Chapter 1

Social Media Analytics for Healthcare

Alexander Kotov
Department of Computer Science
Wayne State University
Detroit, MI
kotov@wayne.edu

1.1 Introduction

The emergence of social media resources in the form social networking sites, blogs/microblogs, forums, question answering services, on-line communities and encyclopedias, which are often collectively referred to as Web 2.0, designated a move from passive consumption to active creation of diverse types of content by the Internet users. Unlike newswire articles, social media goes beyond stating facts and describing events and provides a wealth of information about public opinion on virtually any topic, including healthcare. Recent studies [39] [38] report that 61% of American adults seek health information on-line and 37% have accessed or posted health information on-line. In addition to that, 72% of online adults in the United States are using social media. Of adult social media users, 23% follow their friends' personal health experiences or updates, 17% use social media to remember and memoralize people with a specific health condition and 15% obtain health information from social media sites [7].

Web 2.0 services and platforms have been designed to encourage frequent expression of people's thoughts and opinions on a variety of issues as well as random details of their lives. They have made possible open expression of opinions and exchange of ideas. They have also made measurable what was previously unmeasurable and shed additional light on important questions in public health that have been either too expensive or outright impossible to answer, such as distribution of health information in a population, tracking health information trends over time and identifying gaps between health information supply and demand. The fine granularity and pervasiveness of social media data allows to model phenomena that was previously out of reach, including the probability of a given individual to get sick with a disease. Although most individual social media posts and messages contain little informational value, aggregation of millions of such messages can generate important knowledge. For example, knowing that a certain individual has contracted a flu based on his or her messages on social networking sites may not be an interesting fact by itself, but millions of such messages can be used to track influenza rate in a state or a country.

This chapter provides an overview of recent work that demonstrates that social media data can be mined for patterns and knowledge that can be leveraged in descriptive as well as predictive models of population health. It can also improve the overall effectiveness of public health monitoring and analysis and significantly reduce its latency. Previous research work on social media analytics for healthcare has focused on the following three broad areas:

1. Methods for capturing aggregate health trends from social media data, such as outbreaks of infectious diseases, and analyzing the mechanisms underlying the spread of infectious diseases;

2. Methods for fine-grained analysis and processing of social media data, such as methods to detect reports of adverse drug interactions and medical events and to model the health status and well-being of individuals;

3. Studying how social media can be effectively used as a communication medium between patients, between patients and doctors and how to effectively leverage social media in interventions and health education campaigns.

The primary goal of the first line of work, which we focus on in Section 1.2, is to detect and estimate the magnitude of an infectious disease outbreak in a particular geographical region from social media data, search logs of major Web search engines or access logs to medical Web sites. This direction views Web 2.0 users as the first responders to a disease outbreak in an information sense and attempts to capture the signals from those users to enable faster outbreak discovery. Timely detection of infectious disease outbreaks can significantly decrease their negative effect, while modeling ”what-if” scenarios based on analysis of data from both social media and healthcare agencies can decrease public health response time and increase its effectiveness. A common theme for most early approaches proposed along this direction is establishing the degree of correlation between the data extracted from social media and official federal, state and local public health statistics.

The primary goal of the second line of work, which we focus on in Section 1.3), is to extract knowledge from social media that can be utilized to address specific healthcare problems, such as detecting reports of adverse medical events, summarizing the effects of recreational drugs use and predicting when a particular individual will become afflicted with an illness. Large-scale social media mining in combination with the analysis of on-line social networks, demographic analysis and predictive modeling of risk behaviors can improve our understanding of epidemiological mechanisms and allow public health professionals to tailor awareness, design more effective interventions and better predict their outcome.

The third line of work, which we overview in Section 1.4, is focused on studying how social media is used as a source of health information, such as how ordinary people and healthcare professionals use social media to answer their health-related questions or report their experiences in dealing with medical conditions. In particular, we discuss popular on-line communities for both patients and healthcare professionals and outline the findings that were made by analyzing the textual content posted on those communities.

The vast majority of approaches proposed as part of these three directions are based on combining social network analysis, machine learning, statistical modeling and computational linguistics with epidemiology, sociology, economics and public health research. Such an approach is best illustrated by the pyramid model of public health proposed by Sadilek and Kautz [71]. At the base of the pyramid is the entire population. In the middle of the pyramid are the users of on-line social media, whose data is publicly available. At the top of pyramid is a small but strategically selected sample of individuals from the general population (which includes some of the social media users), for whom the detailed health records are available. This sample includes the subjects who respond to on-line medical surveys, actively monitor their health status at home (e.g. by using glucose or blood pressure monitors, HIV rapid tests, etc.) or at a nearby medical lab and are willing to share their personal health information with other people. Traditionally, epidemiological studies are based on the data collected from the top of this pyramid. Although majority of the work discussed in this chapter uses the data from the middle of the pyramid, machine learning and statistical modeling techniques allow the knowledge gained at any level in the pyramid to “trickle down”. For example, by applying machine learning techniques we can bootstrap from the top of the pyramid to make well-grounded predictions about the general population at the bottom of the pyramid. This will infuse epidemiological models with additional structure and parameters learned from detailed timely data, so that fewer factors need to be modeled via simulation. Information can also “trickle up” the pyramid, where the latent behavior of the general population may influence the predictions even for the individuals at the top. In the following sections, we will examine in detail each one of the three general directions outlined above.

1.2 Social Media Analysis for Detection and Tracking of Infectious Disease Outbreaks

Epidemics of infectious diseases, such as influenza and cholera, are a major public health concern that is difficult to anticipate and model [86]. Seasonal influenza epidemics result in about three to five million cases of severe illnesses and about 250,000 to 500,000 deaths worldwide each year. Although influenza reoccurs each season in regular cycles, geographic location, timing and size of each outbreak varies, complicating the efforts to produce reliable and timely estimates of influenza activity using traditional methods for time series analysis. In general, health organizations require accurate and timely disease surveillance techniques in order to respond to the emerging epidemics by better planning for surges in patient visits, therapeutic supplies and public health information dissemination campaigns. Additional early knowledge of an upward trend in disease prevalence can inform patient capacity preparations and increased efforts to distribute the appropriate vaccine or other treatments, whereas knowledge of a downward trend can signal the effectiveness of these efforts.

Public health monitoring has traditionally relied on surveys and aggregating primary data from healthcare providers and pharmacists (e.g. clinical encounters with healthcare professionals, sick-leave and drug prescriptions). Syndromic surveillance, the monitoring of clinical syndromes that have significant impact on public health, is particularly required for episodic and widespread infections, such as seasonal influenza. Many infectious disease surveillance systems, including those employed by Centers for Disease Control and Prevention (CDC) in the United States, Public Health Agency of Canada, Infectious Disease Surveillance Center in Japan, Health Protection Agency in the United Kingdom, Swedish Institute for Infectious Disease Control and the European Influenza Surveillance Scheme continuously collect virological and clinical reports from designated laboratories and physicians, in a process known as sentinel surveillance. For example, CDC operates the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) and publishes the data collected and aggregated from it on-line via FluView [1]. ILINet is one of the most effective disease surveillance systems, which monitors 2,700 sentinel outpatient health providers and issues weekly reports of the proportion of all visits to those providers that are related to influenza-like illness (ILI) symptoms (temperature 100 degrees Fahrenheit or greater, cough and/or sore throat without any apparent cause). Although survey-based surveillance systems are effective tools in discovering disease outbreaks, they typically incur high operational costs and temporal lags in reporting the outbreaks, since an infectious disease case is recorded only after a patient visits a doctor's office and the information about it is sent to the appropriate public health agency. In case of CDC, the typical lag times for influenza reporting are one to two weeks with even longer lags for less common diseases [29]. During the deadly infectious disease outbreaks, such as cholera, this delay can hinder early epidemiological assessment and result in greater number of fatalities. Previous work in epidemiology [34] has also shown that the most effective way to fight an epidemic in urban areas is to quickly confine infected individuals to their homes. Since this strategy is effective only when applied early, it becomes important to be able to detect the outbreaks of infectious diseases in urban areas as quickly as possible. In general, methods for earlier outbreak detection allow more time to deploy interventions that can lower the morbidity and mortality resulting from the outbreak. Besides longer reporting time lag, sentinel-based surveillance systems suffer from population bias, since people who do not actively seek treatment or do not respond to surveys are virtually invisible to them, and tend to over-report population groups that are more vulnerable to diseases. By contrast, social media data are available in near real-time and therefore can provide much earlier estimates of the magnitude and dynamics of an epidemic. Social media platforms, such as Twitter, offer virtually unlimited volumes of publicly available data and population sample sizes that exceed those of paper surveys by several orders of magnitude. Finding the key symptomatic individuals along with other people, who may have already contracted the disease, can also be done more effectively and in a timely manner by leveraging on-line social network data. Furthermore, geographical metadata in the form of the coordinates associated with some of the social media posts can play an important role in monitoring the impact and the geographical spread of an epidemic. In this section, we provide an extensive overview of the recently proposed methods for detection and tracking of infectious disease outbreaks based only on the analysis of the signals from social media.

1.2.1 Outbreak Detection

Experiments with unconventional methods using pre-clinical "health information seeking" for syndromic surveillance have been conducted before the advent of Web 2.0 and social media. For example, several surveillance systems have been introduced in the past to monitor indirect signals of influenza activity, such as the volume of calls to telephone health advisory lines [23] [33], over-the-counter drug sales [57] and school absenteeism rates [55]. The emergence and rapid widespread adoption of on-line services, such as search engines and social media platforms like Twitter and Facebook, presented an opportunity for nearly real-time Internet-based surveillance for disease outbreaks based only on the analysis of the data from these services. This led to the emergence of a new area of research at the intersection of computer science and public health known as "infodemiology" or "information epidemiology" [35]. Infodemiology is an umbrella term for methods that study the determinants and distribution of health information for public health purposes and are aiming at:

- developing methodologies and measures to understand patterns and trends for general public health research;

[1] http://www.cdc.gov/flu/weekly/

- identifying disease outbreaks based on the analysis of these trends;
- studying and quantifying knowledge translation gaps;
- understanding the predictive value of search and content generation behavior for syndromic surveillance and early detection of emerging diseases.

As a result, several lines of recent work have focused on developing new methods to detect outbreaks of infectious diseases using the data from different types of on-line services, such query logs, microblogs and blogs.

1.2.1.1 Using Search Query and Web Site Access Logs

An increasing number of people around the world are using the Internet to seek and disseminate health-related information. People search for health information for a variety of reasons: concerns about themselves, their family or friends. According to National Library of Medicine, an estimated 113 million people in the United States use the Internet to find health-related information with up to 8 million people searching for health-related information on a typical day [60]. About 90 million American adults are believed to search for on-line information about specific diseases or medical problems each year, making Web search a unique source of information about health trends and major events, such as epidemics. Therefore, an interesting research question from both computer science and public health perspective is whether tracking health information seeking behavior of people over time can be used to monitor public health in general and for syndromic surveillance, in particular.

The general idea behind the proposed methods for monitoring public health based on the analysis of query logs of search engines is that the interest of a general public in a certain public health topic can be approximated by the search query activity related to this topic. Therefore, health information seeking behavior can be captured and transformed into indicators of disease activity. Since some search query data also carries geographical information (generally based on the IP address of the computer, from which a particular query was issued), it may also be possible to detect simple geo-spatial patterns. Eysenbach [36] explored whether an automated analysis of trends in Internet searches could be useful for predicting the outbreaks of infectious diseases, such as influenza. He created a Google advertisement campaign, in which the advertisements were triggered by the influenza-related search terms and experimented with different multivariate models to predict the number of ILI cases based on the advertisement campaign statistics. He found out that the number of clicks on on-line advertisements has the highest correlation with traditional surveillance measures. He also observed that the weekly number of flu-related advertisement clicks has even higher correlation with ILI reports from sentinel physicians for the following week, suggesting systematic mining of search engine logs could be a valuable addition to traditional surveillance methods for those conditions, when the patients consult the Internet before visiting a physician.

A joint study by CDC and Yahoo! suggested that Internet searches for specific cancers correlate with their estimated incidence, mortality and the volume of related news coverage [22]. They concluded that media coverage appears to play a powerful role in prompting on-line searches for cancer information. Ginsberg et al. [45] processed search logs containing hundreds of billions of search queries submitted to Google search engine between 2003 and 2008 and estimated a simple linear regression model to predict the log-odds of the percentage of ILI-related physician visits in a geographical region based only on the log-odds of ILI-related queries for the same geographical region. Estimates produced by this model resulted in the Pearson correlation coefficient of 0.97 with the CDC-reported ILI statistics. This work resulted in creation of the Google Flu Trends service [2], which estimates the current flu activity around the world based on the volume of search queries to the Google search engine. Google Flu Trends is used to successfully track influenza rates on a

[2] `http://www.google.org/flutrends`

daily basis, up to 7 to 10 days faster than CDC's FluView [13]. Nevertheless, similar to sentinel surveillance, systems based on the analysis of search engine query logs also suffer from population bias, as they are restricted to the sample of individuals, who search the Internet for certain types of content when sick.

Pelat et al. [66] compared the search trends related to 3 infectious diseases with clinical surveillance data from the French Sentinel Network and reported the correlation coefficients of 0.82 for influenza-like illnesses, 0.9 for gastroenteritis and 0.78 for chickenpox. They concluded that, for each of these three infectious diseases, one well-chosen query is sufficient to provide time series of searches that is highly correlated with the actual incidence of those diseases reported through the sentinel surveillance system. The highest correlation between the best queries for influenza and gastroenteritis was achieved without any time lag, while the time series of searches for chickenpox was lagging one week behind the incidence time series.

Seifter et al. [80] explored the utility of using Google Trends to study seasonal and geographic patterns for Lyme disease. They found that the search traffic for the query "Lyme disease" reflected the increased likelihood of exposure to this disease during spring and summer months and that the cities and states with the highest search traffic for this query considerably overlapped with those, where Lyme disease was known to be endemic. Following similar idea, Hulth et al. [50] explored the feasibility of using the queries submitted to a Swedish medical Web site [3] for the task of influenza outbreak detection and observed that certain influenza related queries followed the same pattern as the data obtained by the two standard surveillance systems (based on the number of laboratory verified influenza cases and the proportion ILI-related patients visits to sentinel general practitioners). In particular, they used partial least squares regression to identify the most indicative queries for influenza and achieved the correlation coefficients of 0.9 with the sentinel data and 0.92 with the laboratory data. Johnson et al. [51] used the access logs from Healthlink medical Web site [4] to measure the correlation between the number of accesses to selected Influenza-related pages on this Web site and influenza surveillance data from CDC and reported such correlation to be moderately strong.

Although search data is confounded by media reports and "epidemics of fear", even crude (unadjusted) surges in increased search activity on a health topic not triggered by a real pandemic are still important measures for government health agencies and policy makers, as they may, even in the absence of a true epidemic, warrant a public health response into what may be causing an increased information demand. However, the major limitation of search query logs is that they do not provide any additional contextual information, therefore the questions like why the search was initiated in the first place are difficult to answer.

1.2.1.2 Using Twitter and Blogs

The emergence and rapid increase in popularity of Twitter [5] opened up a new research direction in Internet-based disease surveillance. Twitter is a social networking and microblogging platform that enables users to create the posts limited to 140 characters and share them either with the general public or only with a specific group of people designated as "followers". Although the Twitter stream consists largely of useless chatter, self-promotion messages and user-to-user conversations that are only of interest to the parties involved, due to the sheer volume of tweets, it contains enough useful information for any task. For example, Twitter data has been used to measure political opinions [59], national sentiment [4], public anxiety related to stock prices [44] and to monitor the impact of earthquakes [74]. The advantages of Twitter-based approaches for disease outbreak detection over the ones that are based on search query and access logs are two-fold. First, although Twitter messages are fairly short, they are still more descriptive and provide more contextual information

[3] http://www.vardguiden.se
[4] http://www.healthlink.com
[5] http://www.twitter.com

than search engine queries. Second, Twitter profiles often contain rich meta-data associated with the users (e.g their geographical location, gender, age and social network), enabling more sophisticated and detailed analysis. Twitter also has an advantage over other social media services in that it offers larger volume of mostly publicly available messages. In particular, as of January 2014, Twitter is estimated to have over 600 million active registered users worldwide, who create 58 million microblog posts every day. Frequent updates and public data availability open up opportunities for near real-time, demographically and geographically focused disease surveillance.

The work of Ritterman et al. [68] was one of the first to use Twitter for infectious disease surveillance. In particular, they used the dataset consisting of 48 million tweets collected over a period of two months, which covers the timespan between the first time when the news about H1N1 (or Swine Flu) virus first broke out and until the H1N1 pandemic was declared by the World Health Organization on May 11th, 2009. They used the data from Hubdub [6], an on-line prediction market, to model the public belief that H1N1 will become a pandemic using support vector machine (SVM) regression. Their analysis resulted in two major conclusions. The first conclusion is that simple bigram features extracted from the content of Twitter messages within historical contexts of different granularity (1 day, 3 days, 1 week, entire history) can accurately predict health-related beliefs and expectations of the general public. The second conclusion is that combining the features based on the content of Twitter messages and with the ones derived from the Hubdub data results in more accurate prediction model that the one, which relies on the prediction markets data alone. Quincey and Kostkova [31] have demonstrated the potential of Twitter outbreak detection by collecting and characterizing over 135,000 posts pertaining to H1N1 over a period of one week. Culotta [29] identified influenza-related Twitter posts by applying simple and multiple logistic regression-based document classifiers using the occurrence of predefined keywords such as “flu”, “cough”, “sore throat” and “headache” as features to a dataset of over 500,000 posts spanning 10 weeks. In multiple regression model, each keyword had a different weight, whereas in simple regression model all keywords had the same weights. He then calculated the Pearson correlation coefficient between the log-odds of a fraction of influenza-related messages in the overall daily volume of Twitter posts and the log-odds of a fraction of all outpatient visits with ILI-related symptoms reported by CDC. Although multiple regression outperformed simple regression, he found that multiple regression began to overfit when too many keywords were used. The best model in his study achieved the correlation coefficient of 0.78 with CDC statistics. Culotta [30] applied similar methodology to estimate alcohol sales from the volume of tweets related to drinking and found that the most accurate model is the one, which relies only the keyword “drunk”. In particular, this model achieved the correlation coefficient of 0.932 with the U.S. Census Bureau data, which suggests that Twitter can also be used by public health researchers as a useful source for monitoring alcohol consumption trends. Signorini et al. [83] filtered 951,697 tweets containing flu-related keywords (“h1n1”, “swine”, “flu”, “influenza”) from 334,840,972 tweets posted during H1N1 pandemic between April 29th and June 1st of 2009 and observed that that the percentage of such tweets rapidly declined over time as more and more H1N1 cases have been reported. They also constructed the time series of daily counts of tweets related to particular sub-topics of H1N1 pandemic, such as countermeasures (hand hygiene, protective face masks), treatments (antiviral medications used to treat influenza), travel-related issues, vaccination and vaccination side-effects (Guillain-Barré syndrome) and food consumption related concerns, and found that there was no evidence of sustained interest in those topics by Twitter users during the pandemic. They also applied SVM regression based on bag-of-words feature vectors to estimate the national ILI levels as well as ILI levels for specific geographical regions (states) by geo-locating the tweets based on user profiles and reported very high accuracy for both types of estimates.

Lampos and Cristianini [53] proposed a method to calculate the flu-score for a given tweet (or a corpus of tweets) using the tweet’s *n*-grams as features (or “textual markers” in their terminology).

[6] `http://www.hubdub.com`

A set of 1,560 stemmed candidate features was first extracted from both the Wikipedia article about influenza and an entry from an on-line medical directory describing flu symptoms along with the comments from the patients who experienced flu. The most important features were then selected by the least angle regression model (a variant of LASSO) using the daily flu scores reported by the U.K.'s Health Protection Agency as a dependent variable. The estimated regression model was used to determine the projected flu rates, which achieved the correlation coefficient of 0.94 with the actual rates. Additionally, they performed geo-location of tweets and cross-validation of regression models learned for one geographical region on all other regions and reported the total average correlation of 0.89.

Most of the early methods for infectious disease surveillance based on the content of Twitter posts relied on relatively simple methods (e.g. n-gram based models for classifying a tweet as flu-related or not). Although these methods were able to relatively accurately classify the tweets as being related or unrelated to influenza with promising surveillance results, they have ignored many subtle differences between the flu-related tweets. For example, many flu-related tweets express either beliefs related to the flu infection and preventative flu measures (e.g. flu shots) or concerned awareness of increased infections, including the fear of contracting the flu or even a wide-spread panic associated with a pandemic, as opposed to the actual infection-related tweets. Unlike search engine queries, Twitter posts provide more context, which can be leveraged by natural language processing tools to isolate more informative "self-diagnostic" posts from general discussions and opinions, caused by an increased attention towards the subject during the flu season and outright panic during the pandemic. In order to improve the accuracy of Twitter-based surveillance, Lamb et al. [52] proposed two methods for fine-grained classification of tweets based on a large number of lexical, syntactic and stylometric features. One method allows to differentiate the flu awareness tweets from the flu infection-related ones, while the other method allows to distinguish the tweets that correspond to self-reported cases of flu by the people who are in fact infected from the tweets created by healthy people that refer to other individuals sick with flu. The tweets that were identified as flu-related based on their approach achieved higher correlation with the CDC ILI data than the tweets identified as flu-related based on using only lexical features. Achrekar et al. [1] approached the problem of improving the quality of epidemiological signal from Twitter data from a different perspective. They observed that re-tweets and posts from the same users may distort the true number of self-reported cases of influenza infection and reported that excluding such messages improves the correlation coefficient and lowers the root mean-squared error of a linear regression between the number of unique Twitter users self-reporting flu infection and the CDC statistics. They also proposed an auto-regression model combining the Twitter data for the current week with the CDC data from two weeks back (simulating a typical 2 week delay in CDC data reporting) to predict the percentage of ILI-related visits for the current week and observed that the addition of Twitter data improves the accuracy of prediction compared to using past CDC data alone.

In a recent work, Li and Cardie [56] focused on the early detection of flu pandemic and introduced a Bayesian approach based on spatio-temporal Markov Network (which they call Flu Markov Network), which takes into account both the spatial information and the daily fluctuations in the number of posted tweets, for early stage unsupervised detection of flu. Spatial proximity is an important factor in early-stage flu detection, since flu breakouts in many of the neighbors of a non-pandemic geographic location can be indicative of an imminent breakout in this location. Daily fluctuations in the number of tweets, depending on whether a certain day falls on a weekday, weekend or holiday, is a known phenomenon in social media analysis, which needs to be accounted for accurate interpretation of signals from Twitter. Spatial and temporal information is incorporated into a four-state Markov chain, in which the states correspond to non-epidemic, rising epidemic, stationary epidemic and declining epidemic phases and model the progression of a typical pandemic. In the non-epidemic and stationary phases, the number of flu-related tweets is modeled as a Gaussian process taking in account the daily tweet fluctuations, whereas in epidemic and declining epidemic phases the number of tweets is modeled as an autoregressive process incorporating the daily effect.

In contrast to the standard Hidden Markov Model, in Flu Markov Network, the state of the Markov chain for a given location at a given time point is not only dependent on the state of the Markov chain for the same location, but also on the states of the Markov chains of the geographical neighbors of that location at a previous timestamp. For example, the number of ILI-related tweets in a state (country) is influenced by the number of ILI-related tweets in the neighboring states (countries). After filtering out flu-related tweets using SVM with polynomial kernel based on unigram and collocational features, removing re-tweets and tweets of the same user, they reported correlation coefficients with CDC data exceeding 0.98 for some geographical regions. Aramaki et al. [2] experimented with a large number of standard classifiers and feature generation techniques to determine the most accurate method for identifying the tweets about self-reported cases of flu infection and compared it with simple frequency-based and search log analysis-based method. They separated an influenza season into periods of excessive and non-excessive news coverage and found that the performance of Twitter-based surveillance method is sensitive to the intensity of news coverage. In particular, during the periods of non-excessive news coverage, the Twitter-based method slightly outperformed Google Flu Trends (achieving the correlation coefficient of 0.89 versus 0.847). However, Twitter-based surveillance method exhibited dramatic reduction in performance during the periods of excessive news coverage, indicating its vulnerability to "news wire bias". This observation is supported by social amplification of risk framework, which postulates that psychological, social, cultural and institutional factors interact with emergency events and intensify or attenuate risk perceptions. They also found that the Twitter-based method outperformed Google Flu Trends before the peak of the influenza season, but not afterwards, suggesting that the Twitter-based surveillance methods are better suited for early stage influenza detection.

While most research in this direction have correlated social media signals with influenza prevalence metrics in a retrospective way, Broniatowski et al. [8] demonstrated the potential for influenza surveillance with a system built and deployed before the influenza season even started. They found that the accuracy of most social media surveillance systems declines with media attention. The reason is that media attention increases Twitter "chatter" - tweets that are about flu, but do not pertain to the actual infection. They used a staged binary classifier, which first identified whether a tweet was relevant to health, then if it was relevant to influenza and, finally, if it was a report of an actual infection. They also correlated the weekly counts of tweets passing through all the filters with CDC ILI data from 2012-2013 influenza season and reported the correlation coefficient of 0.93. In contrast, the correlation coefficient of the weekly number of tweets containing influenza keywords provided by the U.S. Department of Health and Human Services achieved the correlation coefficient with CDC data of only 0.75. They also applied their method at the level of municipality and reported the correlation coefficient of 0.88 between the number of weekly tweets that pass through all the filters and are geo-located to New York City and the number of weekly ILI-related emergency department visits reported by the New York City Department of Health and Mental Hygiene. Keyword-based selection of tweets resulted in the drop of the correlation correlation coefficient to 0.72. In addition, they analyzed the time series of national CDC ILI rates and counts of tweets related to the flu infection and tweets containing flu-related keywords using Box-Jenkins procedure and found statistically significant effects for a lag of one week, while the lags of two weeks or more were insignificant.

Chew and Eysenbach [15] developed Infovigil, an open-source infoveillance system [7], which continuously gathers flu-related tweets, automatically classifies them into pre-defined content categories and determines temporal trends for each category. Classification is performed according to a coding scheme consisting of three dimensions: the content of a tweet (whether a tweet contains a link to an informational resource, expresses opinion, contains a joke or is about personal experience, etc.); how the content was expressed (humor, relief, downplayed risk, concern, frustration, etc.); type of a link, if a tweet contains any (news, government Web sites, online stores, blogs, etc.) using a simple method based on matching the tweets with a set of pre-defined keywords, emoticons or example

[7] `http://www.infovigil.com`

phrases for each content category. Despite its simplicity, this method was able to achieve significant overlap with the manually created golden standard for most of the content categories. Furthermore, automatically classified tweets demonstrated significant linear trend over time across different content categories, which was generally in the same direction as the manually classified tweets. They also found that the correlation coefficient between the number of tweets in each content category and the number of ILI cases reported by CDC varies significantly depending on the content category (from 0.77 for personal experiences to 0.39 for concerns). Further comparison of the trends across different categories revealed that personal accounts of H1N1 increased over time, while the number of humorous comments decreased, possibly due to increasing perceived seriousness of the situation and the declining popularity of the subject. Analysis of trends across different content categories indicated that the perceived severity, news coverage, viral dissemination of information and Twitter campaigns have considerable effect on the tweet volume and posting behavior over time. Another interesting observation reported in this work is that, contrary to popular belief that misinformation is rampant in social media, only 4.5% of all tweets were manually classified as possible misinformation or speculation, while 90.2% of the tweets provided references to the sources of information they contained, allowing others to confirm its trustworthiness. Overall, this study demonstrated the potential of using Twitter to study public attitudes, perceptions and behaviors during pandemics.

Chunara et al. [19] estimated the correlation between the volume of Twitter posts, news media reports on HealthMap and the data reported by the Haitian Ministry of Public Health during the first 100 days of the 2010 Haitian cholera outbreak. They determined that the volume of information from these informal sources significantly correlated with the official reports during the initial phase of the outbreak (with 1 day lag, the correlation coefficients of HealthMap and Twitter data were 0.76 and 0.86, respectively). They also provided experimental results indicating that social media can be used to accurately estimate the reproductive number of an epidemic, which is used to determine the proportion of the population that needs to be immunized to contain an epidemic or the proportion that will be infected, when the disease reaches its endemic equilibrium.

Corley et al. [25] performed several types of analyses on a collection of 44 million blog posts to study the feasibility of using them for disease surveillance. First, they compared the trends for different type of blog posts with respect to the number of posts per day and observed periodic pattern for both the general and influenza-related blog posts, when bloggers create more posts on a weekday than during the weekend, which was supported by fitting an autocorrelation function with statistically significant weekly time lag. Second, they reported the correlation coefficient of 0.767 between the number of flu related blog posts and the CDC ILI statistics. They also proposed a method for identification of blogger communities by leveraging the links between the blogs based on the closeness, betweenness centrality and PageRank. Closeness is the average of the shortest paths (geodesic distances) between a blog and all other blogs reachable from it via links. Betweenness centrality measures interpersonal influence. More specifically, a blog is central if it lies on a large number of shortest paths between between other blogs. PageRank measures the importance of a blog assuming that the links pointing to it from more central blogs contribute to its ranking more than the links pointing from less central nodes. The general idea of this approach is to identify influential blogs, which can quickly disseminate and broker response strategies and interventions in their respective communities. Readers of these influential blogs can trigger an information cascade, spreading the response to vaccinate, quarantine, and close public places. The blogs with high betweenness could broker information between the communities, synchronizing knowledge, while the blogs with greater closeness and PageRank can quickly disseminate outbreak response strategies.

Although Internet-based surveillance approaches may overcome some limitations of the traditional sentinel-based systems, integration of new and traditional approaches offers the greatest promise for future surveillance of influenza and other infectious diseases. Furthermore, social media, which inherently combines three different types of data (textual, geographical and network) opens up unique opportunities to study the interplay between human mobility, social structure and disease transmission. Although the Internet-based data streams as well as new efforts in syndromic surveillance from repurposed clinical data can fill in some of the critical gaps in traditional approaches to early disease outbreak detection (e.g. first cases or early reports of community-level transmission), they cannot completely describe the epidemiology and global impact of an emerging threat. Traditional surveillance is still necessary to estimate morbidity, mortality and shifts in the incidence of disease according to the demographic factors and changes in case fatality rates. Overall, syndromic surveillance using social media is the leading edge of what will almost certainly evolve into real-time surveillance of data from electronic medical records (EMRs).

1.2.2 Analyzing and Tracking Outbreaks

Besides near real-time surveillance through detection of self-reported cases of infectious diseases, social media analysis can also be used to analyze and track the spread of a pandemic. The lack of timely data and limited understanding of the emergence of global epidemics from day-to-day interpersonal interactions makes monitoring and forecasting global spread of infectious diseases very difficult. Previous research in computational epidemiology has mostly concentrated on coarse-grained statistical analysis of populations, often using synthetic data. Although social media-based surveillance methods can effectively perform passive monitoring and produce coarse, aggregate statistics, such as the expected number of people afflicted by flu in a city or a state, their prediction capabilities are severely limited by the low resolution of the aggregate approach. Therefore, another line of work focused on developing new techniques to provide detailed explanation of the mechanisms underlying infectious disease transmission and, given a pandemic, to predict how rapidly and where it will spread.

The bottom-up approaches proposed in [6] [73] [72] consist of two stages and take into account fine-grained interactions between individuals. In the first stage, a classifier is applied to detect sick individuals based on the content of their tweets. In the second stage, physical interactions between sick and healthy people are estimated via their on-line activities and the large-scale impact of these interactions on public health is predicted.

In particular, Brennan et al. [6] proposed a method to accurately predict the prevalence of an infectious disease in a geographical region (e.g. a city) by modeling fine-grained behavior and interactions of the residents in that region with an outside world. In the first stage of their method, individuals are classified as either healthy or symptomatic based on the content of their tweets using SVM. In the second stage, classification results for individual users are aggregated into two probabilistic variables capturing the flux of healthy and sick travelers as well as their physical interactions within predefined geographical locations based on using the GPS coordinates of their tweets and publicly available airline travel statistics to track geographical movements of people. They estimated that the first variable, which corresponds to the expectation of the number of sick users on a given day in a given geographical region has the correlation coefficients of 0.8 with the CDC statistics and 0.87 with Google Flu Trends. Additionally, they found that using only travel statistics in regression model explains 56% of the variance in Google Flu Trends, while adding the expected number of sick travelers to the model explains 73% of the variance. Including the second variable, which models the number of physical interactions as a function of people traveling to the same airport at the same time explains an additional 5% of the variance.

Sadilek et al. [72] focused on fine-grained analysis of the spread of infectious diseases and studied how it is influenced by geographical co-location, social ties and interpersonal interactions. In their method, two individuals are considered to be co-located, if they visited the same 100 by 100 meter area within the same time window. To identify the tweets indicating that their author was infected by flu at the time of posting, they proposed a cascading process of training an SVM classifier working only with bag-of-words features (unigrams, bigrams and trigrams) that is optimized to overcome an imbalance between positive and negative samples and maximize the area under ROC curve (i.e. to consistently have both high precision and recall). After identifying the tweets likely posted by the infected people, they used the GPS coordinates of these tweets and the Twitter friendships of their authors (in their work Twitter friendship is defined as two users, who follow each other) to quantify the effect of geographical co-locations and social ties on disease transmission. In both cases, they observed strong exponential dependencies: in case of co-locations, between probable physical encounters with sick individuals and ensuing sickness and in case of social ties, between the number of sick friends and the probability of getting sick. For example, they established that having 40 encounters with sick individuals within 1 hour or having 10 sick friends on a given day makes one ill with 20% probability on the next day. At the same time, the number of friends in any health state (i.e. the size of a person's friends list) has no impact on that person's health status.

In [73] Sadilek et al. further developed their previous work [72] and proposed a model, which in addition to predicting whether an individual will fall ill can predict when exactly that will happen. Their method simultaneously captures the effect of collocations as well as their duration on disease transmission and the delay between contagion and the onset of symptoms. After applying SVM to detect individuals afflicted by flu based on the content of their posts, they used a dynamic conditional random field (CRF) model to predict an individual's health status in the future using the 7-day prior history of co-location events, the number of unique sick individuals encountered and the number of sick Twitter friends of this individual as features. They observed that the performance of CRF is significantly enhanced by including the features that are not only based on the health status of Twitter friends, but also on the estimated encounters with already sick, symptomatic individuals, including non-friends. Moreover, when using social ties and co-locations individually, CRF performs inconsistently when making predictions into the future. By contrast, when considering friendships and co-locations jointly, along with using the Viterbi algorithm to infer the most likely sequence of a person's health states over time, performance of the CRF improves and stabilizes, achieving up to 0.94 precision and 0.18 recall. The authors explained the low recall by the fact that about 80% of infections occur without any evidence in social media. They concluded that although many complex events and interactions take place "behind the scenes" and are not directly recorded in social media, they can still exhibit themselves in the activity of a sample of people we can observe. For example, although Twitter friendships themselves do not cause or even facilitate the spread of an infection, they can be proxies and indicators of a complex set of phenomena that may not be directly accessible. For example, friends often often eat out together, meet in classes, share items and travel together. While most of these events are never explicitly mentioned on-line, they are crucial from the disease transmission perspective.

These results can have direct and immediate implications for public health. For example, a person predicted to be at high risk of contracting flu could be specifically encouraged to get a flu vaccination. Additionally, recommendations can be made regarding the places that pose a high risk of getting infected. Finally, the proposed models are not limited only to healthcare domain. Similar approaches can be used to model and predict the transmission of political ideas, purchasing preferences and many other complex behavioral phenomena.

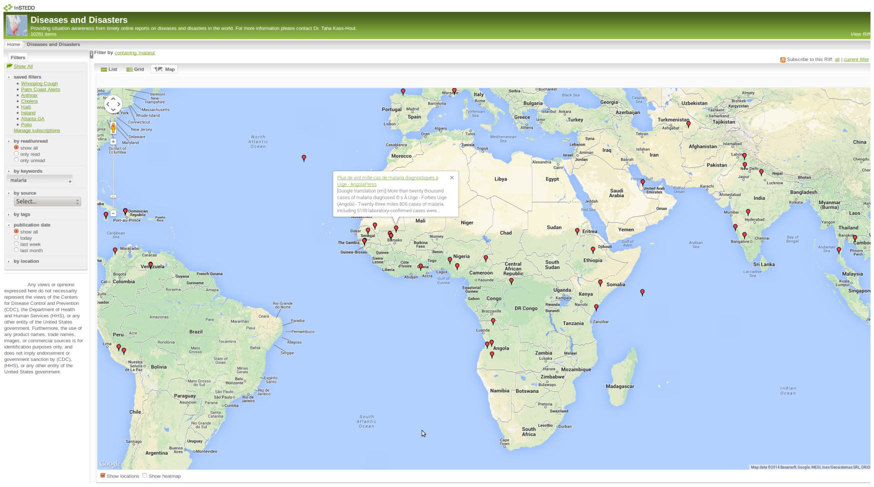
1.2.3 Syndromic Surveillance Systems based on Social Media

Many of the techniques that we overviewed in this section have been implemented in existing on-line syndromic surveillance systems. InSTEDD's Riff [8] is an open source on-line platform for detection, prediction and response to health-related events (such as disease outbreaks) and humanitarian disasters. Riff synthesizes information about public health-related events from a variety of sources (e.g. news, social media, blogs) and visualizes them on a map to assist public health authorities with investigation and response (Figure 1.1).

HealthMap [9] [40] [11] is a system that monitors global media sources such as news wires and web sites to provide a comprehensive view of ongoing disease activity around the world (Figure 1.2).

[8] http://instedd.org/technologies/riff

[9] http://www.healthmap.org



Plus de vint mille cas de malaria diagnostiqués à Uige - AngolaPress
[Google translation (en)] More than twenty thousand cases of malaria diagnosed © s Ã Uige - Forbes Uige (Angola) - Twenty-three miles 806 cases of malaria, including 5159 laboratory-confirmed cases were...

Any views or opinions expressed here do not necessarily represent the views of the Centers for Disease Control and Prevention (CDC), the Department of Health and Human Services (HHS), or any other entity of the United States government. Furthermore, the use of any product names, trade names, images, or commercial sources is for identification purposes only, and does not imply endorsement or government sanction by (CDC), (HHS), or any other entity of the United States government.

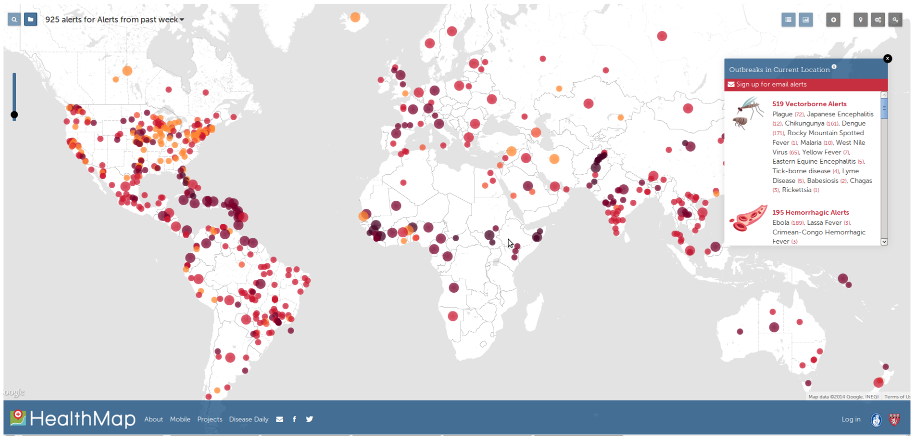
FIGURE 1.1: **User interface of InSTEDD's Riff**

It combines automated, around-the-clock data collection and processing with expert review and analysis. Visitors to the site could filter reports according to the suspected or confirmed cases of deaths from a disease and select a time interval to show its spread. All reports are entered into the HealthMap system along with their geographic location, allowing for easy tracking of both regional and global spread of infectious diseases. During 2009 H1N1 pandemic, HealthMap created an interactive map [10] to provide information about disease outbreaks around the world using information from both informal sources (e.g. news media, mailing lists and contributions from individual users) and formal announcements (primarily from the World Health Organization, the Centers for Disease Control and Prevention and the Public Health Agency of Canada). Brownstein et al. [10] analyzed the geographical pattern for the spread of H1N1 and observed that the countries that are international travel hubs (e.g. France and the United Kingdom) reported flu infections earlier than the countries with less international traffic (e.g. Eastern European nations). They also found that the countries with a high Gross Domestic Product per capita tended to have shorter time lags between the issue dates of reports of suspected and confirmed cases of H1N1 influenza infection. Systems like HealthMap allow anyone with a mobile phone to get involved in responding to a epidemic or humanitarian crisis by contributing relevant information. As an example, during the 2010 Haitian earthquake and cholera outbreak, HealthMap allowed the individuals affected by this crisis to post information about their lost relatives and track the disease activity in their communities.

FluNearYou [10] is an on-line system that integrates different types of data (weekly surveys completed by volunteers, CDC Flu Activity data and Google Flu Trends ILI data) to visualize the current and retrospective flu activity in the United States and Canada (Figure 1.3). It is a joint project between HealthMap, the American Public Health Association, Skoll Global Threats Fund and Boston Children's Hospital.

Crowdbreaks [11] is a surveillance system that automatically collects the disease-related tweets, determines their location and visualizes them on a map (Figure 1.4). It employs machine learning algorithm to assess whether a given tweet contains a reported case of a disease. Crowdbreaks uses crowdsourcing to generate the labeled training data for this algorithm by asking the site visitors to answer simple questions about randomly selected tweets. This systems is based on the idea that social media data are not only provided by the crowd, but can also be assessed and curated by the crowd for their relevance to the issue at hand.

[10] http://flunearyou.org
[11] http://www.crowdbreaks.com


925 alerts for Alerts from past week
HealthMap
About
Mobile
Projects
Disease Daily
Log in
Map data ©2014 Google, INEGI
Terms of Use

Outbreaks in Current Location

Sign up for email alerts

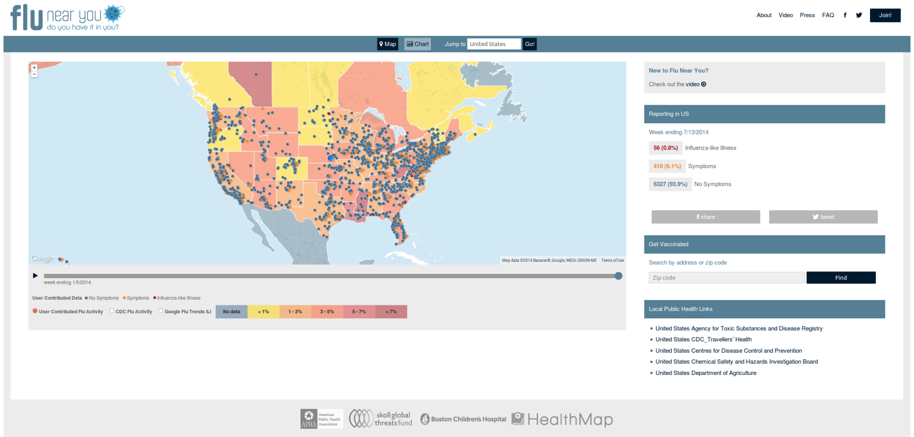
519 Vectorborne Alerts

Plague (72), Japanese Encephalitis (12), Chikungunya (161), Dengue (171), Rocky Mountain Spotted Fever (1), Malaria (10), West Nile Virus (65), Yellow Fever (7), Eastern Equine Encephalitis (5), Tick-borne disease (4), Lyme Disease (5), Babesiosis (2), Chagas (3), Rickettsia (1)

195 Hemorrhagic Alerts

Ebola (189), Lassa Fever (3), Crimean-Congo Hemorrhagic Fever (3)

FIGURE 1.2: User interface of Healthmap



New to Flu Near You?

Check out the video

Reporting in US

Week ending 7/13/2014

56 (0.8%) Influenza-like Illness

410 (6.1%) Symptoms

6327 (93.9%) No Symptoms

Get Vaccinated

Search by address or zip code

Zip code

Local Public Health Links

- United States Agency for Toxic Substances and Disease Registry
- United States CDC_Travellers' Health
- United States Centres for Disease Control and Prevention
- United States Chemical Safety and Hazards Investigation Board
- United States Department of Agriculture

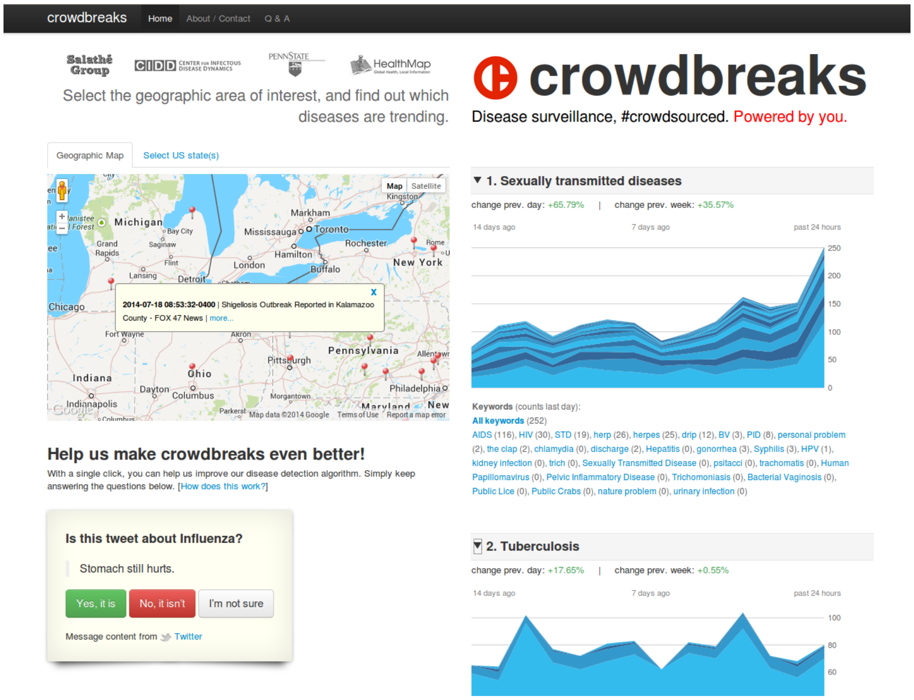
FIGURE 1.3: User interface of FluNearYou

1.3 Social Media Analysis for Public Health Research

While the majority of recent work on social media analysis for healthcare has focused on identifying posts related to particular diseases and correlating their volume with the data reported by the government healthcare agencies, social media analytics can potentially have a far greater impact on healthcare than just disease monitoring. Social media posts are not just isolated textual snippets – they are created at specific times and locations by users from a wide variety of socioeconomic groups, often with known social networks. In this section, we overview the proposed approaches addressing different public health research problems based on the analysis of the content generated by social media users and the structure of their on-line social networks.

1.3.1 Topic Models for Analyzing Health-related Content

Methods capable of aggregating healthcare-related content created by millions of social media users can provide extensive near real-time information about population health and different popu-



Select the geographic area of interest, and find out which diseases are trending.

2014-07-18 08:53:32-0400 | Shigellosis Outbreak Reported in Kalamazoo County - FOX 47 News | more...

Help us make crowdbreaks even better!

With a single click, you can help us improve our disease detection algorithm. Simply keep answering the questions below. [How does this work?]

Is this tweet about Influenza?

Stomach still hurts.

Message content from Twitter

crowdbreaks

Disease surveillance, #crowdsourced. Powered by you.

▼ **1. Sexually transmitted diseases**

change prev. day: +65.79% | change prev. week: +35.57%

Keywords (counts last day):
All keywords (252)
AIDS (116), HIV (30), STD (19), herp (26), herpes (25), drip (12), BV (3), PID (8), personal problem (2), the clap (2), chlamydia (0), discharge (2), Hepatitis (0), gonorrhea (3), Syphilis (3), HPV (1), kidney infection (0), trich (0), Sexually Transmitted Disease (0), psitacci (0), trachomatis (0), Human Papillomavirus (0), Pelvic Inflammatory Disease (0), Trichomoniasis (0), Bacterial Vaginosis (0), Public Lice (0), Public Crabs (0), nature problem (0), urinary infection (0)

▼ **2. Tuberculosis**

change prev. day: +17.65% | change prev. week: +0.55%

FIGURE 1.4: **User interface of Crowdbreaks**

lation characteristics, which is invaluable to public health researchers. Topic models, such as Latent Dirichlet Allocation (LDA) [5], are probabilistic latent variable generative models, which associate hidden variables controlling topical assignments with terms in document collections. They were designed to summarize information in large textual corpora by revealing their latent thematic structure in the form of clusters of semantically related words. In the generative process of topic models, topics are represented as multinomial distributions over the vocabulary of a given collection, such that more probability mass is allocated to the words that frequently co-occur within the same documents, and documents are represented as multinomial distributions over topics. Being an effective mechanism for textual data exploration, new and existing topic models have been extensively used to facilitate the analysis of social media data for healthcare research.

Prier et al. [67] applied LDA to a large corpus of health-related tweets and identified several prevalent topics: physical activity, obesity, substance abuse and healthcare. They observed that the topics related to obesity and weight loss correspond to advertisements, while healthcare topics mostly correspond to political discourse. Besides identifying general health-related topics in the Twitter stream, they also applied basic keyword filtering to create a corpus consisting only of the tweets related to tobacco use and determined the fine-grained topics in this corpus. By examining these topics they found out that besides tobacco promotions, smoking cigarettes or cigars and substance abuse (such as smoking marijuana and crack cocaine), Twitter users also typically discuss strategies to quit smoking and recover from smoking addiction.

Paul and Dredze [61] also applied standard LDA to a corpus of health-related tweets (filtered out from the general Twitter stream using SVM classifier based on bag-of-words features) and reported that, although LDA was able to generate some disease-related topics, most of them did not clearly indicate specific ailments. For example, while many topics discovered by LDA contained surgery terms, it was not clear whether these surgeries were associated with a certain illness, physical injury or cancer. To overcome this problem, they proposed the Ailment Topic Aspect Model (ATAM), which assumes that each health-related tweet corresponds to a latent ailment (e.g. flu, allergy or cancer). Given a collection of such tweets, ATAM identifies a background topic, general health-related topics as well as general, symptom or treatment-related aspects (sub-topics) for each latent ailment. Similar to standard LDA, topics and ailment aspects correspond to multinomial distributions over words. ATAM includes different topic types to account for the fact that even in the tweets about health topics, users often provide additional context, which may not fit into the symptom-treatment-ailment topic structure (e.g. in a tweet "sick today so playing video games" general topics account for "playing video games"). In order to determine which topic type each word is assigned to, the model relies on two binomially distributed latent switch variables. The first switch variable determines if a word is generated from a background topic. If it is not, then the second switch determines whether a word is generated from one of the general health topics or from one of the aspects of an ailment associated with the tweet.

Paul and Dredze later proposed ATAM+ [62], an extension to ATAM that incorporates prior knowledge in the form of multinomial language models, which correspond to the symptoms and treatments for 20 diseases obtained from the articles on WebMD.com, as asymmetric Dirichlet priors for the corresponding ailment aspects in ATAM. They reported the correlation coefficients between the CDC ILI statistics and the proportion of the tweets assigned to the flu ailment by ATAM and ATAM+ to be 0.935 and 0.968 respectively, while the correlation coefficient between Google Flu Trends data and the CDC statistics for the same time period was 0.932. They also evaluated the feasibility of applying ATAM+ to several population health analysis tasks. One such task is monitoring behavioral risk factors by geographical region. To demonstrate the potential of ATAM+ for this task, they calculated the correlation coefficient between the proportion of tweets in each U.S. state that were assigned to a particular ATAM ailment and the state's risk factor rate for the same ailment represented by the corresponding variable in the BRFSS dataset published by the National Center for Chronic Disease Prevention and Health Promotion at the CDC [12]. The strongest reported positive correlation was between the proportion of residents in each state, who are smokers, and the cancer ailment, while the strongest negative correlation was between exercise and the frequency of posting a tweet associated with any ailment, indicating that the Twitter users in the states, where people generally exercise more, are less likely to become sick. The other tasks include geographical syndromic surveillance, when the ailments are tracked both over time and per geographic region (ATAM+ was able to detect several known patterns of allergies) and analyzing correlation of symptoms and treatments with ailments. The latter task is particularly important, since for many health conditions patients prefer not to visit their doctors, managing an illness on their own. The illness, symptoms and chosen treatments for people not visiting healthcare providers remain unreported and obtaining these statistics requires extensive polling of large populations. Therefore, ATAM provides an opportunity to quickly and easily collect these statistics from Twitter. One disadvantage common to both ATAM and ATAM+, however, is that they require specifying the number of general health topics and ailments a priori, which is not always feasible.

Social media can also be a source of accurate and up-to-date information on recreational drugs, such as their usage profiles and side effects, which is crucial for supporting a wide range of healthcare activities, including addiction treatment programs, toxin diagnosis, prevention and awareness campaigns and public policy. Recreational drug use is an important public health problem, as it imposes a significant burden on businesses (via absenteeism or presenteeism of employees), healthcare infrastructure and society in general. Paul and Dredze proposed factorial LDA (f-LDA) [64], a topic model, in which each word is associated with a K-tuple of latent factors (e.g. topic, perspective), in contrast to LDA, which associates only one latent topic variable with each word. In f-LDA, each K-tuple corresponds to its own multinomial distribution over the collection vocabulary and each document is represented as a multinomial distribution over all possible K-tuples. f-LDA can jointly capture these factors as well as interesting interactions between them, producing fine-

[12] `http://apps.nccd.cdc.gov/gisbrfss/`

grained topical summaries of user discussions related to particular combinations of factors. Factorial LDA uses a novel hierarchical prior over model parameters and can be used to automatically extract textual snippets that correspond to fine-grained information patterns, a simple form of extractive multi-document summarization. In [63] and [65], Paul and Dredze reported the results of applying f-LDA to the task of mining recreational drug usage trends from on-line forums. In particular, they collected the data from drugs-forum.com and organized f-LDA topics along three dimensions: drug type (e.g. amphetamines, beta-ketones, LSD, etc.), route of intake (injection, oral, smoking, etc.) and aspect (cultural setting, drug pharmacology, usage and side effects). For example, in their three-dimensional model a tuple (*cannabis*, *smoking*, *effects*) corresponds to a topic summarizing the health effects of smoking cannabis. They focused on five drugs (mephedrone, Bromo-Dragonfly, Spice/K2 and salvia divinorum), which have been only recently discovered and studied, and used tuple-specific word distributions estimated by f-LDA to create a summary for each aspect of using these drugs.

1.3.2 Detecting Reports of Adverse Medical Events and Drug Reactions

Adverse drug reaction (ADR) is defined as a “harmful reaction, resulting from an intervention related to the use of medical product, which predicts hazard from future administration and warrants prevention of specific treatment, or alteration of the dosage regimen, or complete withdrawal of the product from the market” [32]. ADRs and drug-related adverse medical events (or adverse drug events, ADEs) pose substantial risks to patients, who consume post-market or investigational drugs, since they can complicate their medical conditions, increase the likelihood of hospital admission and even cause death. Despite post market drug surveillance, ADEs remain the fourth leading cause of death in the United States. A large portion of adverse medical events have been ascribed to adverse interactions between different drugs, which are often caused by their shared action mechanisms and metabolic pathways. Unknown drug-drug interactions (DDIs) constitute a significant public health problem, as they account for up to 30% of unexpected ADRs. Most discovered adverse DDIs result in additional prescription precautions and contraindications or even complete withdrawal of a drug from the market. Traditionally, ADRs and DDIs have been detected based on four data sources: clinical trial data, chemical/pharmacological databases, EMRs and spontaneous reporting systems (SRSs), which have been developed and deployed by different countries around the world as part of their pharmacovigilance process. SRSs mostly rely on self-reports by patients: the Food and Drug Administration’s MedWatch site [13] and Adverse Event Reporting System (AERS) in the United States, EudraVigilance by the European Medicines Agency and International Pharmacovigilance system by the World Health Organization. All of these sources, however, have inherent limitations, since clinical trials suffer from the cohort bias and passive nature of spontaneous reports leads to low reporting ratios (only 1 to 10 percent of all reportable ADRs are normally reported through MedWatch). Although pharmaceutical companies are required to report all known adverse events and reactions, majority of such events are detected by physicians and patients, for whom the reporting is voluntary. As a result, many serious or rare ADRs or DDIs may not be timely detected and their overall number may be significantly underestimated [92].

Due to the high frequency, diversity, public availability and volume, user posts on social media platforms have a great potential to become a new resource for Internet-based near real-time pharmacovigilance and complement the existing surveillance methods based on using natural language processing techniques to analyze electronic health records [41]. In particular, Chee et al. [14] proposed a machine learning method for identifying potential watchlist drugs from the messages on Health and Wellness Yahoo! groups. They experimented with ensemble methods consisting of standard classifiers (Naïve Bayes and SVM) and two feature sets (bag-of-words lexical features and an expanded set based on drug and side effect lexicons as well as sentiment vocabularies) and were

[13] `http://www.fda.gov/Safety/MedWatch/`

able to identify the drugs that were actually withdrawn from the market. Bian et al. [3] proposed a method for large-scale mining of adverse drug events from Twitter consisting of two classification steps. In the first step, the tweets posted by the users of investigational drugs of interest are identified. In the second step, the historical posts of those users are accessed to identify their previous tweets about adverse side-effects of using those drugs. They used SVM with the Gaussian radial basis kernel as a classification model and experimented with both standard bag-of-words lexical features and semantic features derived by mapping the tweet terms into the concept codes from the Unified Medical Language System Metathesaurus (UMLS) [77]. Despite using standard techniques to optimize the classification model, such as scaling, grid-based kernel parameter searching and feature selection using one-way analysis of variance F-test, they were only able to achieve the classification accuracy of 0.74 and the mean AUC of 0.82 for the the first classification step and the classification accuracy of 0.74 and the mean AUC of 0.74 for the second step. The relatively low performance of the classification models was attributed to the noisiness of Twitter data (the abundance of fragmented and non-grammatical sentences, mis-spellings and abbreviations), which degraded the performance of the standard part-of-speech tagger that was trained on proper medical documents and used to map the terms in tweets to the UMLS concepts. Scanfeld et al. [76] analyzed the tweets mentioning antibiotics to identify the major types of their proper use as well as misunderstanding and misuse. Yang and Yang [92] proposed a method to identify DDIs directly from social media content. In particular, they first extracted the *n*-grams from the posts and comments on the drug forums of MedHelp on-line patient community and identified the drug names and drug reactions by matching the extracted *n*-grams to the ADR lexicon derived from the Consumer Health Vocabulary (CHV) Wiki [14]. CHV is a collection of terms and phrases commonly used by non-specialists to refer to medical concepts (e.g. “irregular heartbeat” for “arrhythmia”) that is compiled, reviewed and validated by healthcare professionals. After extracting drug names and reactions, adverse DDIs were identified by applying association rule mining. Using the data from the DrugBank database as a golden standard, this method was able to identify the known adverse DDIs between 10 popular drugs with 100% recall and 60% precision.

Frost et al. [43] proposed to use on-line patient communities to determine the prevalence of on-label versus off-label drug use (when healthcare providers prescribe a drug for non-FDA approved purpose). In particular, they examined the patient data for amitriptyline and modafinil, two medications that are widely prescribed off-label, and conducted a post-hoc analysis of how patients reported using these drugs and their side effects that informed even broader understanding of these already well-understood medications. Nkhasi et al. [58] provided an analysis of preventable and adverse medical events that were reported and explicitly ascribed to the actions or procedures of healthcare professionals on Twitter. They classified such errors according to the type (procedural versus medication) and the source (physicians, nurses or surgeons) and reported the proportions of each error class. They also found out that the majority of such events are either self-reported by patients or reported by their relatives, demonstrating the potential of leveraging social media platforms to obtain the first-hand patient perspectives on the errors at different levels of the healthcare system. Such information is extremely valuable in developing the strategies to improve patient safety procedures for the entire healthcare teams. They also found that patients and relatives reacted to the safety errors in a wide variety of manners. While some people expressed anger and frustration in response to errors, others found them humorous and had an easy time moving on or used humor as a coping mechanism.

1.3.3 Characterizing Life Style and Well-being

While there is an ongoing argument among psychologists about how happiness should be defined, few would deny that people desire it. Governments around the world are starting to put more

[14] http://consumerhealthvocab.chpc.utah.edu/CHVwiki/

and more effort into measuring subjective well-being in their countries, moving beyond the common economic-based indicators, such as gross domestic product. Surveys by organizations like Gallup and government agencies are increasingly including one or more subjective well-being questions into their questionnaires. Subjective well-being, as measured by life satisfaction is also an important public health issue. Its importance goes far beyond the obvious attraction of positive emotion, as it affects the overall health, productivity and other positive life outcomes. Studying life satisfaction and well-being is conceptually different from predicting flu or allergies, since its goal is to not only predict regional variation in happiness, but to also understand the factors and mechanisms contributing to it. In particular, Schwartz et al. [79] focused on the cognitive-based estimation of overall life satisfaction by studying the language of well-being. They collected one billion tweets over nearly one year timespan and mapped them to the U.S. counties, from which the tweets were sent. They used LASSO linear regression to predict subjective life satisfaction scores derived from the questionnaires using the words from either clusters of semantically related words derived from Facebook status updates using LDA [78] or manually constructed word lists (based on LIWC [87] and PERMA dictionaries [82]) as features and controlling for demographic information (age, sex, ethnicity) and indicators of socioeconomic status (income and education). Based on the analysis of the estimated model, it was determined that socioeconomic status control variables are more predictive than the terms from social media-specific LDA topics alone, which are in turn more useful than hand-crafted lists of words from LIWC and PERMA dictionaries. However, all three feature sets combined produced significantly more accurate results than either of them alone, confirming that the words in tweets convey more information than just the control variables. The key conclusion of this work is that the words used in a sample of social media posts created by a particular community (e.g. county) reflect the well-being of individuals belonging to this community. In particular, the words belonging to categories and topics related to physical activity and exercise (“training”, “gym”, “fitness”, “zumba”), pro-social activities (“money”, “support”, “donate”), community engagement (“meeting”, “conference”, “council”, “board”), work and achievement (“skills”, “management”, “learning”), religion and spirituality (“universe”, “existence”, “spiritual”, “nature”) are the strongest positive predictors, while the words associated with disengagement (e.g. “sleepy”, “tired”, “bored”) are the strongest negative predictors of life satisfaction.

Although to date there has been a wealth of behavioral science research examining the role of face-to-face interactions and real-life social networks in influencing a broad range of behavioral (such as alcohol consumption [70] and obesity [16]) and emotional (such as happiness [37], loneliness [12] and depression [69]) changes in individuals, very few works studied similar effects of on-line social networks. One notable exception is the work of Sadilek and Kautz [71], who applied regression decision tree using 62 features derived from both the text of 16 million geo-tagged tweets authored by 630,000 unique users located in New York City and social network structure of that sample to the task of predicting the cumulative probability of sickness of Twitter users based on the number of calendar days, during which they wrote at least one “sick” tweet in the past. They used several types of features covering different aspects of life and well-being: the on-line social status of individuals, their behavior and lifestyle, socioeconomic characteristics, intensity of contacts with sick individuals and exposure to pollution. The on-line social status of individuals was measured based on the properties of their Twitter social network, such as PageRank, reciprocity of following, various centrality measures (degree, communicability, eigenvector, betweenness, load, flow and closeness) and interactions with other users (e.g. how many times person’s messages got forwarded or “liked” and how many times other people mentioned that person in their messages). Lifestyle and behavior of users, such as how often they visit bars as opposed to gyms as well as how much time they spend in crowded public transportation, was measured by juxtaposing the GPS coordinates of their tweets with a database of 25,000 different venues and major public transportation routes in New York City. The main findings in this work are that physical encounters with sick individuals within different time windows (1, 4, 12, 24 hours) are strongly negatively correlated, while on-line social status is strongly positively correlated with a person’s health status. Besides that, the distance to pollution sources and the visits to polluted cites are the two single features having the strongest positive and negative correlations with the health status, respectively. Additionally, measures of social rank are highly cross-correlated and have almost identical high predictive power, together explaining over 24% of the variance in health status. Other highly predictive types of features include life style, pollution, number of sick friends and encounters with sick individuals. At the same time, individual contributions of census-based features such as poverty, education and race were found to be small, jointly accounting only for 8.7% of the variance unexplained by other factors.

1.4 Analysis of Social Media Use in Healthcare

Communication between patients and clinicians is at the heart of healthcare. The emergence of new social media resources such as social networks, instant messaging platforms and video chats has the potential to completely change the way doctors and patients interact. Hawn [48] points out that using social media in health education "is about changing the locus of control to the patient" and altering the relationships between care givers and care receivers, in which patient portals, EHR platforms, blogs and microblogs won't merely substitute for many one-on-one encounters with providers, but will also allow for deeper doctor-patient relationships. Besides helping to establish better doctor-patient relationships, leveraging social media in healthcare has the following benefits:

- Social media platforms can make it easier for severely ill patients who are home-bound or bed-bound to regularly communicate with their providers, since written communication may take less energy/effort than phone calls and can be paused if the patient needs to take a break during the communication to rest;
- Such platforms can narrow the information gap between providers and patients and make patients more engaged in their healthcare management and decision-making;
- Communications via social media would also be beneficial for patients who are seeing experts located in different parts of the state or even country for their health conditions.

The public health community is also considering how social media can be used to spread health information, with applications including health education and promotions, as well as larger scale scenarios, in which patients can "friend" their doctors and constantly share health information with them and receive advise.

1.4.1 Social Media as a Source of Public Health Information

Personal health data has been traditionally considered as private. However, with the emergence of collaboratively generated content platforms dedicated specifically to healthcare that view started to change. While health information systems vary in complexity and purpose, the predominant model is that of a central repository for all health information generated within clinical contexts (health history, diagnoses, allergies, current treatments) that is kept securely for view only by patients and their healthcare providers. And while there is growing demand by patients for access to their own health data, little is known about how other people with similar medical concerns can effectively use these data, if they are made available to them. A medical informatics working group asserted that the ideal personal health record is more than just a static repository for patient data. It should combine data, knowledge and software tools to help patients become active participants in their own care. Framing on-line patient interaction around sharing personal health information resulted in the emergence of healthcare-related Web 2.0 communities, in which the members exchange their knowledge and experience, educating each other. This way patients can be viewed as individual data stores, which if linked together with on-line social networks, can become part of global, dynamic and shared healthcare knowledge repository.

The popularity of social media resources can be leveraged to disseminate health information and conduct interventions. For example, in the dermatology community, the Sulzberger Institute for Dermatologic Education is sponsoring an Internet contest for the best video promoting sun safe behavior. Other examples include Twitter groups dedicated to certain medical conditions (e.g. a group for mothers of children with attention deficit disorder), YouTube videos on tobacco cessation and human papillomavirus vaccination campaigns. Vance et al. [88] analyzed the pros and cons of using social media to spread public health information to young adults and concluded that the pros include low cost and rapid transmission, while the cons include blind authorship, lack of source citation and frequent presentation of opinions as facts.

Verhaag [89] studied experiences, expectations and strategies of 84 healthcare organizations in using social media for external communication. In particular, she studied the activity, popularity and presence of these organizations on Facebook, Twitter, LinkedIn, YouTube, Google+, Pinterest as well as blogs and found that different social media platforms are not equally utilized and that the activity of organization on these platforms differs by their specific area. She found that health organizations generally have a Facebook and/or Twitter account, however, other social media platforms, such as Google+, blogs and YouTube are hardly used at all. In addition, health organizations most commonly use social media to spread information about themselves. Interviews with the employees of those organizations responsible for social media relations indicated that there is a need for “closed platforms”, in which the members have different levels of access to the content. Such platforms will be more suitable for private and sensitive information, which is common in the healthcare industry.

As behavioral interventions are becoming increasingly important in public health, the potential of using social media to study dissemination of health behaviors, sentiments and rumors among very large populations is unparalleled. Salathé and Khandelwal [75] assessed the spread of vaccination sentiments from person to person during the unfolding of H1N1 pandemic and found that anti-vaccination sentiments could be reliably assessed across time and that those sentiments tend to cluster in certain parts of on-line social networks. Their analysis also indicated that negative sentiments spread more effectively than positive ones. They also identified strong positive correlation between anti-vaccination sentiments and CDC estimates of H1N1 vaccination rates (i.e. vaccination coverage was higher in the regions with more positive sentiment).

1.4.2 Analysis of Data from On-line Doctor and Patient Communities

Fast and easy access to on-line health information resulted in patients relying on social media and the Internet more frequently than their physicians as a source of health information. In particular, Lau et al. [54] conducted an extensive study of social media use by Generation Y, people with low socioeconomic status and chronically ill populations. Emerging healthcare-related social media platforms also play an increasing role in on-line health searches. In many cases, people prefer to turn to social media groups, discussion forums and patient communities to express and discuss their fears and concerns for several reasons. The patients either may not feel comfortable disclosing their fears to providers or may wish to find other individuals in similar situation, who will listen to them, provide support and address their everyday issues and fears that healthcare providers may not realize. This particularly applies to the issues that are traditionally related to stigma, ridicule and rejection in a society. Social interaction through computer mediated communication services resembles face-to-face interactions, but offers greater anonymity and intimacy, which in turn results in higher levels of trust.

Both patients and doctors naturally seek to meet and interact with a community of other patients and doctors either to share their knowledge and experience or to receive support and advice. This type of dynamic on-line communication (called Health 2.0, by analogy with Web 2.0) now offers patients a unique opportunity to learn about their illness and gain support and knowledge from others with similar experiences. As a result, on-line patient communities can be used as a source of clinical data and patients' insights on the functioning of different aspects of the healthcare system. These platforms are based on two assumptions. First, given appropriate tools, patients will be able to interpret and learn from their own and others health data. Second, sharing personal health data and collaboratively reviewing and critiquing it will enhance the utility of the data for each contributor. A list of popular on-line patient communities is provided in Table 1.1.

Community	Description	Website
PatientsLikeMe	On-line community for patients to share their experiences and progress or to get input from others, who suffer from the same condition	`www.patientslikeme.com`
MedHelp	On-line patient community that partners with hospitals and medical research institutions to deliver on-line discussion boards on a variety of healthcare topics	`www.medhelp.org`
DailyStrength	Social networking platform centered on support groups, where users provide one another with emotional support by discussing their struggles and successes with each other	`www.dailystrength.org`
Inspire	Patient community organized around support groups related to medical conditions that are represented as a hierarchy	`www.inspire.com`
MediGuard	Patient and consumer network that helps patients to track their medications and exchange information with others	`www.mediguard.org`

TABLE 1.1: Popular on-line patient communities

PatientsLikeMe is an on-line platform built to support information exchange between the patients with life-changing diseases, which is organized around patient communities designated for specific conditions. PatientsLikeMe has more than 20 disease communities formed by more than 50,000 patients that anonymously share treatment options, symptoms, progression and outcome data for complex diseases. To make health information more accessible, this Web site provides visualization tools that help the patients understand and share information about their health status. Upon joining the site, patients enter a combination of structured and unstructured information about their health status and history, which is then processed and represented as a set of graphical displays on their profiles: a personal picture, an autobiographical statement, a diagram that maps a functional impairment to specific areas of the body, a diagnosis history and a series of charts. The "nugget" summary diagram displays the current function score as a color code mapped onto the affected areas of the body as well as the number of years with the disease, an iconic representation of the equipment currently used, and stars indicating the level of participation on the site. Each member can also see a graphical representation of their own and others health status, treatments and symptoms over time and can view reports of aggregated data. The site includes an interactive report for each treatment, medication and intervention that patients add to the system. Such reports include dosages taken, time on treatment, evaluations of treatment, including perceived efficacy, side effects and burden. Members can locate other patients in similar circumstances and with shared medical experiences using searching and browsing tools and discuss the profiles, reports and general health concerns through the forum, private messages and comments they post on one another's profiles.

Frost and Massagli [42] identified and analyzed how the users of PatientsLikeMe with incurable or rare life-altering disease reference personal health information of each other in patient-to-patient dialogues and found that discussions on the site fall into 3 major categories: targeted questions to other patients with relevant experience, proffering personally acquired disease-management knowledge or coping strategies, and forming and solidifying relationships based on shared health concerns.

On-line patient networks open up new ways of testing treatments and can speed up patient recruitment into clinical trials for new drugs [9]. Recent studies have also demonstrated that using on-line patient network data in clinical studies can accelerate discoveries related to complex conditions such as Parkinson's disease [90], amyotrophic lateral sclerosis (ALS) [91] and rheumatoid arthritis [85]. These platforms can also be used to identify shifts in patients' perceptions and behaviors in response to public health policies.

Many disease-specific groups have arisen on Facebook, representing important sources of information, support and engagement for patients with chronic diseases. Greene et al. [46] identified the 15 largest groups focused on diabetes management and evaluated a sample of discussions within them. They found that Facebook diabetes communities contain a plurality of participants, including patients, their family members, advertisers and researchers, with divergent interests and modes of communication. They use Facebook to share personal clinical information, request disease-specific guidance and feedback and receive emotional support. They also found that users posted individual concerns about possible adverse effects of medications and diet supplements in an attempt to see if their own experiences correlated with those of others. Furthermore, nearly a quarter of all the posts shared sensitive aspects of diabetes management unlikely to be revealed in doctor-patient interactions.

Many blogging and Twitter communities are also dedicated to specific health conditions. Chung et al. [20] studied dietdiaries.com, the community of bloggers focused on weight management, and compared the effectiveness of two approaches for the task of predicting weight loss from natural language use in blogs. The first approach is based on manually categorizing blog posts based on the degree of weight loss or gain reported in them and then using standard multinomial Naïve Bayes textual classifier with bag-of-words features to classify them into those categories. The second approach is based on the detailed linguistic analysis of blog posts leveraging Linguistic Inquiry and Word Count (LIWC) [87] categories. In this method, textual feature vectors are mapped into linguistic categories that are known to be associated with psychological constructs. The proposed method first computes correlations between LIWC categories and weight change and then uses linear regression to predict the percent of body weight change based on the distribution of LIWC categories, which have statistically significant correlations with weight change. The authors observed that the LIWC-based regression approach generally outperformed Naïve Bayes-based classification approach. In particular, they found that using more sadness-related words and fewer food ingestion-related words is a statistically significant predictor of weight loss, whereas the percent of body weight change was unrelated to the usage of positive emotion words (e.g. "awesome", "happy"), health words (e.g. "nausea", "sick") or social words (e.g. "friend", "hug"). The author's interpretation of these results was that sharing negative emotions is a more successful strategy in blogging about weight loss than simply keeping a food intake diary. Harris et al. [47] studied communication about childhood obesity on Twitter using descriptive statistics and exponential random graph modeling to examine the content of tweets, characteristics of users tweeting about childhood obesity and the types of Twitter followers receiving tweets about childhood obesity. They concluded that Twitter may provide an important channel for reaching traditionally difficult-to-reach populations, including lower income, Hispanic, and non-Hispanic Black groups facing significantly higher rates of childhood obesity than their higher income and non-Hispanic White counterparts.

Several researchers also focused on studying the content and social network structure of on-line communities for smoking cessation. Selby et al. [81] analyzed the content of the posts on StopSmokingCenter.net, an on-line social support network moderated by trained program health educators, as well as characteristics of the users who created them. They found that the majority of posters were female and that the most common theme of the posts was seeking support or advice with quitting. However, only 15% of the new members made at least one post on the support group boards and an even smaller fraction of users were active and consistent posters, suggesting that other self-quit program aspects (e.g. developing a strong sense of community) might be more appealing to the participants. Additional analysis revealed that 50% the the first-time posts were made relatively quickly (within three hours after joining the site). In their first posts, members most frequently conveyed that they were seeking support and advice. Replies to the first posts from other support group members were also quick, with 25% of the first posts receiving reply within 12 minutes and 50% within 29 minutes. Responses were even faster for the posts from the members that were actively seeking support, revealing that the support group board did function to provide members with an immediate source of support not available with most traditional interventions. Cobb et al. [21] used network analysis techniques to identify structural and functional characteristics of QuitNet [15], one of the largest and most popular continuously operating on-line communities focused on smoking cessation. They found that the members in the strongly and densely connected cores of QuitNet's social network are mostly older females (over 40 years old), that have been active and abstinent community members for more than a year. In a recent study by Corazza et al. [24], social media was also used to study a new drug methoxetamine.

Chuang and Yang [18] [17] identified and compared the level of different types of social support (informational, emotional and instrumental) across three different types of computer-mediated communication tools (discussion forums, personal journals and notes) on the MedHelp alcoholism support community and found that the patients use these communication tools for different purposes. Forum users are more likely to seek and provide informational support, while journals and notes are primarily used to express higher levels of emotional support. Similar qualitative content analyses of posts on on-line communities for health conditions such as irritable bowel syndrome [26], Huntington's disease [27] and HIV [28] have been conducted and identified that all five subtypes of social support (emotional, informational, esteem and social network) are evident in the posts, with informational and emotional support being offered most frequently. Silenzio et al. [84] studied characteristics of young lesbian, gay and bisexual population on Twitter and proposed several methods for effective peer-driven information diffusion and preventive care, specifically focusing on suicide prevention.

Besides patient communities, there is also a growing number on-line communities for healthcare professionals, which foster and facilitate the exchange of information, insights and knowledge about current medical practices, treatments and medications and generate epidemiological and clinical data that were previously dispersed between the physicians' charts, EMRs and clinical histories. A list of popular on-line platforms dedicated to healthcare professionals is presented in Table 1.2. Sermo is the largest such community with over 200,000 registered licensed MDs and DOs. DataGenno is a Web portal for healthcare professionals and researchers along with patients and their relatives to exchange information about rare genetic and complex diseases. It provides a database with sample and disease information along with the images for each sign or symptom, a search engine for differential diagnosis and features for information exchange between healthcare professionals. It has been designed to bridge the gap between healthcare professionals, scientists, genetic counselors, nurses and patients by combining clinical, genetic and genomic information for specific diseases. eMERGE is an NIH-funded collaborative project linking medical records data with genomic data.

A community-based social network for health professionals that combines traditional drug discovery informatics with Web 2.0 platforms and strong privacy is believed to be the key to facilitate richer collaborations between healthcare professional with the same interests [49].

[15] http://www.quitnet.com

Community	**Description**	**Website**
DataGenno	Interactive database containing molecular and clinical genetic information from diseases targeted to healthcare professionals, research scientists and patients	`www.datagenno.com`
eMERGE	The Electronic Medical Records and Genomics (eMERGE) network combines DNA repositories with electronic medical record systems for large-scale genetic research	`emerge.mc.vanderbilt.edu`
Sermo	Online network for physicians with panel discussions about specific topics	`www.sermo.com`
Ozmosis	Provides several solutions for physicians to share their knowledge and clinical experiences with each other	`www.ozmosis.com`

TABLE 1.2: Popular on-line communities for doctors and clinical researchers

1.5 Conclusions and Future Directions

As we have seen in this chapter, social media can be considered both as a comprehensive source of individual-level experiences to be used by patients for health self-education or by providers to inform clinical practice and as a nearly unlimited source of aggregate data for large-scale population studies. The main advantages of social media based approaches to public health research are that they do not require to explicitly recruit participants and can provide large volumes of data in near real-time at virtually no cost. Their major disadvantages are sample bias and trustworthiness of the data.

The issue of sample bias has to do with the fact that the demographics of social media does not fully represent the general population, with the elderly and young children being particularly underrepresented. For example, previous studies reported that Twitter users tend to be younger (nearly half are under 35 and only 2% are 65 or older). Twitter is also known to be centric to the United States: as of June 2009, 62% of Twitter users were physically located in the U.S. Furthermore, the exact demographics of social media population that actively generates health-related content is generally unknown and not easy to estimate. As a result, population bias may limit the type of public health research questions that can be answered with the help of social media data. Therefore, an interesting future direction is to study how the estimates obtained from social media can be adjusted to reflect the properties of the general population.

Another challenge limiting the use of social media in healthcare is related to inability of the existing methods to identify all health-related posts and messages with 100% accuracy as well as the general reliability of user-generated content. As more and more people become the users of social media, communication channels grow exponentially more diffused and the possibility of spreading inaccurate or problematic information increases accordingly. As a result, social media users have to aggregate often contradictory information from multiple sources and judge their credibility. It is also known that some users may never self-report any health condition no matter how serious it is (stoics), while others may report being sick even when in fact they are not (hypochondriacs). While some of this bias can be partially mitigated using heuristics, such as counting the number of days, during which a user posts “sick” tweets, it is very hard to remove it completely. Although for the purpose of identifying general population-level trends and important insights into an epidemic, the sample bias may not have a large effect, the lack of confirmation for the diagnosis presents certain challenges for validation, detailed analysis and interpretation of results obtained from smaller scale social media-based studies. Public health officials typically have reservations about integrating social media data into their reports as it could result in an additional burden to their surveillance responsibilities. However, social media is, by nature, a venue for two-way information exchange, where the users can verify and evaluate the quality of information shared by other users. Therefore, many existing on-line systems actively leverage this property of social social media by requiring the messages to be reviewed by a moderator (either before or after their public dissemination) and enabling the users to provide feedback and even corroboration of submissions, a strategy that has proven successful with Wikipedia. The initial work on automatically establishing the trustworthiness of social media data through cross validation with official sources was extensively discussed in this chapter. Nevertheless, this problem is far from being solved and constitutes another interesting and challenging future research direction.

Despite these limitations, social media-based methods clearly have the potential to become valuable additions to traditional public health monitoring and analysis systems, which can uncover the detailed biological mechanisms behind diseases and capture the signals that are presently too weak to be detected on-line.

Bibliography

[1] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Workshop on Cyber-Physical Networking Systems (CPNS 2011)*, 2011.

[2] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 1568–1576, 2011.

[3] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 International Workshop on Smart Health and Well-being*, pages 25–32, 2012.

[4] Celeste Biever. Twitter mood maps reveal emotional states of america. *The New Scientist*, 207(2771):14, 2010.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(12):993–1022, 2003.

[6] Sean Brennan, Adam Sadilek, and Henry Kautz. Towards understanding global spread of disease from everyday interpersonal interactions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*, pages 2783–2789, 2013.

[7] Joanna Brenner and Aaron Smith. 72% of online adults are social networking site users. `http://www.pewinternet.org/2013/08/05/72-of-online-adults-are-social-networking-site-users/`, 2013. Accessed 08-08-2014.

[8] David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS ONE*, 8(12):e83672, 2013.

[9] Catherine A. Brownstein, John S. Brownstein, Davis S. Williams, Paul Wicks, and James A Heywood. The power of social networking in medicine. *Nature Biotechnology*, 27:888–890, 2009.

[10] John S. Brownstein, Clark C. Freifeld, Emily H. Chan, Mikaela Keller, Amy L. Sonricker, Sumiko R. Mekaru, and David L. Buckeridge. Information technology and global surveillance of cases of 2009 h1n1 influenza. *The New England Journal of Medicine*, 362(18):1731–1735, 2010.

[11] John S. Brownstein, Clark C. Freifeld, Ben Y. Reis, and Kenneth D Mandl. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS Medicine*, 5(7):e151, 2008.

[12] John T. Cacioppo, James H. Fowler, and Nicholas A. Christakis. Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, 97(6):977–991, 2009.

[13] Herman Anthony Carneiro and Eleftherios Mylonakis. Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49(10):1557–1564, 2009.

[14] Brant W. Chee, Richard Berlin, and Bruce Schatz. Predicting adverse drug events from personal health messages. In *Proceedings of the 2011 AMIA Annual Symposium*, pages 217–226, 2011.

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

[16] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357:370–379, 2007.

[17] Katherine Y. Chuang and Christopher C. Yang. Helping you to help me: Exploring supportive interaction in online health community. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.

[18] Katherine Y. Chuang and Christopher C. Yang. Social support in online healthcare social networking. In *Proceedings of the 2010 iConference*, 2010.

[19] Rumi Chunara, Jason R. Andrews, and John S. Brownstein. Social and news media enable estimations of epidemiological patterns early in the 2010 haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene*, 86(1):39–45, 2012.

[20] Cindy K. Chung, Clinton Jones, and Alexander Liu. Predicting success and failure in weight loss blogs through natural language use. In *Proceedings of the 2nd International AAAI Conference on Weblogs and Social Media (ICWSM'08)*, pages 180–181, 2008.

[21] Nathan K. Cobb, Amanda L. Graham, and David B. Abrams. Social network structure of a large online community for smoking cessation. *American Journal of Public Health*, 100(7):1282–1289, 2010.

[22] Crystale Purvis Cooper, Kenneth P. Mallon, Steven Leadbetter, Lori A. Pollack, and Lucy A. Peipins. Cancer internet search activity on a major search engine. *Journal of Medical Internet Research*, 7(3):e36, 2005.

[23] D. L. Cooper, G. E. Smith, V. A. Hollyoak, C. A. Joseph, L. Johnson, and R. Chaloner. Use of nhs direct calls for surveillance of influenza - a second year's experience. *Communicable Diseases and Public Health*, 5(2):127–131, 2002.

[24] Ornella Corazza, Fabrizio Schifano, Pierluigi Simonato, Suzanne Fergus, Sulaf Assi, Jacqueline Stair, John Corkery, Giuseppina Trincas, Paolo Deluca, Zoe Davey, Ursula Blaszko, Zsolt Demetrovics, Jacek Moskalewicz, Aurora Enea, Giuditta di Melchiorre, Barbara Mervo, Lucia di Furia, Magi Farre, Liv Flesland, Manuela Pasinetti, Cinzia Pezzolesi, Agnieszka Pisarska, Harry Shapiro, Holger Siemann, Arvid Skutle, Elias Sferrazza, Marta Torrens, Peer van der Kreeft, Daniela Zummo, and Norbert Scherbaum. Phenomenon of new drugs on the internet: the case of ketamine derivative methoxetamine. *Human Psychopharmacology: Clinical and Experimental*, 27(2):145–149, 2012.

[25] Courtney D. Corley, Armin R. Mikler, Karan P. Singh, and Diane J Cook. Monitoring influenza trends through mining social media. In *Proceedings of the International Conference on Bioinformatics and Computational Biology (BIOCOMP'09)*, pages 340–346, 2009.

[26] Neil S. Coulson. Receiving social support on-line: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. *Cyberpsychology and Behavior*, 8(6):580–584, 2005.

[27] Neil S. Coulson, Heather Buchanan, and Aimee Aubeeluck. Social support in cyberspace: A content analysis of communication within a huntington's disease online support group. *Patient Education and Counseling*, 68(2):173–178, 2007.

[28] Constantinos K. Coursaris and Ming Liu. An analysis of social support exchanges in on-line hiv/aids self-help groups. *Computers in Human Behavior*, 25(4):911–918, 2009.

[29] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *KDD 1st Workshop on Social Media Analytics*, 2010.

[30] Aron Culotta. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Language Resources and Evaluation*, 47:217–238, 2013.

[31] Ed de Quincey and Patty Kostkova. Early warning and outbreak detection using social networking websites: The potential of twitter. In *2nd International eHealth Conference*, pages 21–24, 2009.

[32] I Ralph Edwards and Jeffrey K Aronson. Adverse drug reactions: Definitions, diagnosis and management. *The Lancet*, 356(9237):1255–1259, 2000.

[33] Jeremy U. Espino, William R. Hogan, and Michael M. Wagner. Telephone triage: A timely data source for surveillance of influenza-like diseases. In *Proceedings of the 2003 AMIA Annual Symposium*, pages 215–219, 2003.

[34] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[35] Gunther Eysenbach. Infodemiology: The epidemiology of (mis)information. *The American Journal of Medicine*, 113(9):763–765, 2002.

[36] Gunther Eysenbach. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *Proceedings of the 2006 AMIA Annual Symposium*, pages 244–248, 2006.

[37] James H. Fowler and Nicholas A. Christakis. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337:a2338, 2008.

[38] Susannah Fox. The social life of health information. `http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/`, 2011. Accessed 08-08-2014.

[39] Susannah Fox and Sydney Jones. The social life of health information. `http://www.pewinternet.org/2009/06/11/the-social-life-of-health-information/`, 2009. Accessed 08-08-2014.

[40] Clark C. Freifeld, Kenneth D Mandl, Ben Y. Reis, and John S. Brownstein. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157, 2008.

[41] Carol Friedman. Discovering novel adverse drug events using natural language processing and mining of the electronic health records. In *Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME'09)*, pages 1–5, 2009.

[42] Jeana Frost and Michael P Massagli. Social uses of personal health information within patientslikeme, an online patient community: What can happen when patients have access to one another's data. *Journal of Medical Internet Research*, 10(3):e15, 2008.

[43] Jeana Frost, Sally Okun, Timothy Vaughan, James Heywood, and Paul Wicks. Patient-reported outcomes as a source of evidence in off-label prescribing: Analysis of data from patientslikeme. *Journal of Medical Internet Research*, 13(1):e6, 2011.

[44] Jim Giles. Blogs and tweets could predict the future. *The New Scientist*, 206(2675):20–21, 2010.

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

[46] Jeremy A. Greene, Niteesh Choudhry, Elaine Kilabuk, and William H. Shrank. Dissemination of health information through social networks: Twitter and antibiotics. *Journal of General Internal Medicine*, 26(3):287–292, 2011.

[47] Jenine K. Harris, Sarah Moreland-Russell, Rachel G. Tabak, Lindsay R. Ruhr, and Ryan C. Maier. Communication about childhood obesity on twitter. *American Journal of Public Health*, 104(7):e62–e69, 2014.

[48] Carleen Hawn. Take two aspirin and tweet me in the morning: How twitter, facebook and other social media are reshaping health care. *Health Affairs*, 28(2):361–368, 2009.

[49] Moses Hohman, Kellan Gregory, Kelly Chibale, Peter J. Smith, Sean Ekins, and Barry Bunin. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discovery Today*, 14(5-6):261–270, 2009.

[50] Anette Hulth, Gustaf Rydevik, and Annika Linde. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4(2):e4378, 2009.

[51] Heather A. Johnson, Michael M. Wagner, William R. Hogan, Wendy Chapman, Robert T Olszewski, John Dowling, and Gary Barnas. Analysis of web access logs for surveillance of influenza. In *11th World Congress on Medical Informatics (MEDINFO 2004)*, pages 1202–1206, 2004.

[52] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 789–795, 2013.

[53] Vasileios Lampos and Nello Cirstianini. Tracking the flu pandemic by monitoring the social web. In *IAPR 2nd Workshop on Cognitive Information Processing (CIP 2010)*, 2010.

[54] A.Y.S. Lau, K.A. Siek, L. Fernandez-Luque, H. Tange, P. Chhanabhai, S. Y. Li, P. L. Elkin, A. Arjabi, L. Walczowski, C.S. Ang, and G. Eysenbach. The role of social media for patients and consumer health. contribution of the imia consumer health informatics working group. *Yearbook of Medical Informatics*, 6(1):131–138, 2011.

[55] Dennis D. Lenaway and Audrey Ambler. Evaluation of a school-based influenza surveillance system. *Public Health Reports*, 110(3):333–337, 1995.

[56] Jiwei Li and Claire Cardie. Early stage influenza detection from twitter. arXiv:1309.7340v3, 2013.

[57] Steven F. Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins University APL Technical Digest*, 24:349–353, 2003.

[58] Atul Nakhasi, Ralph J. Passarella, Sarah G. Bell, Michael J. Paul, Mark Dredze, and Peter Pronovost. Malpractice and malcontent: Analyzing medical complaints in twitter. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, volume FS-12-05 AAAI Technical Report, 2012.

[59] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, pages 122–129, 2010.

[60] National Library of Medicine / National Institutes of Health. Nlm technical bulletin: Mla 2006, nlm online users' meeting remarks. `http://www.nlm.nih.gov/pubs/techbull/ja06/ja06_mla_dg.html`, 2006. Accessed 04-20-2014.

[61] Michael J. Paul and Mark Dredze. A model for mining public health topics from twitter. Technical report, Johns Hopkins University, 2011.

[62] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, pages 265–272, 2011.

[63] Michael J. Paul and Mark Dredze. Experimenting with drugs (and topic models): Multi-dimensional exploration of recreational drug discussions. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.

[64] Michael J. Paul and Mark Dredze. Factorial lda: Sparse multi-dimensional text models. In *Proceedings of the Conference on Nueral Information Processing Systems (NIPS'12)*, pages 2591–2599, 2012.

[65] Michael J. Paul and Mark Dredze. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'13)*, pages 168–178, 2013.

[66] Camille Pelat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. More diseases tracked by using google trends. *Emerging Infectious Diseases*, 15(8):1327–1328, 2009.

[67] Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. In *Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP'11)*, pages 18–25, 2011.

[68] Joshua Ritterman, Miles Osborne, and Ewan Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st International Workshop on Mining Social Analytics*, 2010.

[69] J. Niels Rosenquist, James H. Fowler, and Nicholas A. Christakis. Social network determinants of depression. *Molecular Psychiatry*, 16:273–281, 2011.

[70] J. Niels Rosenquist, Joanne Murabito, James H. Fowler, and Nicholas A. Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, 2010.

[71] Adam Sadilek and Henry Kautz. Modeling the impact of lifestyle on health at scale. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*, pages 637–646, 2013.

[72] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM'12)*, pages 322–329, 2012.

[73] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geotagged micro-blog data. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pages 136–142, 2012.

[74] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*, pages 851–860, 2010.

[75] Marcel Salathé and Shashank Khandelwal. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology*, 7(10):e1002199, 2011.

[76] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188, 2010.

[77] Peri L. Schuyler, William T. Hole, Mark S. Tuttle, and David D. Sherertz. The umls metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217–222, 1993.

[78] H. Andrew Schwartz, Johannes C. Eichstaedt, Lukasz Dziurzynski, Margaret L. Kern, Martin E. P. Seligman, Lyle Ungar, Eduardo Blanco, Michal Kosinski, and David Stillwell. Toward personality insights from language exploration in social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, pages 72–79, 2013.

[79] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J. Park, Shrinidhi Lakshmikanth, Sneha Jha, Martin E. P. Seligman, Lyle Ungar, and Richard E. Lucas. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13)*, pages 583–591, 2013.

[80] Ari Seifter, Alison Schwarzwalder, Kate Geis, and John Aucott. The utility of google trends for epidemiological research: Lyme disease as an example. *Geospatial Health*, 4(2):135–137, 2010.

[81] Peter Selby, Trevor van Mierlo, Sabrina C. Voci, Danielle Parent, and John A. Cunningham. Online social and professional support for smokers trying to quit: An exploration of first time posts from 2562 members. *Journal of Medical Internet Research*, 12(3):e34, 2010.

[82] Martin E. P. Seligman. *Flourish: A Visionary New Understanding of Happiness and Well-being*. Free Press, 2011.

[83] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The use of twitter to track levels of disease activity and public concerns in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467, 2011.

[84] Vincent Michael Bernard Silenzio, Paul R. Duberstein, Xin Tu, Wan Tang, Naiji Lu, and Christopher M. Homan. Connecting the invisible dots: Network-based methods to reach a hidden population at risk for suicide. *Social Science and Medicine*, 69(3):469–474, 2009.

[85] Christof Specker, Jutta Richter, Ayako Take, Oliver Sangha, and Matthias Schneider. Rheumanet – a novel internet-based rheumatology information network in germany. *British Journal of Rheumatology*, 37(9):1015–1019, 1998.

[86] Donna F. Stroup, Stephen B. Thacker, and Joy L. Herndon. Application of multiple time series analysis to the estimation of pneumonia and influenza mortality by age 1962-1983. *Statistics in Medicine*, 7(10):1045–1059, 1988.

[87] Yla R. Tauszik and James W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[88] Karl Vance, William Howe, and Robert Dellavalle. Social internet sites as a source of public health information. *Dermatologic Clinics*, 27(2):133–136, 2009.

[89] Melissa L. Verhaag. Social media and healthcare - hype or future? Master's thesis, University of Twente, 2014.

[90] Paul Wicks and Graeme J. A. MacPhee. Pathological gambling amongst parkinson's disease and als patients in an online community (patientslikeme.com). *Movement Disorders*, 24(7):1085–1088, 2009.

[91] Paul Wicks, Timothy E. Vaughan, Michael P. Massagli, and James Heywood. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology*, 29:411–414, 2009.

[92] Haodong Yang and Christopher C. Yang. Harnessing social media for drug-drug interactions detection. In *Proceedings of the 2013 IEEE International Conference on Healthcare Informatics*, pages 22–29, 2013.