to track public opinions. *Signorini et al.* (2011) used twitter to track levels of disease activity and public concerns during influenza season. Their results show that Twitter can be effectively used as a measure of public concern about health-related events. *Gonzalez-Bailon et al.* (2010) analyzed changes in public opinion by tracking political discussions. They emphasized that their method was different from polls or surveys in the aspect that they approximated public opinions by analyzing the discussions which people voluntarily participated in.

More relevant to this paper, *Skea et al.* (2008) conducted a thematic analysis of discussions about MMR in an online chat forum for parents. *Salath and Khandelwal*

(2011) examined the sentiments of tweets towards influenza A (H1N1) vaccination. They trained a machine learning algorithm on around 50,000 manually rated tweets, and then used the trained classifier to predict the remaining tweets. Comparing to these two papers, our methods are unsupervised in a way that no human labor is required to label or code the text. Specifically, summarization techniques can extract most representative sentences from a large collection of user-generated content. Topic models can further cluster the content into meaningful themes to help people get more insights of the corpus. The same techniques can be easily applied to analyze public response during other public health crisis without human supervision.

Finally, one potential application of our methods is to assist in the development of survey instruments. Comparing to questions developed by researchers themselves alone or from pilot interview responses with a small number of participants, questions generated based on summaries of large volume of online user-generated content can be less biased, and more representative and comprehensive. Survey designers thus can leverage our methods to effectively develop survey questionnaires with smaller cost.

Comparing to traditional survey studies, our methods have a much larger sample size. The participants are voluntary and their answers are free text instead of choosing from pre-designed categories. There are several limitations of our study, however. First, there is a selection bias of the sample as all participants are online news site users. In addition, we ignored the thread structure of the comments, and treated comments from different websites equally. It is also difficult to determine the true polarity of the opinion of the person who left the comment by looking at individual sentences outside of the context of the original full comment. In the future, this can be improved by weighting the comments based on thread structure and source of the comments. In addition, I will perform sentiment analysis at comment level incorporating thread structure information.

5.1.5 Conclusion

Vaccination is arguably one of the most important public health interventions ever developed; yet parental concerns continue to prevent many children from receiving recommended vaccines. Previous studies have used surveys or interviews to solicit parents opinions towards childhood vaccination. In this study, I instead leveraged the online news comments to automatically extract users opinions about the purported link between autism and vaccines. I used two computational methods, topic modeling and text summarization, to characterize the divided opinions on the topic. Both methods yielded higher coverage of public concerns of vaccination compared to previous qualitative studies. The results demonstrate that social media content such as online news comments provide a useful source of information for understanding the public's reaction to important public health issues such as the linkage between vaccination and autism.

5.2 Public Response to Obamacare on Twitter

5.2.1 Introduction

The Affordable Care Act (ACA) is one of the most significant and controversial healthcare reform efforts in the US history. Monitoring public response to new laws and regulations, such as those included in the ACA, is of considerable interest to health policymakers, government agencies, and the media. Traditionally, measuring public response has relied on expensive and time-consuming surveys administered by polling agencies such as the Pew Research Center and the Kaiser Family Foundation. However, the advent of social media introduces new opportunities for tracking public response. While the use of social media data has some limitations,(*Mitchell and Hitlin*, 2013) it is inexpensive, immediate, and can offer more contextual insights not captured by survey questionnaires. Therefore, I conducted a study to explore the use of Twitter for measuring public response to the rollout of the ACA and compared it conventional polling data collected by the Kaiser Family Foundation Health Tracking Poll.[1]

5.2.2 Methods

5.2.2.1 Twitter Data

The study employed a dataset consisting of 10% of all tweets from July 10, 2011 to July 31, 2015 collected using Twitter Gardenhose streaming API. To retrieve relevant tweets relate to the ACA, we developed a list of key search terms. We surveyed three resources, namely Google Trends[2], Wikipedia PPACA page[3], and a random sample of comments to the ACA lay media articles, to explore the words and phrases commonly used online when discussing the ACA. We also expanded our list by adding two ACA-

[1] `http://kff.org/interactive/tracking-opinions-aca/`
[2] `https://www.google.com/trends`
[3] `http://en.wikipedia.org/wiki/Patient_Protection_and_Affordable_Care_Act`

related hashtags. Table 5.7 lists the final search terms and hashtags.

Search Terms and Hashtags
affordable care act or ACA
healthcare act or bill
healthcare insurance act or bill
healthcare insurance exchanges
healthcare reform
obamacare
patient protection and affordable care act or PPACA
#ACA
#obamacare

Table 5.7: Search terms and hashtags used to identify tweets about the Affordable Care Act

To check the validity of this method for identifying ACA tweets, we pulled a random sample of 100 tweets from our final sample of ACA tweets. Two separate members of the research team reviewed the tweets to determine if indeed it was relevant to the ACA. Only six of these tweets were not in fact ACA-related.

5.2.2.2 Sentiment of ACA Tweets

This study used lexicon-based sentiment analysis to assign each ACA tweet a measure of positive to negative sentiment. Lexicon-based sentiment analysis uses a dictionary of sentiment words and phrases each with previously assigned numeric measures of emotion to determine the sentiment of a document.(*Liu*, 2012) Specifically, the study used a lexicon called 'language assessment by Mechanical Turk 1.0 (labMT 1.0)'.(*Dodds et al.*, 2011) It was developed based on a list of most frequent words used in Twitter, Google Books, music lyrics, and the New York Times rated by Amazon Mechanical Turk contributors. Words with highest sentiment scores include laughter, happiness and love. Terrorist and suicide are among the most negative words. Given the scores of all sentiment words in a Tweet, this study followed the original paper and computed the sentiment score of a Tweet as

$$score(T) = \frac{\sum_{i=1}^{N} score(w_i) f_i}{\sum_{i=1}^{N} f_i}, \tag{5.2}$$

where f_i is the frequency of the ith word in a tweet. Since some of the words in our actual query are words associated with sentiment scores (e.g. "care" and "bill"), we excluded our query terms from the assignment of sentiment score for both lexicons.

5.2.2.3 Kaiser Family Foundation Health Tracking Poll

Since the inception of the ACA bill in March 2010, the Kaiser Family Foundations (KFF) Health Tracking Poll has been conducted monthly to evaluate the public views of the ACA. From the KFF Poll data, we were able to determine the percent of respondents who reported being favorable versus unfavorable towards the ACA by month. Five months of KFF data were missing from the data source (December 2012, January 2013, May 2013, July 2013, August 2014, and February 2015). For these months we assigned them the average percent based on the entire time period.

5.2.2.4 Correlation Analysis

We used Spearman correlation to evaluate the associations between public response measured using ACA-relevant tweets and using the KFF Poll. As young adults tend to use Twitter more often than older adults,[4] we also examined correlations stratified by the age of KFF Poll respondents. We performed a sub-analysis to test the robustness of the associations we observed to determine whether tweets from political and special interest groups impacted our results. To do so we re-analyzed the associations after excluding clearly political ACA tweets including hashtags such as #teapart, #p2 (Progressive 2.0), #PJNET (Patriot Journalist Network), and #tcot (Top Conservatives on Twitter).

[4] `http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/`

5.2.3 Results

Overall, 3,375,781 tweets matched one of our query terms or hashtags in the four-year span. Most frequently appearing words and hashtags are listed in Table 5.8.

Word	Frequency
obamacare	2,535,879
aca	548,030
care	337,945
obama	336,952
health	316,792
insurance	170,049
gop	166,212
tcot	157,452
people	156,438
affordable	137,096

(a) words

Hashtag	Frequency
#Obamacare	484,316
#ACA	179,018
#tcot	173,818
#ObamaCare	146,056
#obamacare	79,558
#p2	47,573
#PJNET	41,663
#aca	35,656
#pjnet	34,419
#teaparty	34,259

(b) hashtags

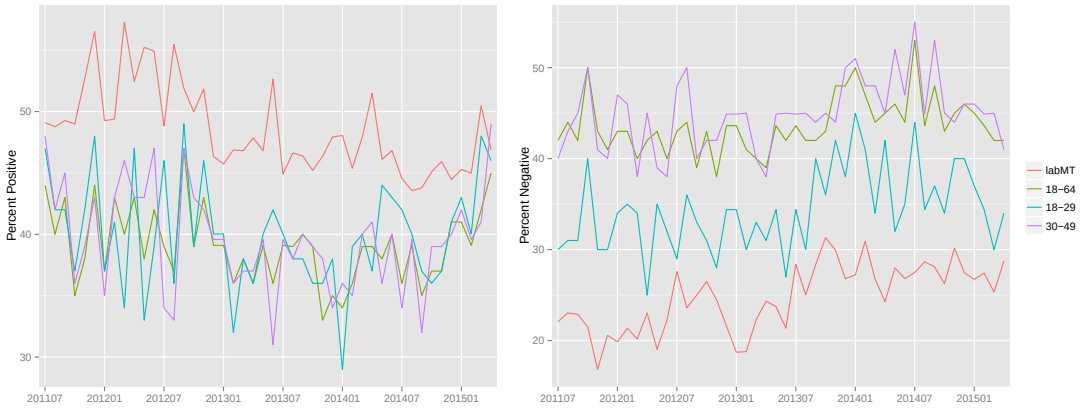
Table 5.8: Most frequent words and hashtags

Age Group	Positive Percentage		Negative Percentage	
	No Lag	1-month Lag	No Lag	1-month Lag
18-64	0.31**	**0.41*** **	0.38***	0.47***
18-29	0.12	0.33**	0.45***	**0.49*** **
30-49	0.32**	0.33**	0.38***	0.45***
50-64	0.21	0.27**	0.03	0.20
65+	0.23	0.15	0.03	0.07

Significant at the: * 10%, ** 5%, and *** 1% levels.

Table 5.9: Correlation test results between time series of Kaiser polling results and tweets sentiment score

Table 5.9 summarizes the correlation test results between the ACA tweets sentiment and the KFF Poll results in different configurations. In general, public response to the ACA on Twitter correlated with the KFF Poll results among majority KFF respondents (age 18-64). The correlation coefficient between percentage of negative ACA tweets and percentage of unfavorable KFF respondents in this age group was 0.38, p-value < 0.01 over the study time period. Likewise, the correlation coefficient between percentage of positive ACA tweets and percentage of favorable KFF



(a) (b)

Figure 5.1: Positive (a) versus negative (b) public response to the Affordable Care Act using tweets compared to results from the Kaiser Family Foundation Poll.

respondents in this age group was 0.31, p-value < 0.05.

When examined by the age category of the KFF respondents, public response to the ACA on Twitter correlated strongest with younger adults. (Figure 5.1) Correlation coefficients for positive and negative views between the two approaches were 0.32 and 0.38 respectively for KFF respondents between 30 to 49 years old; whereas among older KFF respondents, correlations were weak and statistically insignificant. By KFF respondent age category, the strongest correlation was for unfavorable public response (i.e., between percent negative ACA tweets and percent unfavorable KFF respondents) among 18 to 29 year old.

Considering the possible time lag of KFF Poll results, we also computed the correlations between Twitter sentiment with one month lag and the Kaiser Poll results. Table 5.9 shows that Twitter data with one month lag almost always have higher correlation coefficients than those without time lag.

Finally, the results of our sub-analysis show that excluding tweets from political and special interest groups have no impact to the correlation test results.

5.2.4 Conclusion

Overall, we found evidence that Twitter data can be leveraged to effectively estimate public response (and reaction to) specific events such as healthcare reform. The overall positive/negative response measured by tweet sentiment is comparable to the results of the KFF Poll. Furthermore, tweet sentiment correlates better with the KFF Poll results among younger age groups. Adding one month lag to the tweet sentiment time series achieves higher correlations.

5.3 Summary

In this chapter, I applied text-mining and natural language processing techniques including topic models, text summarization and sentiment analysis on news comments and Twitter data to understand public opinions on controversial public health topics. I also demonstrated the effectiveness of these tools by evaluating the results on human-annotated data and public survey results.

CHAPTER VI

System Design: News Comments Analyzer

According to a 2013 Pew survey[1], 50% of all Americans and 71% of the young people between 18-29 use the internet as their main source for news. Different from traditional sources of news such as newspapers and television, online news readers can actively leave comments related to a news article and others may reply to a comment a user posted through the social media component many news websites have incorporated. This functionality provides an interactive platform for users across the world to share and exchange their personal opinions about a news story and its related policy implications. The news commenting feature has become so popular that it is not uncommon to have thousands of reader-contributed comments below a headline news article on a popular news website. Such comments provide us a unique source of information to solicit public opinions in real time with very low cost. The insights generated from user comments can further inform public policy decision making and targeted education on common misperceptions and concerns about important public health issues.

In Section 5.1, I have demonstrated the effectiveness of understanding public opinions regarding vaccine and autism linkage from analyzing online news comments. I specifically showed that text summarization and topic modeling can be applied effectively to distill key opinions from a large number of user comments. In Section 5.2, I

[1] http://www.pewresearch.org/fact-tank/2013/10/16/12-trends-shaping-digital-news/

further applied sentiment analysis to gauge public opinion on the ACA. The output of three algorithms were compared against traditional survey or survey study results. High coverage or strong correlation were found in all cases.

Despite the potential, text-mining toolkits, in addition to news comments data are not easily accessible to public policy and public health researchers. In this chapter, I present a news comments analyzing system that I designed and developed, which automates the data collection and data analysis process to facilitate the utilization of this valuable information.[2]

6.1 System Design

In this section I describe the design and implementation of the news comments analyzer system. Figure 6.1 shows the architecture of my system. There are two major components of the system - a data collection component that collects news articles and reader comments from major news websites, and an analytical component that preprocesses users comments and applies text-mining techniques to analyze the data collected. Details of two components are described below.

6.1.1 The Data Collection Component

To facilitate users to identify related news articles efficiently, the system provides a news search engine which allows users to search for news articles on different websites. The news search engine leverages Google News Search API[3] to retrieve a list of relevant news articles based on user's query. The system currently supports eight popular news websites (shown in Figure 6.2) in the United States that provide user commenting function. Then the system can aide users to further select relevant articles from the retrieved news article list, as shown in Figure 6.3. A crawler built into

[2] `http://newton.si.umich.edu/owenliu/news_comments` (University of Michigan VPN required)

[3] `https://developers.google.com/news-search/?hl=en`

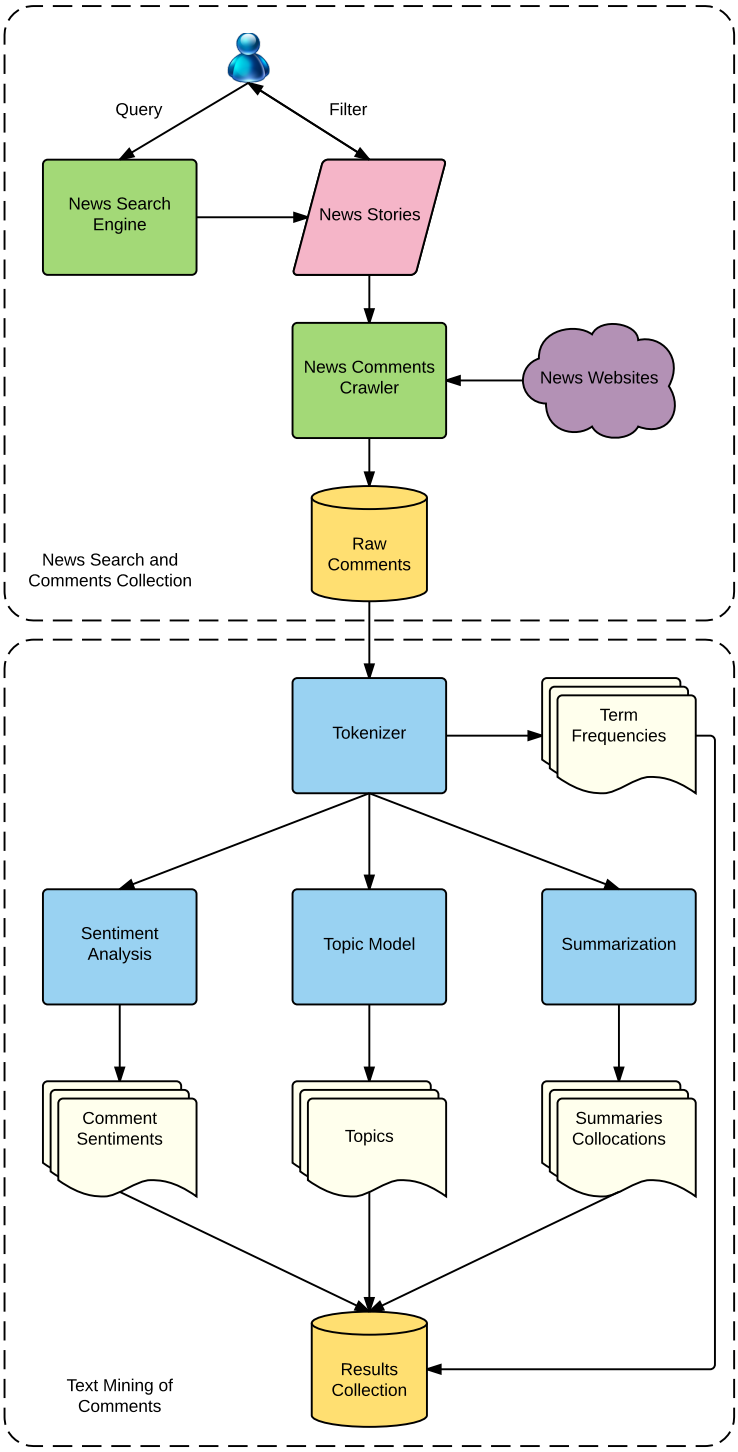


Figure 6.1: The system architecture of news comments analyzer

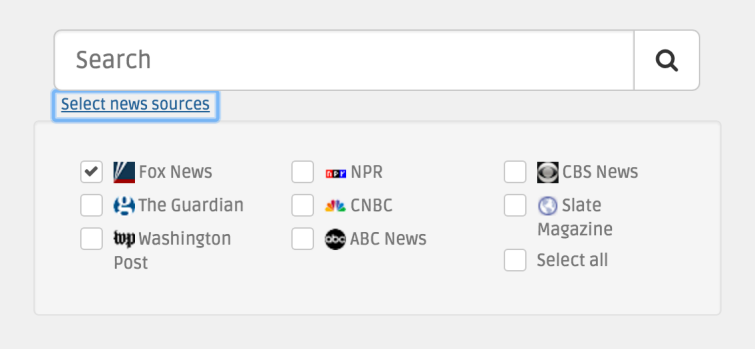


Figure 6.2: Screenshot of the news search interface

the system will then efficiently download all the comments along with the metadata of all news articles the user has selected. The raw comments will be stored in JSON format to be used as input to the analytical engine.

6.1.2 The Analytical Engine Based on Text Mining

Preprocessing: After all raw comments are downloaded by the crawler, a preprocessing module will first split the comments into sentences. It will then tokenize and lemmatize words in each sentence. Stop words will also be removed during this process. A table of lemmatized words and their term frequencies will be generated for users to quickly examine most common words in the collection.

Sentiment Analysis: This module performs lexicon-based sentiment analysis of all comments collected to help users estimate the sentiment of public opinions of selected news stories. Two lexicons are currently built into the system: labMT 1.0 (*Dodds et al.*, 2011) and the lexicon provided in pattern.en package (*De Smedt and Daelemans*, 2012). Each comment will be assigned a sentiment score, which is determined by the weighted average of sentiment scores of all sentiment words in the comment (equation 5.2) provided by one of the lexicons. To make the comment sentiment scores determined by two lexicons comparable, the system will normalize both scores to the range of -1 to 1, with -1 represents most negative sentiment and 1 represents most positive sentiment. The system also allows users to sort the comments

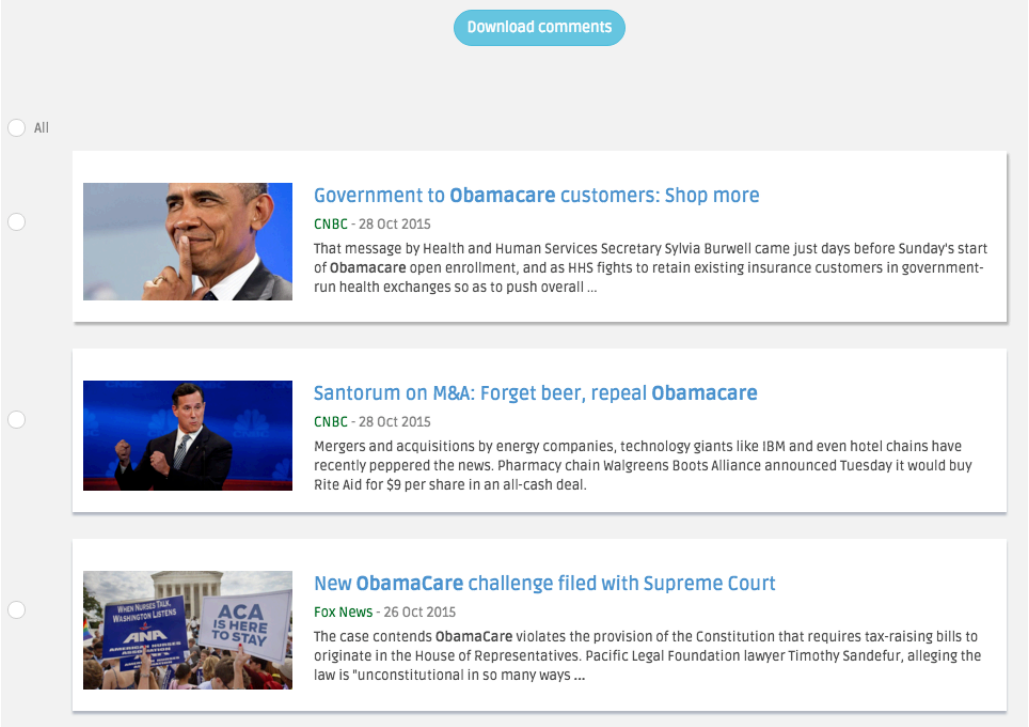


Figure 6.3: Screenshot of the news selection interface. Users can select a news article by checking the checkbox in front of the news article panel.

based on one of the sentiment scores to review the comments using most positive or negative sentiment words. Finally, the system provides a stacked histogram of sentiment scores based on two lexicons to help user understand the overall distribution of sentiment scores of all comments collected. Figure 6.4 shows a histogram of sentiment scores of comments on recent news articles about Obamacare.

Topic Modeling: The topic modeling module is designed to help users explore

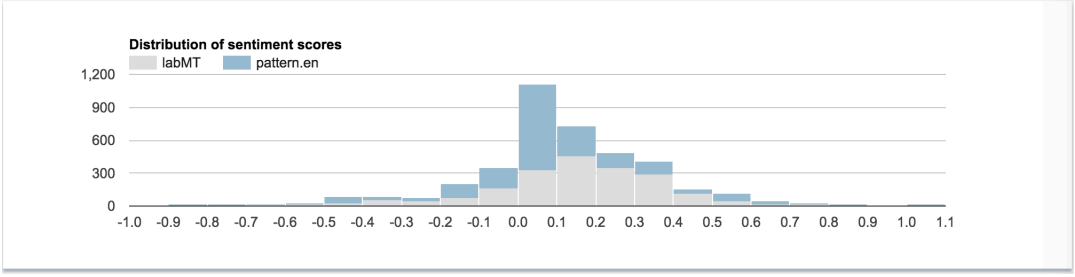


Figure 6.4: A histogram of sentiment scores of comments on news articles about Obamacare.

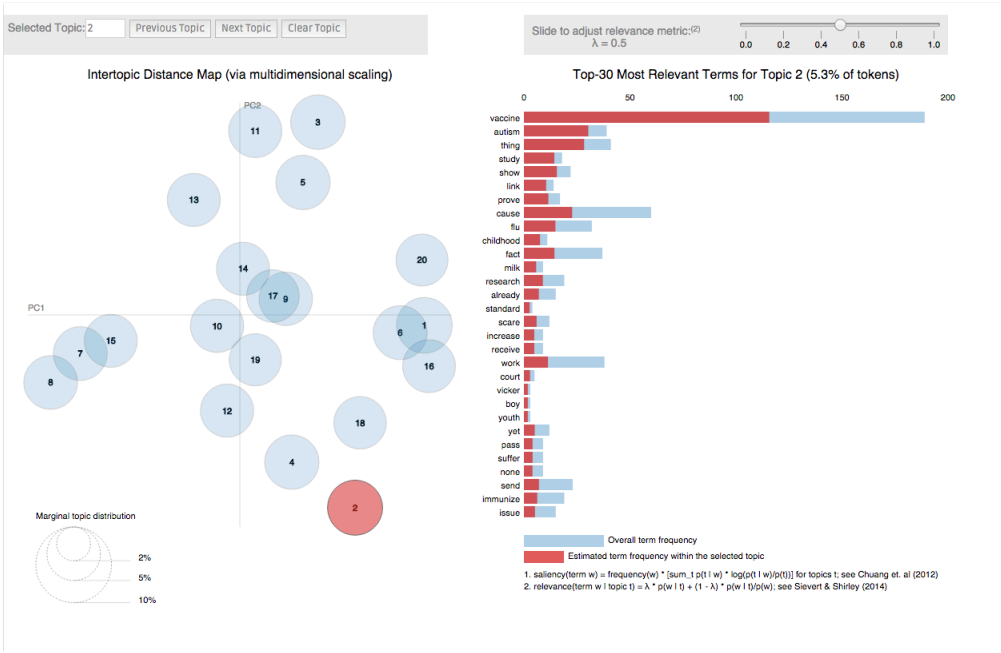


Figure 6.5: LDAvis visualization of LDA results of MMR vaccination news comments

the topics which online news readers expressed in their comments. Specifically, it will apply the LDA topic model (*Blei et al.*, 2003) to the collection of comments. A description of topic modeling and LDA can be found in Section 5.1. Based on the output of LDA, the system will output the top ten topic words and group comments based on their dominant topic. To help user better interpret and evaluate the topic model results, the system further incorporates an interactive visualization of LDA results developed by *Sievert and Shirley* (2014). As shown in Figure 6.5, it visualizes the semantic differences between topics and displays the top topic terms based on a so-called "relevance" metric to help users better interpret the topics by punishing universal common terms that have high probabilities to be generated by all topics.

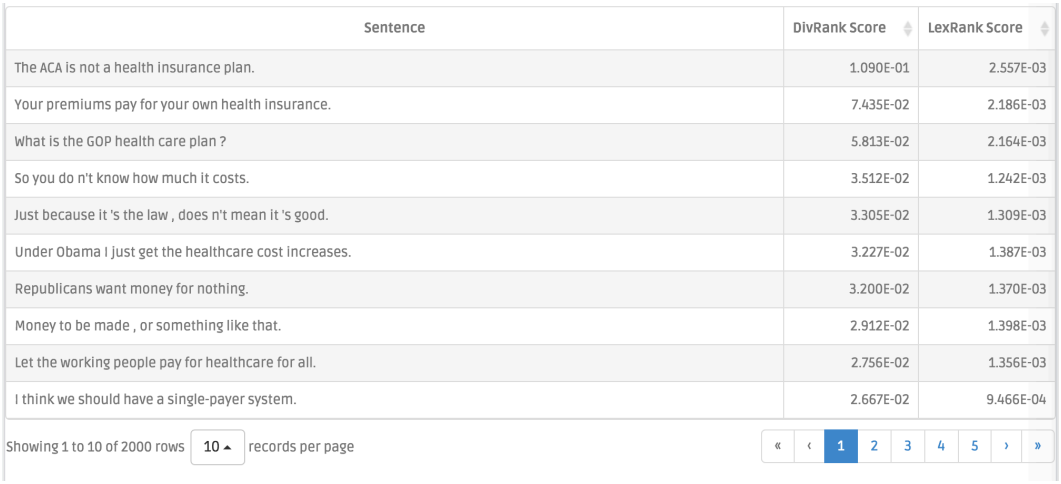
Text summarization: Text summarization module provides users with representative public opinions from collected news comments. It will generate a short summarization by selecting most important sentences among all comments. Two graph-based summarization methods, namely LexRank (*Erkan and Radev*, 2004) and DivRank (*Mei et al.*, 2010), are implemented in the system. Graph-based text sum-

marization methods first generate a graph of sentences based on sentence similarity and rank the sentences by preferred centrality measures. One disadvantage of these methods is that they require computing pairwise similarities between all sentences in quadratic time, which can be resource consuming. To speed up this process, the system will first select a small subset of most informative sentences and only compute pairwise similarities between the sentences in this subset. Similar to *Hsu et al.* (2009), the informativeness of a sentence is defined by:

$$inform(c_j) = \sum_{t_i \in c_j} tf_{i,j} \times idf_i, \tag{6.1}$$

where $tf_{i,j}$ is the normalized term frequency of term i in comment j and idf_i is the inverse document frequency of term i. Different from *Hsu et al.* (2009), L2 norm is used here to normalize term frequency instead of L1 norm which favors very short sentences with few terms with high inverse document frequency.

Two graph-based summarization methods will then be applied to rank the subset of most informative sentences. Details about two algorithms can be found in Section 5.1. Figure 6.6 shows the most representative sentences extracted from a collection of Obamacare news comments along with ranking scores determined by DivRank and LexRank.



Sentence	DivRank Score	LexRank Score
The ACA is not a health insurance plan.	1.090E-01	2.557E-03
Your premiums pay for your own health insurance.	7.435E-02	2.186E-03
What is the GOP health care plan ?	5.813E-02	2.164E-03
So you do n't know how much it costs.	3.512E-02	1.242E-03
Just because it 's the law , does n't mean it 's good.	3.305E-02	1.309E-03
Under Obama I just get the healthcare cost increases.	3.227E-02	1.387E-03
Republicans want money for nothing.	3.200E-02	1.370E-03
Money to be made , or something like that.	2.912E-02	1.398E-03
Let the working people pay for healthcare for all.	2.756E-02	1.356E-03
I think we should have a single-payer system.	2.667E-02	9.466E-04

Figure 6.6: Top sentences from Obamacare news comments based on DivRank and LexRank

CHAPTER VII

Conclusions

Health-related information has become increasingly popular in social media. It presents enormous opportunities for public health researchers and policy makers to monitor population health or solicit public opinion effectively and efficiently. However, retrieving relevant information from heterogeneous sources of social media data remains challenging, and whether useful insights can be generated from short, informal texts in social media using computational methods is still unclear.

In this thesis, I first examine users' identities and their intent to participate in online health-related discussions (Chapter III). The results shed light on the availability and characteristics of health-related information in social media, which becomes a motivation to explore how to best use such information from Twitter (Chapter IV). Through a case study of eye-related disease, I then investigate the performance of a state-of-the-art medical NLP tool, MetaMap, on social media data, and improve its accuracy by using machine-learning classifier to further filter out medically-irrelevant information. In the subsequent two chapters, I demonstrate the effectiveness of distilling public opinions on controversial medical or ethical issues from online news comments using various text-mining and natural language processing techniques, and report the development of a system that automates the process to facilitate researchers collecting and analyzing such data.

As health-related information become ubiquitous in social media, one key factor to the success of leveraging this valuable source of information at the population level is to address potential sampling biases appropriately. Researchers should investigate the user population and validate the results with other data sources before making important inferences. Finally, while this thesis mainly focuses on textual form of health-related information in social media authored by users themselves, in the future I hope to explore other types of information, such as digital health status logs published automatically by mobile devices as they become more available.

APPENDIX

APPENDIX A

Most Frequent UMLS Concepts of Signs/Symptoms Identified by MetaMap

Concept	Frequency
malaise	826608
tired	449258
flatulence	368211
pallor	330857
catch - finding of sensory dimension of pain	320312
chills	272620
hunger	241165
seizures	238400
pain	198435
sore to touch	184845
clumsiness	148244
tremor	128794
laziness	120345
gastrointestinal gas	103886

asthenia	102999
nervousness	89285
headache	84383
thirsty	71328
sighing respiration	70504
spells (neurological symptom)	67957
signs and symptoms	56782
hangover from alcohol	51246
whooping respiration	44084
coughing	43890
visual halos (disorder)	42911
exhaustion	42722
breathiness	41954
burning sensation	30861
sleeplessness	29069
muscle twitch	28817
charmed	25117
breath-holding spell	19179
snoring	18978
moaning	18157
pruritus	17874
spots on skin	16979
blackout - symptom	16364
muscle cramp	15612
ache	14447
retching	14270
harsh voice quality	14144

blushing	13787
dizziness	12728
drooling	12182
syncope	12030
fatigue	11861
vomiting	11157
gait, drop foot	9430
sore throat	8530
stomach ache	7926
flushing	7861
bad dreams	7322
nausea	7062
eructation	7049
early waking	6818
agitation	6612
clubbing	6480
feeling tense	6406
feeling sick	6110
photopsia	5705
workaholic	5293
sick to stomach	5096
welts	4919
red nose	4843
breathing abnormally deep	4762
flare	4458
lividity	4444
gasping for breath	4269

trembling	4260
back pain	4219
out	4200
paraneoplastic opsoclonus ataxia	4070
pregnancy mood swing	3854
rundown	3720
heartburn	3609
symptoms	3471
giddy mood	3445
oversleeps	3411
overweight	3404
redness of eye	3108
indifferent mood	3085
toothache	3043
suffocated	3041
body ache	2835
failure to gain weight	2651
chest pain	2582
diarrhea	2545
feeling despair	2443
nasal congestion (finding)	2429
halitosis	2368
menstrual spotting	2171
rhinorrhea	2144
drugged state	2089
grunting respiration	1993
constipation	1925

lightheadedness	1900
tired feeling	1809
rales	1781
abdominal bloating	1755
hallucinations, visual	1667
shaking of hands	1578
dyspnea	1540
fumbling	1516
forgetful	1385
floppy	1383
rolling of eyes	1350
cluttering	1318
perfectionism	1278
stinging sensation	1265
jitteriness	1201
spasmodic movement	1164
has tingling sensation	1160
sluggishness	1144
spastic	1005
earache	1001
spasm	996
circling gait	991
flasher - visual manifestation	979
pyrexia of unknown origin (excl puerperal)	973
withdrawal symptoms	927
taste sweet	904
urinary hesitation	904

sensory discomfort	902
abdominal colic	900
giving-way	887
growing pains	867
coarse hair	836
wheezing	817
jumpiness	762
morning sickness	759
put weight	758
hot flushes	741
greasy hair	733
redness of face	722
hoarseness	679
cardiac pain	677
feeling cold	664
feeling strange	661
muscle rigidity	655
cold sweat	648
loss of scalp hair	648
rigor - temperature-associated observation	648
memory loss	642
unrest	638
dry skin	624
heart problem	614
myalgia	606
hypersomnia	598
leg cramps	592

abdominal pain	564
sore eye	563
eye swelling	556
stomach cramps (finding)	535
headache associated with sexual activity	530
watery eyes	516
the runs	511
bleached hair	506
lethargy	498
lump in throat	490
blurred vision	484
grimaces	483
verbal auditory hallucinations	463
sharp pain	462
chapping of lips	459
head ache	456
gassiness	450
eyes twitching	437
imbalance	433
neck pain	424
pounding in head	422
feeling hot	414
dyspepsia	410
sleep disturbances	404
xerostomia	394
crowning	385
neck stiffness	385

blanching	384
change voice	375
cramp in foot	373
sore back	370
dryness of eye	353
high weight	351
projectile vomiting	351
knee pain	351
abnormal coordination	351
physical appearance	348
pain in back	348
hunger pain	341
excruciating pain	338
low back pain	337
numbness	326
hemoptysis	325
mannerism	323
abnormal breathing	320
night pain	320
pain in eyes	318
f.u.o.	312
flaccid muscle tone	310
feeling dizzy	307
sense smell	303

BIBLIOGRAPHY

BIBLIOGRAPHY

Ache, K. A., and L. S. Wallace (2008), Human papillomavirus vaccination coverage on youtube, *American journal of preventive medicine*, *35*(4), 389–392.

Achrekar, H., Z. Fang, Y. Li, C. Chen, B. Liu, and J. Wang (2011), *A spatio-temporal approach to the discovery of online social trends*, pp. 510–524, Springer.

Ahmed, O. H., S. J. Sullivan, A. G. Schneiders, and P. Mccrory (2010), isupport: do social networking sites have a role to play in concussion awareness?, *Disability and rehabilitation*, *32*(22), 1877–1883.

Alias-i (2008), Lingpipe 4.1.0.

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

Aronson, A. R. (2006), Metamap: Mapping text to the umls metathesaurus, *Bethesda, MD: NLM, NIH, DHHS*, pp. 1–26.

Bazzano, A., A. Zeldin, E. Schuster, C. Barrett, and D. Lehrer (2012), Vaccine-related beliefs and practices of parents of children with autism spectrum disorders, *Am J Intellect Dev Disabil*, *117*(3), 233–42, doi:10.1352/1944-7558-117.3.233.

Bender, J. L., M.-C. Jimenez-Marroquin, and A. R. Jadad (2011), Seeking support on facebook: a content analysis of breast cancer groups, *Journal of medical Internet research*, *13*(1).

Bernard, H. R. (2012), *Social research methods: Qualitative and quantitative approaches*, Sage Publications, Incorporated.

Bhattacharya, S., H. Tran, and P. Srinivasan (2012), Discovering health beliefs in twitter., in *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*.

Bian, J., U. Topaloglu, and F. Yu (2012), Towards large-scale twitter mining for drug-related adverse events, in *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pp. 25–32, ACM.

Bilge, U., S. Bozkurt, B. Yolcular, and D. Ozel (2012), Can social web help to detect influenza related illnesses in turkey?, *Studies in health technology and informatics*, *174*, 100.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), Latent dirichlet allocation, *J. Mach. Learn. Res.*, *3*, 993–1022.

Bollen, J., A. Pepe, and H. Mao (2011), Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 450–453.

Bosley, J. C., N. W. Zhao, S. Hill, F. S. Shofer, D. A. Asch, L. B. Becker, and R. M. Merchant (2013), Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication, *Resuscitation*, *84*(2), 206–212.

Briones, R., X. Nan, K. Madden, and L. Waks (2012), When vaccines go viral: an analysis of hpv vaccine coverage on youtube, *Health communication*, *27*(5), 478–485.

Brownstein, J. S., C. C. Freifeld, B. Y. Reis, and K. D. Mandl (2008), Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project, *PLoS Med*, *5*(7), e151.

Burton, S., R. Morris, M. Dimond, J. Hansen, C. Giraud-Carrier, J. West, C. Hanson, and M. Barnes (2012a), Public health community mining in youtube, in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 81–90, ACM.

Burton, S. H., K. W. Tanner, C. G. Giraud-Carrier, J. H. West, and M. D. Barnes (2012b), "right time, right place" health communication on twitter: value and accuracy of location information, *Journal of medical Internet research*, *14*(6).

Cairns, P., and A. L. Cox (2008), *Research methods for human-computer interaction*, Cambridge University Press.

Caplan, A. L. (2009), Retractionileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, *Wkly Epidemiol Rec*, *84*, 301–08.

Capstrat (2010), Health care information – where do you go? who do you trust?, in *Capstrat Public Policy Poll, conducted Apr 2010*.

Carneiro, H. A., and E. Mylonakis (2009), Google trends: a web-based tool for real-time surveillance of disease outbreaks, *Clinical infectious diseases*, *49*(10), 1557–1564.

Carroll, M. V., A. Shensa, and B. A. Primack (2012), A comparison of cigarette-and hookah-related videos on youtube, *Tobacco control*, pp. tobaccocontrol–2011.

Casiday, R., T. Cresswell, D. Wilson, and C. Panter-Brick (2006), A survey of uk parental attitudes to the mmr vaccine and trust in medical authority, *Vaccine*, *24*(2), 177–184, doi:http://dx.doi.org/10.1016/j.vaccine.2005.07.063.

Chen, L., H. Achrekar, B. Liu, and R. Lazarus (2010), Vision: towards real time epidemic vigilance through online social networks: introducing sneft–social network enabled flu trends, in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond*, p. 4, ACM.

Chen, Y.-W., and C.-J. Lin (2006), Combining svms with various feature selection strategies, in *Feature extraction*, pp. 315–324, Springer.

Chew, C., and G. Eysenbach (2010), Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak, *PLoS One*, *5*(11), e14,118.

Chou, W.-Y. S., Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse (2009a), Social media use in the United States: Implications for health communication, *Journal of Medical Internet Research*, *11*(4), e48.

Chou, W. Y. S., Y. M. Hunt, E. B. Beckjord, R. P. Moser, and B. W. Hesse (2009b), Social media use in the united states: Implications for health communication, *Journal of Medical Internet Research*, *11*(4).

Chou, W. Y. S., Y. Hunt, A. Folkers, and E. Augustson (2011), Cancer survivorship in the age of youtube and social media: A narrative analysis, *Journal of Medical Internet Research*, *13*(1), 108–116.

Chuang, K. Y., and C. C. Yang (2012), Interaction patterns of nurturant support exchanged in online health social networking, *Journal of Medical Internet Research*, *14*(3), e54.

Chunara, R., J. R. Andrews, and J. S. Brownstein (2012), Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak, *The American Journal of Tropical Medicine and Hygiene*, *86*(1), 39–45.

Clerici, C. A., L. Veneroni, G. Bisogno, A. Trapuzzano, and A. Ferrari (2012), Videos on rhabdomyosarcoma on youtube: an example of the availability of information on pediatric tumors on the web, *J Pediatr Hematol Oncol*, *34*(8), e329–31.

Cohen, J., et al. (1960), A coefficient of agreement for nominal scales, *Educational and psychological measurement*, *20*(1), 37–46.

Colineau, N., and C. Paris (2010), Talking about your health to strangers: understanding the use of online social networks by patients, *New Review of Hypermedia and Multimedia*, *16*(1-2), 141–160.

Collier, N., N. T. Son, and N. M. Nguyen (2011), Omg u got flu? analysis of shared health messages for bio-surveillance, *J Biomed Semantics*, *2 Suppl 5*, S9.

Collier, N., et al. (2008), Biocaster: detecting public health rumors with a web-based text mining system, *Bioinformatics*, *24*(24), 2940–2941.

Corley, C. D., D. J. Cook, A. R. Mikler, and K. P. Singh (2010), Text and structural data mining of influenza mentions in web and social media, *Int J Environ Res Public Health*, *7*(2), 596–615.

Cortes, C., and V. Vapnik (1995), Support-vector networks, *Machine learning*, *20*(3), 273–297.

Crutzen, R., J. de Nooijer, W. Brouwer, A. Oenema, J. Brug, and N. K. de Vries (2011), Strategies to facilitate exposure to internet-delivered health behavior change interventions aimed at adolescents or young adults: a systematic review, *Health Educ Behav*, *38*(1), 49–62.

Culotta, A. (2010), Towards detecting influenza epidemics by analyzing twitter messages, in *Proceedings of the first workshop on social media analytics*, pp. 115–122, ACM.

Culotta, A. (2012), Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages, *Language Resources and Evaluation*, pp. 1–22.

Dales, L., S. J. Hammer, and N. J. Smith (2001), Time trends in autism and in mmr immunization coverage in california, *Jama*, *285*(9), 1183–5.

De la Torre-Diez, I., F. J. Diaz-Pernas, and M. Anton-Rodriguez (2012), A content analysis of chronic diseases social groups on facebook and twitter, *Telemed J E Health*, *18*(6), 404–8.

De Smedt, T., and W. Daelemans (2012), Pattern for python, *The Journal of Machine Learning Research*, *13*(1), 2063–2067.

Deer, B. (2011a), Secrets of the mmr scare . how the vaccine crisis was meant to make money, *Bmj*, *342*, c5258, doi:10.1136/bmj.c5258.

Deer, B. (2011b), How the case against the mmr vaccine was fixed, *BMJ*, *342*.

Deluca, P., et al. (2012), Identifying emerging trends in recreational drug use; outcomes from the psychonaut web mapping project, *Prog Neuropsychopharmacol Biol Psychiatry*, *39*(2), 221–6.

Denecke, K., and A. Stewart (2011), *Learning from Medical Social Media Data: Current State and Future Challenges*, Social Media Tools and Platforms in Learning Environments, 353-372 pp.

Denecke, K., P. Dolog, and P. Smrz (2012), Making use of social media data in public health, in *Proceedings of the 21st international conference companion on World Wide Web*, pp. 243–246, ACM.

DeStefano, F. (2007), Vaccines and autism: evidence does not support a causal association, *Clin Pharmacol Ther*, *82*(6), 756–9, doi:10.1038/sj.clpt.6100407.

DeStefano, F., T. K. Bhasin, W. W. Thompson, M. Yeargin-Allsopp, and C. Boyle (2004), Age at first measles-mumps-rubella vaccination in children with autism and school-matched control subjects: a population-based study in metropolitan atlanta, *Pediatrics*, *113*(2), 259–66.

Diaz-Aviles, E., and A. Stewart (2012), Tracking twitter for epidemic intelligence: case study: Ehec/hus outbreak in germany, 2011, in *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 82–85, ACM.

Diaz-Aviles, E., A. Stewart, E. Velasco, K. Denecke, and W. Nejdl (), Epidemic intelligence for the crowd, by the crowd, in *Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 2012)*.

Diaz-Aviles, E., A. Stewart, E. Velasco, K. Denecke, and W. Nejdl (2012), Towards personalized learning to rank for epidemic intelligence based on social media streams, in *Proceedings of the 21st international conference companion on World Wide Web*, pp. 495–496, ACM.

Divecha, Z., A. Divney, J. Ickovics, and T. Kershaw (2012), Tweeting about testing: do low-income, parenting adolescents and young adults use new media technologies to communicate about sexual health?, *Perspect Sex Reprod Health*, *44*(3), 176–83.

Dobson, R. (2003), Media misled the public over the mmr vaccine, study says, *Bmj*, *326*(7399), 1107, doi:10.1136/bmj.326.7399.1107-a.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011), Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter, *PloS one*, *6*(12), e26,752.

Doing-Harris, K. M., and Q. Zeng-Treitler (2011), Computer-assisted update of a consumer health vocabulary through mining of social network data, *Journal of Medical Internet Research*, *13*(2).

Dredze, M. (2012), How social media will change public health, *Intelligent Systems, IEEE*, *27*(4), 81–84.

Dreesman, J., and K. Denecke (2011), Challenges for signal generation from medical social media data, *Stud Health Technol Inform*, *169*, 639–43.

Dumbrell, D., and R. Steele (2012), What are the characteristics of highly disseminated public health-related tweets?, in *Proceedings of the 24th Australian Computer-Human Interaction Conference*, pp. 115–118, ACM.

Egan, K. G., and M. A. Moreno (2011), Prevalence of stress references on college freshmen facebook profiles, *Comput Inform Nurs*, *29*(10), 586–92.

Elkin, L., G. Thomson, and N. Wilson (2010), Connecting world youth with tobacco brands: Youtube and the internet policy vacuum on web 2.0, *Tob Control*, *19*(5), 361–6.

Erkan, G., and D. Radev (2004), Lexrank: Graph-based lexical centrality as salience in text summarization, *J. Artif. Intell. Res. (JAIR)*, *22*, 457–479.

Eysenbach, G. (2006), Infodemiology: tracking flu-related searches on the web for syndromic surveillance, *AMIA Annu Symp Proc*, pp. 244–8.

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

Farmer, A. D., C. E. Bruckner Holt, M. J. Cook, and S. D. Hearing (2009), Social networking sites: a novel portal for communication, *Postgrad Med J*, *85*(1007), 455–9.

Fernandez-Luque, L., N. Elahi, and r. Grajales, F. J. (2009), An analysis of personal medical information disclosed in youtube videos created by patients with multiple sclerosis, *Stud Health Technol Inform*, *150*, 292–6.

Fernandez-Luque, L., R. Karlsen, and J. Bonander (2011a), Review of extracting information from the social web for health personalization, *J Med Internet Res*, *13*(1), e15.

Fernandez-Luque, L., R. Karlsen, and G. B. Melton (2011b), Healthtrust: trust-based retrieval of you tube's diabetes channels, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1917–1920, ACM.

Fisher, J., and M. Clayton (2012), Who gives a tweet: assessing patients' interest in the use of social media for health care, *Worldviews Evid Based Nurs*, *9*(2), 100–8.

Fox, S. (2011), *The social life of health information 2011*, Pew Internet & American Life Project Washington, DC.

Freeman, B., and S. Chapman (2010), British american tobacco on facebook: undermining article 13 of the global world health organization framework convention on tobacco control, *Tob Control*, *19*(3), e1–9.

Frohlich, D. O., and A. Zmyslinski-Seelig (2012), The presence of social support messages on youtube videos about inflammatory bowel disease and ostomies, *Health Commun*, *27*(5), 421–8.

Fry, J. P., and R. A. Neff (2009), Periodic prompts and reminders in health promotion and health behavior interventions: systematic review, *J Med Internet Res*, *11*(2), e16.

Gellin, B. G., E. W. Maibach, and E. K. Marcuse (2000), Do parents understand immunizations? a national telephone survey, *Pediatrics*, *106*(5), 1097–102.

Gill, P. S., and B. Whisnant (2012), A qualitative assessment of an online support community for ovarian cancer patients, *Patient Related Outcome Measures*, *3*, 51–58.

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

Godlee, F., J. Smith, and H. Marcovitch (2011), Wakefield's article linking mmr vaccine and autism was fraudulent, *Bmj*, *342*, c7452, doi:10.1136/bmj.c7452.

Goeuriot, L., J.-C. Na, W. Y. M. Kyaing, S. Foo, C. Khoo, Y.-L. Theng, and Y.-K. Chang (2011), Textual and informational characteristics of health-related social media content: A study of drug review forums.

Goldacre, B. (2007), Medicine and the media: Mmr: the scare stories are back, *Bmj*, *335*(7611), 126–7, doi:10.1136/bmj.39280.447419.59.

Golder, S. A., and M. W. Macy (2011), Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures, *Science*, *333*(6051), 1878–81.

Gonzalez-Bailon, S., R. Banchs, and A. Kaltenbrunner (2010), Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions, *arXiv preprint arXiv:1009.4019*.

Greene, J., N. Choudhry, E. Kilabuk, and W. Shrank (2011), Online social networking by patients with diabetes: A qualitative evaluation of communication with facebook, *Journal of General Internal Medicine*, *26*(3), 287–292.

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

Gruzd, A., F. A. Black, T. N. Le, and K. Amos (2012), Investigating biomedical research literature in the blogosphere: a case study of diabetes and glycated hemoglobin (hba1c), *J Med Libr Assoc*, *100*(1), 34–42.

Guillaume, L., and P. A. Bath (2008), A content analysis of mass media sources in relation to the mmr vaccine scare, *Health Informatics J*, *14*(4), 323–34, doi: 10.1177/1460458208096654.

Gust, D. A., T. W. Strine, E. Maurice, P. Smith, H. Yusuf, M. Wilkinson, M. Battaglia, R. Wright, and B. Schwartz (2004), Underimmunization among children: effects of vaccine safety concerns on immunization status, *Pediatrics*, *114*(1), e16–22.

Hackett, A. J. (2008), Risk, its perception and the media: the mmr controversy, *Community Pract*, *81*(7), 22–5.

Hagan 3rd, J. C., and M. J. Kutryb (2009), Cataract and intraocular implant surgery concerns and comments posted at two internet eye care forums, *Missouri Medicine*, *106*(1), 78–82.

Hallmayer, J., et al. (2011), Genetic heritability and shared environmental factors among twin pairs with autism, *Arch Gen Psychiatry*, *68*(11), 1095–102, doi:10.1001/archgenpsychiatry.2011.76.

Halsey, N. A., and S. L. Hyman (2001), Measles-mumps-rubella vaccine and autistic spectrum disorder: report from the new challenges in childhood immunizations conference convened in oak brook, illinois, june 12-13, 2000, *Pediatrics*, *107*(5), E84.

Heaivilin, N., B. Gerbert, J. E. Page, and J. L. Gibbs (2011), Public health surveillance of dental pain via twitter, *J Dent Res*, *90*(9), 1047–51.

Herman Tolentino, M., M. Raoul Kamadjeu, M. Michael Matters PhD, M. Marjorie Pollack, and M. Larry Madoff (2007), Scanning the emerging infectious diseases horizon-visualizing promed emails using epispider, *Advances in disease surveillance*, *2*, 169.

Hofmann, T. (1999), Probabilistic latent semantic indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, ACM.

Holton, A., B. Weberling, C. E. Clarke, and M. J. Smith (2012), The blame frame: media attribution of culpability about the mmr-autism vaccination scare, *Health Commun*, *27*(7), 690–701, doi:10.1080/10410236.2011.633158.

Hsu, C.-F., E. Khabiri, and J. Caverlee (2009), Ranking comments on the social web, in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4, pp. 90–97, IEEE.

Hussain, H., S. B. Omer, J. A. Manganello, E. E. Kromm, T. C. Carter, L. Kan, S. Stokley, N. A. Halsey, and D. A. Salmon (2011), Immunization safety in us print media, 19952005, *Pediatrics*, *127*(Supplement 1), S100–S106, doi:10.1542/peds.2010-1722O.

Hviid, A., M. Stellfeld, J. Wohlfahrt, and M. Melbye (2003), Association between thimerosal-containing vaccine and autism, *Jama*, *290*(13), 1763–6, doi:10.1001/jama.290.13.1763.

Jamison-Powell, S., C. Linehan, L. Daley, A. Garbett, and S. Lawson (2012), I can't get no sleep: discussing# insomnia on twitter, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1501–1510, ACM.

Jiang, K., and Y. Zheng (2013), Mining twitter data for potential drug effects, in *Advanced Data Mining and Applications*, pp. 434–443, Springer.

Kamel Boulos, M. N., A. P. Sanfilippo, C. D. Corley, and S. Wheeler (2010), Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance, *Comput Methods Programs Biomed*, *100*(1), 16–23.

Kanhabua, N., S. Romano, A. Stewart, and W. Nejdl (2012), Supporting temporal analytics for health-related events in microblogs, in *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2686–2688, ACM.

Katz, J. A., T. Capua, and J. Bocchini, J. A. (2012), Update on child and adolescent immunizations: selected review of us recommendations and literature, *Curr Opin Pediatr*, *24*(3), 407–21, doi:10.1097/MOP.0b013e3283534d11.

Keelan, J., V. Pavri, R. Balakrishnan, and K. Wilson (2010), An analysis of the human papilloma virus vaccine debate on myspace blogs, *Vaccine*, *28*(6), 1535–40.

Keller, M., M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein (2009), Use of unstructured event-based reports for global infectious disease surveillance, *Emerging infectious diseases*, *15*(5), 689.

Kendall, L., A. Hartzler, P. Klasnja, and W. Pratt (2011), Descriptive analysis of physical activity conversations on twitter, in *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 1555–1560, ACM.

Kennedy, A., K. LaVail, G. Nowak, M. Basket, and S. Landry (2011), Confidence about vaccines in the united states: Understanding parents' perceptions, *Health Affairs*, *30*(6), 1151–1159, doi:10.1377/hlthaff.2011.0396.

Kramer, A. D. (2010), An unobtrusive behavioral model of gross national happiness, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 287–290, ACM.

Krieck, M., J. Dreesman, L. Otrusina, and K. Denecke (2011), A new age of public health: Identifying disease outbreaks by analyzing tweets, in *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*.

Kumar, R. A., and S. L. Christian (2009), Genetics of autism spectrum disorders, *Curr Neurol Neurosci Rep*, *9*(3), 188–97.

Kushin, M. J., and M. Yamamoto (2010), Did social media really matter? college students' use of online media and political decision making in the 2008 election, *Mass Communication and Society*, *13*(5), 608–630.

Lamb, A., M. J. Paul, and M. Dredze (2012), Investigating twitter as a source for studying behavioral responses to epidemics, in *AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*.

Lampos, V., and N. Cristianini (2010), Tracking the flu pandemic by monitoring the social web, in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pp. 411–416.

Lampos, V., T. De Bie, and N. Cristianini (2010), *Flu Detector - Tracking Epidemics on Twitter*, *Lecture Notes in Artificial Intelligence*, vol. 6323, pp. 599–602.

Larson, H. J., L. Z. Cooper, J. Eskola, S. L. Katz, and S. Ratzan (2011), Addressing the vaccine confidence gap, *Lancet*, *378*(9790), 526–35, doi:10.1016/s0140-6736(11)60678-8.

Liu, B. (2012), Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Liu, X., and H. Chen (2013), Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums, in *Smart Health*, pp. 134–150, Springer.

Lyles, C. R., A. López, R. Pasick, and U. Sarkar (2013), "5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": An exploratory qualitative analysis of cervical and breast cancer screening dialogue on twitter, *Journal of Cancer Education*, *28*(1), 127–133.

Lynch, M. (2010), Healthy habits or damaging diets: an exploratory study of a food blogging community, *Ecol Food Nutr*, *49*(4), 316–35.

Lyon, A., M. Nunn, G. Grossel, and M. Burgman (2012), Comparison of web-based biosecurity intelligence systems: Biocaster, epispider and healthmap, *Transbound Emerg Dis*, *59*(3), 223–32.

MacLean, D. L., and J. Heer (2013), Identifying medical terms in patient-authored text: a crowdsourcing-based approach, *Journal of the American Medical Informatics Association*, pp. amiajnl–2012.

Madsen, K. M., A. Hviid, M. Vestergaard, D. Schendel, J. Wohlfahrt, P. Thorsen, J. Olsen, and M. Melbye (2002), A population-based study of measles, mumps, and rubella vaccination and autism, *N Engl J Med*, *347*(19), 1477–82, doi:10.1056/NEJMoa021134.

Marcus, M. A., H. A. Westra, J. D. Eastwood, and K. L. Barnes (2012), What are young adults saying about mental health? an analysis of internet blogs, *J Med Internet Res*, *14*(1), e17.

Marshall, V., and N. W. Baylor (2011), Food and drug administration regulation and evaluation of vaccines, *Pediatrics*, *127 Suppl 1*, S23–30, doi:10.1542/peds.2010-1722E.

Mawudeku, A., and M. Blench (2006), Global public health intelligence network (gphin), in *7th Conference of the Association for Machine Translation in the Americas*, pp. 8–12.

McNeil, K., P. M. Brna, and K. E. Gordon (2012), Epilepsy in the twitter era: a need to re-tweet the way we think about seizures, *Epilepsy Behav*, *23*(2), 127–30.

Mei, Q., J. Guo, and D. Radev (2010), Divrank: the interplay of prestige and diversity in information networks, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1009–1018, ACM.

Meystre, S. M., G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle, et al. (2008), Extracting information from textual documents in the electronic health record: a review of recent research, *Yearb Med Inform*, *35*, 128–44.

Mitchell, A., and P. Hitlin (2013), Twitter reaction to events often at odds with overall public opinion, *Pew Research Center*, *4*.

Mogadala, A., and V. Varma (2012), Twitter user behavior understanding with mood transition prediction, in *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, pp. 31–34, ACM.

Mohammad, S. M. (2012), #emotional tweets, in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 246–255, Association for Computational Linguistics.

Moher, D., A. Liberati, J. Tetzlaff, D. G. Altman, and T. P. Group (2009), Preferred reporting items for systematic reviews and meta-analyses: The prisma statement, *PLoS Med*, *6*(7), e1000,097, doi:10.1371/journal.pmed.1000097.

Moreno, M. A., M. Parks, and L. P. Richardson (2007), What are adolescents showing the world about their health risk behaviors on myspace?, *MedGenMed*, *9*(4), 9.

Moreno, M. A., M. R. Parks, F. J. Zimmerman, T. E. Brito, and D. A. Christakis (2009), Display of health risk behaviors on myspace by adolescents: prevalence and associations, *Arch Pediatr Adolesc Med*, *163*(1), 27–34.

Moreno, M. A., L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker (2011), Feeling bad on facebook: depression disclosures by college students on a social networking site, *Depress Anxiety*, *28*(6), 447–55.

Mrozek-Budzyn, D., A. Kieltyka, and R. Majewska (2010), Lack of association between measles-mumps-rubella vaccination and autism in children: a case-control study, *Pediatr Infect Dis J*, *29*(5), 397–400, doi:10.1097/INF.0b013e3181c40a8a.

Murthy, D., A. Gross, and S. Longwell (a), Twitter and e-health: A case study of visualizing cancer networks on twitter, in *Information Society (i-Society), 2011 International Conference on*, pp. 110–113, IEEE.

Murthy, D., A. Gross, and D. Oliveira (b), Understanding cancer-based networks in twitter using social network analysis, in *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pp. 559–566.

Neuman, Y., Y. Cohen, D. Assaf, and G. Kedma (2012), Proactive screening for depression through metaphorical and automatic text analysis, *Artif Intell Med*, *56*(1), 19–25.

Newman, D., E. V. Bonilla, and W. Buntine (2011a), Improving topic coherence with regularized topic models, in *Advances in neural information processing systems*, pp. 496–504.

Newman, M. W., D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris (2011b), It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health, in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 341–350, ACM.

Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (2000), Text classification from labeled and unlabeled documents using em, *Machine Learning*, *39*(2), 103–134, doi:10.1023/a:1007692713085.

Nikfarjam, A., and G. H. Gonzalez (2011), Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments, *AMIA Annu Symp Proc*, *2011*, 1019–1026.

Page, L., S. Brin, R. Motwani, and T. Winograd (1999), The pagerank citation ranking: Bringing order to the web.

Pandey, A., N. Patni, M. Singh, A. Sood, and G. Singh (2010), Youtube as a source of information on the h1n1 influenza pandemic, *Am J Prev Med*, *38*(3), e1–3.

Pang, B., and L. Lee (2008), *Opinion mining and sentiment analysis*, Now Pub.

Parker, S. K., B. Schwartz, J. Todd, and L. K. Pickering (2004), Thimerosal-containing vaccines and autistic spectrum disorder: a critical review of published original data, *Pediatrics*, *114*(3), 793–804, doi:10.1542/peds.2004-0434.

Patrick, J., and M. Li (2010), High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, *Journal of the American Medical Informatics Association*, *17*(5), 524–527.

Paul, M. J., and M. Dredze (2011), You are what you tweet: Analyzing twitter for public health, in *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*.

Paul, M. J., and M. Dredze (2012), A model for mining public health topics from twitter, *HEALTH*, *11*, 166.

Pew (2014), Social networking fact sheet, `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`, accessed: 2015-11-29.

Prier, K., M. Smith, C. Giraud-Carrier, and C. Hanson (2011), *Identifying Health-Related Topics on Twitter Social Computing, Behavioral-Cultural Modeling and Prediction*, *Lecture Notes in Computer Science*, vol. 6589, pp. 18–25, Springer Berlin / Heidelberg.

Prochaska, J. J., C. Pechmann, R. Kim, and J. M. Leonhardt (2012), Twitter=quitter? an analysis of twitter quit smoking social networks, *Tob Control*, *21*(4), 447–9.

Quincey, E., and P. Kostkova (2010), *Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter Electronic Healthcare*, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 27, pp. 21–24, Springer Berlin Heidelberg.

Ressler, P. K., Y. S. Bradshaw, L. Gualtieri, and K. K. Chui (2012), Communicating the experience of chronic pain and illness through blogging, *J Med Internet Res*, *14*(5), e143.

Richardson, A., and D. M. Vallone (2012), Youtube: a promotional vehicle for little cigars and cigarillos?, *Tob Control*.

Richardson, C. G., L. Vettese, S. Sussman, S. P. Small, and P. Selby (2011), An investigation of smoking cessation video content on youtube, *Subst Use Misuse*, *46*(7), 893–7.

Ritterman, J., M. Osborne, and E. Klein (2009), Using prediction markets and Twitter to predict a swine flu pandemic, in *Proceedings of the 1st International Workshop on Mining Social Media*.

Sadilek, A., and H. Kautz (2013), Modeling the impact of lifestyle on health at scale, in *Sixth ACM International Conference on Web Search and Data Mining*.

Sadilek, A., H. Kautz, and V. Silenzio (2012a), Modeling spread of disease from social interactions, in *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.

Sadilek, A., H. Kautz, and V. Silenzio (2012b), Predicting disease transmission from geo-tagged micro-blog data, in *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Saha, S., D. Chant, and J. McGrath (2007), A systematic review of mortality in schizophrenia: is the differential mortality gap worsening over time?, *Arch Gen Psychiatry*, *64*(10), 1123–31.

Salathe, M., et al. (2012), Digital epidemiology, *PLoS computational biology*, *8*(7), e1002,616.

Salath, M., and S. Khandelwal (2011), Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control, *PLoS Comput Biol*, *7*(10), e1002,199, doi:10.1371/journal.pcbi.1002199.

Sarah, A., G. Bell, M. Paul, and M. Pronovost (2012), Malpractice and malcontent: Analyzing medical complaints in twitter, in *AAAI 2012 Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text*.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute (2010), Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications, *Journal of the American Medical Informatics Association*, *17*(5), 507–513.

Scanfeld, D., V. Scanfeld, and E. L. Larson (2010), Dissemination of health information through social networks: twitter and antibiotics, *Am J Infect Control*, *38*(3), 182–8.

Seidenberg, A. B., E. J. Rodgers, V. W. Rees, and G. N. Connolly (2012), Youth access, creation, and content of smokeless tobacco ("dip") videos in social media, *J Adolesc Health*, *50*(4), 334–8.

Seifter, A., A. Schwarzwalder, K. Geis, and J. Aucott (2010), The utility of google trends for epidemiological research: Lyme disease as an example, *Geospatial Health*, *4*(2), 135–137.

Shaw, R. J., and C. M. Johnson (2011), Health information seeking and social media use on the internet among people with diabetes, *Online J Public Health Inform*, *3*(1).

Sievert, C., and K. E. Shirley (2014), Ldavis: A method for visualizing and interpreting topics, in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

Signorini, A., A. M. Segre, and P. M. Polgreen (2011), The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic, *PLoS One*, *6*(5), e19,467.

Simunaniemi, A. M., H. Sandberg, A. Andersson, and M. Nydahl (2011), Laypeople blog about fruit and vegetables for self-expression and dietary influence, *Health Commun*, *26*(7), 621–30.

Skea, Z. C., V. A. Entwistle, I. Watt, and E. Russell (2008), 'avoiding harm to others' considerations in relation to parental measles, mumps and rubella (mmr) vaccination discussions - an analysis of an online chat forum, *Soc Sci Med*, *67*(9), 1382–90, doi:10.1016/j.socscimed.2008.07.006.

Smith, A., J. Yarwood, and D. M. Salisbury (2007), Tracking mothers' attitudes to mmr immunisation 19962006, *Vaccine*, *25*(20), 3996–4002, doi:http://dx.doi.org/10.1016/j.vaccine.2007.02.071.

Smith, C. A. (2011), Consumer language, patient language, and thesauri: a review of the literature, *Journal of the Medical Library Association*, *99*(2), 135–144.

Smith, C. A., and P. J. Wicks (2008), Patientslikeme: Consumer health vocabulary as a folksonomy, *AMIA Annu Symp Proc*, pp. 682–6.

Smrz, P., and L. Otrusina (2011), Finding indicators of epidemiological events by analysing messages from twitter and other social networks, in *Proceedings of the second international workshop on Web science and information exchange in the medical web*, pp. 7–10, ACM.

Sofean, M., and M. Smith (2012), A real-time architecture for detection of diseases using social networks: design, implementation and evaluation, in *Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 309–310, ACM.

Sofean, M., K. Denecke, A. Stewart, and M. Smith (2012), Medical case-driven classification of microblogs: characteristics and annotation, in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 513–522, ACM.

Speers, T., and J. Lewis (2004), Journalists and jabs: media coverage of the mmr vaccine, *Commun Med*, *1*(2), 171–81, doi:10.1515/come.2004.1.2.171.

Steele, R. (2011), Social media, mobile devices and sensors: categorizing new techniques for health communication, in *Sensing Technology (ICST), 2011 Fifth International Conference on*, pp. 187–192, IEEE.

Steele, R., and K. Min (), Health system zeitgeist: How tweets can provide real-time insight into the health system, in *Industrial Informatics (INDIN), 2012 10th IEEE International Conference on*, pp. 846–851.

Stephen, K., and G. P. Cumming (2012), Searching for pelvic floor muscle exercises on youtube: what individuals may find and where this might fit with health service programmes to promote continence, *Menopause Int*, *18*(3), 110–5.

Steyvers, M., and T. Griffiths (2007), Probabilistic topic models, *Handbook of latent semantic analysis*, *427*(7), 424–440.

Sullivan, S. J., A. G. Schneiders, C. W. Cheang, E. Kitto, H. Lee, J. Redhead, S. Ward, O. H. Ahmed, and P. R. McCrory (2012), 'what's happening?' a content analysis of concussion-related traffic on twitter, *Br J Sports Med*, *46*(4), 258–63.

Tausczik, Y., K. Faasse, J. W. Pennebaker, and K. J. Petrie (2012), Public anxiety and information seeking following the h1n1 outbreak: blogs, newspaper articles, and wikipedia visits, *Health Commun*, *27*(2), 179–85.

Tian, Y. (2010), Organ donation on web 2.0: content and audience analysis of organ donation videos on youtube, *Health Commun*, *25*(3), 238–46.

Uno, Y., T. Uchiyama, M. Kurosawa, B. Aleksic, and N. Ozaki (2012), The combined measles, mumps, and rubella vaccines and the total number of vaccines are not associated with development of autism spectrum disorder: the first case-control study in asia, *Vaccine*, *30*(28), 4292–8, doi:10.1016/j.vaccine.2012.01.093.

van der Velden, M., and K. El Emam (2013), "not all my friends need to know": a qualitative study of teenage patients, privacy, and social media, *J Am Med Inform Assoc*, *20*(1), 16–24.

van Uden-Kraan, C., C. Drossaert, E. Taal, E. Seydel, and M. van de Laar (2009), Participation in online patient support groups endorses patients' empowerment, *Patient Education and Counselling*, *74*(1), 61–69.

Villiard, H., and M. A. Moreno (2012), Fitness on facebook: advertisements generated in response to profile content, *Cyberpsychol Behav Soc Netw*, *15*(10), 564–8.

Von Muhlen, M., and L. Ohno-Machado (2012), Reviewing social media use by clinicians, *Journal of the American Medical Informatics Association*, *19*(5), 777–781.

Vydiswaran, V. V., Y. Liu, K. Zheng, D. A. Hanauer, and Q. Mei (2014), User-created groups in health forums: What makes them special?, in *Proc Conf Weblogs and Social Media (ICWSM) Assoc Adv Artif Intell*, pp. 515–24.

Wakefield, A., et al. (1998), Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, *THE LANCET*, *351*, 637–41.

Wang, Y.-C., R. Kraut, and J. M. Levine (2012), To stay or leave? The relationship of emotional and informational support to commitment in online heath support groups, in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*, pp. 833–842.

White, R. W., R. Harpaz, N. H. Shah, W. DuMouchel, and E. Horvitz (2014), Toward enhanced pharmacovigilance using patient-generated data on the internet, *Clinical Pharmacology & Therapeutics*, *96*(2), 239–246.

Xu, J.-M., K.-S. Jun, X. Zhu, and A. Bellmore (2012), Learning from bullying traces in social media, in *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pp. 656–666, Association for Computational Linguistics.

Yang, C. C., H. Yang, L. Jiang, and M. Zhang (2012), Social media mining for drug safety signal detection, in *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pp. 33–40, ACM.

Zeng, Q. T., and T. Tse (2006), Exploring and developing consumer health vocabularies, *J Am Med Inform Assoc*, *13*(1), 24–29, doi:10.1197/jamia.M1761.

Zeng, Q. T., T. Tse, G. Divita, A. Keselman, J. Crowell, A. C. Browne, S. Goryachev, and L. Ngo (2007), Term identification methods for consumer health vocabulary development, *J Med Internet Res*, *9*(1), e4, doi:10.2196/jmir.9.1.e4.

Zheluk, A., J. A. Gillespie, and C. Quinn (2012), Searching for truth: internet search patterns as a method of investigating online responses to a russian illicit drug policy debate, *J Med Internet Res*, *14*(6), e165.

Ziebland, S., and S. Wyke (2012), Health and illness in a connected world: how might sharing experiences on the internet affect people's health?, *Milbank Q*, *90*(2), 219–49.