Social Data Integration and Analytics for Health Intelligence

Xiang Ji
New Jersey Institute of Technology
Supervised by James Geller & Soon Ae Chun
Newark, NJ, 07032, USA
xj25@njit.edu

ABSTRACT

The aim of this Ph.D. work is to develop a social data analytics framework for healthcare. The user-generated data can provide direct experience and opinions on medical conditions, treatments and insights on population health that can benefit clinical doctors, public health officials as well as patients and researchers. To utilize these social data, we address following research problems: (1) Semantic integration of data sources: Open health data are distributed and can vary from structured to highly unstructured. Different data sources need to be integrated. (2) How to mine sentiments and quantify the Degree of Concern (DOC) about contagious diseases, and how to extend sentiment mining by utilizing topic modeling to detect and quantify topic sentiments (3) Individual disease prediction based on social data. The social health analytics framework can provide healthcare providers with a tool to better utilize social data for evidence-based practice.

1. INTRODUCTION

1.1 Problems and Objectives

The research problems and objectives are described in detail as follows:

(1) The open health datasets such as PatientsLikeMe, PubMed, and WebMD, accessible through different platforms vary from structured to highly unstructured. An information seeker has to spend time visiting many, possibly irrelevant, websites, and has to select information from each and integrate it into a coherent mental model. The social health analytics framework will provide the semantic integration data model [1] that represents the semantic relationships of streaming data from distributed health data sources.

Monitoring and extracting of social media data involves large volumes of data in a highly unstructured format. The social health analytics framework incorporates a social media data ETL (Extract-Transform-Load) component that will be used to build the integrated data store. The data store will feed various visualization tools for public health status monitoring.

The existing public health sentiment surveillance methods [2], such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays. My social health analytics framework aims at providing algorithms [3] to perform real-time public health sentiment mining to supplement the current public health surveillance. Public health specialists would like to retrieve the general trend of health-related topics and topic-sentiments from social media, instead of reading through massive amounts of messages over a period of time. The social health analytics framework contains a *topic-modeling component* for social health data that extracts topics from health-related tweets and automatically generates overall sentiment polarity judgments for these health topics.

Due to the similar molecules, gene structures, and patients' life styles, the appearance of some conditions indicates the occurrence of other conditions [4]. This correlation is called *comorbidity relationship.* A disease prediction component based on the publicly available social network data will be developed to represent these comorbidity relationships, and to help doctors as well as patients to anticipate potential health problems.

1.2 Approaches

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

- Data Integration Component
- Population Analytics Component
- Predictive Analytics Component

The overall architecture is shown in Figure 1. The Data Integration component contains the Health Knowledge Semantic Model sub-component and the Term-Matching Algorithm sub-component. The Population Analytics component contains the Correlation Analysis sub-component, the Sentiment Mining sub-component and the Topic Modeling sub-component. The Predictive Analytics component consists of developing new temporal collaborative filtering technique to predict individual conditions.

The rest of the paper is organized as follows. In Section 2, the Data Integration component will be introduced. Section 3 will discuss the Population Analytics component. In Section 4, the Predictive Analytics component will be introduced. Finally, the conclusions and working directions are discussed in Section 5.

2. SOCIAL AND OPEN HEALTH DATA INTEGRATION

The Data Integration component will be introduced in this Section. There are different "open" health data sources available

Proceedings of the VLDB 2014 PhD Workshop

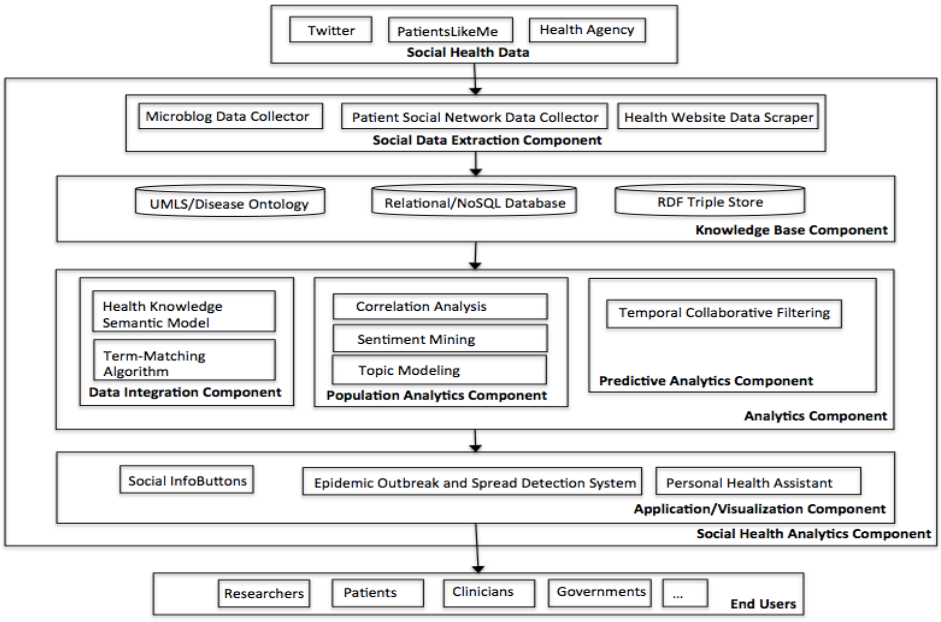

Figure 1 The component architecture of the social health analytics framework

online, the available information varies in formats and platforms, from unstructured to highly structured. When this information needs to be integrated, the data formats are usually not compatible. In this section, to address this heterogeneity problem, Semantic Web technologies are used to model the integrated knowledge network by linking health data from multiple data sources. The RDF triples provide a lightweight integration method through linking entities from different sources to represent the relationships among them.

The research goals in this section are:

(1) Extract patient-generated or reported social health data, open government health data, and clinical research data from heterogeneous data sources, and then link and combine them to develop an integrated semantic network model. (2) Develop a semantic web-based network model to represent, link, and combine the health knowledge from heterogeneous data sources. (3) Provide a set of analytic tools called "Social InfoButtons" that can help end-users to become aware of socially distributed health information.

2.1 Related Work

The Semantic Web has been used as a framework for data integration in previous research. Sheth et al. [5] reviewed the viability of the Semantic Web for data integration. In the field of Health Informatics, the study of Chun et al. [6] proposed a preliminary semantic integration model of different health data sources, that can help annotating social health blogs. In the work of Tofferi et al. [7], clinical trial data is integrated with drug data to support end users at finding an appropriate clinical trial for them to participate in, but their study does not include social data. InfoButtons was developed by Cimino et al. [8] to meet the clinician's information needs in the context of patient care. The HL7 InfoButton Standard [9] provides a standard mechanism for Electronic Health Records systems and knowledge resources to communicate, and implement InfoButton capabilities.

2.2 Methods

2.2.1 Semantic Model for Linking Health Data

To capture the variety of available health information, including patients' profile information, treatments, symptoms, side-effects, and their inter-relationships from social networks, a conceptual model was developed to capture the relationships among health data entities extracted from different sources. For PatientsLikeMe entities, the Condition class and Treatment class are the central concepts in the model. The Condition class has the "isConditionOf" relationship to the Patient class, the "hasSymptom" relationship to the Symptom class, and the "hasTreatment" relationship to the Treatment class. The Treatment class has the "hasSideEffect" relationship to the SideEffect class, and the "hasPurpose" relationship to the Purpose class. These semantic nodes are also linked to concepts from web sources other than PatientsLikeMe. For example, the Condition class can be linked to the web resource in WebMD that describes the corresponding condition. This link can be represented with the "hasResource" relationship to denote that the Condition in PatientsLikeMe has a related resource in WebMD. Similarly, the Condition class can be linked to a research paper that reports a study on the condition at the PubMed publication site by the relationship "isReportedBy."

The semantic model provides the integration schema with classes (nodes) and relationships (links), and has the flexibility of schema modification during integration, compared with relational databases, which requires "refactoring" operation.

2.2.2 Entity Extraction and Triple Representation

As reported in a survey of Luque et al. [10], most of the health websites do not provide APIs for researchers to retrieve data. In this paper, publicly available data was collected from PatientsLikeMe, PubMed, WebMD, the CDC website, and the UMLS Metathesaurus. The PHP HTML DOM Parser [11] was

utilized to scrape relevant information from the above websites. The retrieved structured data is stored in a Jena triple store.

The extracted health information is represented as triples <subject, predicate, object> to assert a statement, as described in the semantic knowledge model. To represent relationships between two instances, both subject and object are viewed as entities and identified by URIs, and linked together with a named relationship.

2.3 Social InfoButtons

Based on the semantic integration approach and perceived information needs, a prototype of the "Social InfoButtons" system was developed to answer a list of predefined questions, based on information needs of clinicians [12]. The questions are summarized in Table 1.

Table 1 Questions Answered by Social InfoButtons

Category	Questions
Statistics	Top conditions with the most patients?
Statistics	# of patients suffering from this condition?
Demographics	Patients suffering from this condition?
Demographics	Gender distribution of the patients?
Location	Patients distribution in state/country level?
Location	Where is the individual patient?
Condition	What are the symptoms of the condition?
Condition	What are the treatments of the condition?
Correlation	Difference between social and official data?

3. SENTIMENT MINING

The Sentiment Mining and Monitoring Disease Outbreak sub-components of the Population Analytics component will be discussed in this section. It addresses the problem of how to monitor public sentiments and disease outbreaks. Traditionally, it is hard to detect and monitor health-related concerns and changes in public attitudes to health-related issues. The existing surveillance methods, such as questionnaires and clinical tests, can only cover a limited number of people and results often appear with significant delays.

A novel framework, "Epidemics Outbreak and Spread Detection System" (EOSDS) [3][17], was developed for mining social network data, such as Twitter, to provide a tool for public health specialists and government decision makers to gauge the degree of concern (DOC) expressed in the tweets of Twitter users under the impact of spreading diseases.

3.1 Related Work

3.1.1 Twitter Sentiment Mining

In sentiment analysis of Twitter, Pandey and Iyer [13] stressed the significance of domain specific features other than common text features used in traditional information retrieval tasks. Barbosa and Feng [14] focused on automation of the training data generation process. Their work combined sentiment-labeled tweets coming from three sources: Twendz, Twitter Sentiment, and Tweet Feel. Twendz, Twitter Sentiment, and Tweet Feel are websites that allow users to monitor the sentiments of Twitter messages.

The above research achieved relatively good precision and recall on certain datasets, but it did not include a measure to quantify the sentiments, which is important for end users. In addition, all approaches focused on how to classify the tweets into positive/negative or positive/negative/neutral. However, we focused on the problem of how to distinguish between news tweets and personal tweets, and how to classify personal tweets into negative tweets and non-negative tweets. By developing a novel algorithm to address this new problem, we can provide a measure to quantify the temporal trend of negative sentiments.

3.1.2 Topic Modeling

In topic modeling, the state-of-the-art model is the Latent Dirichlet Allocation (LDA) method, developed by Blei et al. [24]. This model has been applied to problems in text mining and requires no manually labeled data, and it distills a collection of text documents into a distribution of words that tend to co-exist in similar documents. A topic is represented as a set of words. For extracting topics from short texts such as tweets, Ramage et al. [25] proposed a method called Labeled LDA, which incorporated the implied tweet-level labels, such as hashtags, replies, emoticons, etc. In this paper, LDA will be used to extend the sentiment mining to topic sentiment mining.

3.2 Sentiment Mining

We developed a novel two-step tweet sentiment classification algorithm [3] to quantify the degree of public concern (DOC) and to allow tracking the temporal trends of the DOC about a specific disease with a timeline chart. It also provides a concern map to explore the spatial distribution of the DOC. The key challenge is to identify tweets expressing negative emotions caused by each disease. We proposed to use an Asymmetric Four-Point Likert scale, where a tweet must be at one of the following four Likert Steps: (1) Strongly Negative, (2) Negative, (3) Neutral, (4) Positive. The goal of the two-step classification algorithm is to identify negative and personal tweets.

3.2.1 Collecting Health Tweets

The overall EOSDS data collection process can be described as ETL (Extract-Transform-Load) approach. EOSDS contains 8,043 inherited, developmental and acquired human disease names from the DO open-source medical ontology [15]. The extracted disease vocabulary is used as the source of keywords to monitor diseases for signs of becoming epidemics. The current prototype system monitors infectious diseases (1) listeria and (2) tuberculosis (3) swine flu (4) measles (5) influenza (6) meningitis; Mental Disorders: (1) Major depression (2) Generalized anxiety disorder (3) Obsessive-compulsive disorder (4) bipolar disorder; and Disasters: (1) natural disaster (2) air disaster. Any number of diseases can be potentially monitored given sufficient computational resources. The core component uses the Twitter Streaming API for collecting epidemics-related real-time tweets.

3.2.2 Two-Step Sentiment Classification

The sentiment analysis problem is approached in two steps. First, personal tweets are separated from news (non-personal) tweets; then the personal tweets are further classified into negative and non-negative (neutral) tweets. The clue-based classification [16] is integrated with a profanity list. This

combination improved the classification accuracy because of the unique style of Twitter's casual messages. After the personal tweets are extracted by the most successful of the personal/news classification algorithms, these personal tweets are used to recognize negative tweets by the Machine Learning algorithms. The two-step classification algorithm is illustrated in Figure 2.

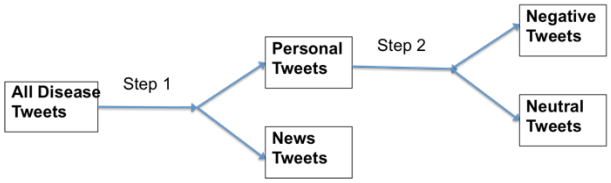

Figure 2 Two-step classification algorithm

3.2.3 Quantification of the Degree of Concern

After we identified the personal negative tweets, in order to quantify the negative sentiments in disease-related personal tweets, a concern map and a timeline chart of concerns were generated. These visual displays of negative sentiments toward a particular disease help detect the degrees of concern and high concern regions, which can help with preparing appropriate public health responses. The degree of concern for a certain disease at a particular date, *DOC [d, t]* is defined as:

$$DOC[d,t] = \frac{NN^2}{PN}$$

Where PN denotes the number of personal tweets and NN denotes the number of negative tweets for a particular disease "d" for a certain time "t." These are computed on a daily basis for each disease. The time, t, is in a daily scale, and the disease, d, can be any particular disease of interest. The DOC increases with the relative growth of negative tweets to personal tweets and with the absolute growth of negative tweets.

3.3 Monitoring Disease Outbreaks and Health Concerns

EOSDS contains a module to preprocess noisy geographic names. To enhance the visual analytics of the disease outbreak detection, EOSDS provides three mapping techniques for tweets. A static map displays tweet instances about a relevant disease, while a distribution map displays numbers of tweets from different states in the US relative to population size. A third display mode, the filter map, includes combined geospatial, temporal and user influence filters that provide a dynamic interface for a user to track tweet patterns for monitoring epidemic activity. The details and results of this module can be seen in the paper by Ji et al. [17].

According to the best available knowledge, there is no other similar tool to monitor degree of concern about the same epidemics. A qualitative study was performed on the DOC caused by listeria, measles, and TB from 10/18/2011 to 04/06/2012 and compared with related real-world incidents. Figure 3 shows the timeline chart of the 0-1 normalized DOC (blue line) and the 0-1 normalized number of tweets (red line) for these three diseases. In Figure 3A, peak 1 was recorded on 11/4/2011. On that day, the 29th death due to listeria was confirmed, making that outbreak the deadliest food borne illness in US history [18]. Unsurprisingly, the general public showed a large DOC because of it. In Figure 3B, peak 2 was caused by the reaction to a girl from Delaware that was infected with measles [19]. Peak 3 appeared, because at least one person confirmed to have measles was believed to have attended Super Bowl festivities [20]. In Figure 3C, peak 4 was the result of a TB case confirmed for a student from the University of New Brunswick (Canada) [21]. A TB outbreak among Occupy Atlanta protesters [22] caused peak 5. Also note that the DOC is not necessarily correlated with the volume of particular disease tweets. However, it indicates significant events related to the specific disease that public officials should pay attention to. Comparing the DOCs for the three diseases, listeria, measles and TB, in the given time range, peak 5 represents a DOC above 122 (before normalization). Thus, TB appears to be a greater reason of concern than measles and listeria in the observed geographic and time ranges. It is possible that listeria is not that well known in the US. Most people with measles get better by themselves [23]. On the other hand, TB needs to be treated, and the discovery of a new untreatable strain is certainly a reason for major concern.

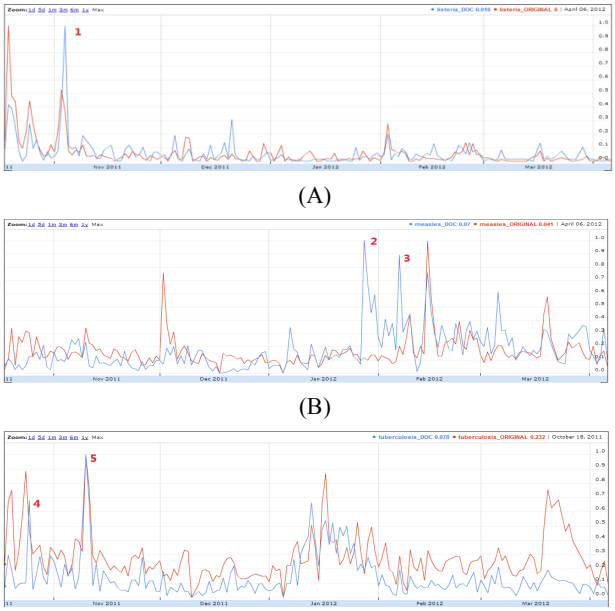

Figure 3 Timeline charts of Degree of Concerns (blue), raw number of tweets (red) of (A) listeria (B) measles (C) TB

3.4 Topic Sentiment Mining

These social network messages reflect people's sentiments and activities, but with a high degree of noise. Extracting the health-related topics and modeling the temporal trend of them from social media provides unique opportunities for public health specialists to investigate what are the trends in health topics and how the importance of different topics changes over a period of time. Sentiment mining can benefit from the extracted topics. Topic modeling is used to summarize the "important" topics, which are usually a set of words, from raw tweets. If the sentiment mining technique is applied to the content of topics instead of to the raw text, the performance can potentially improve, since the extracted topics are sets of "important" words and trivial information has already been eliminated. The research goal is to develop a topic modeling method based on Latent Dirichlet Allocation to model the temporal trend of events in epidemic-related tweets. In addition, we will develop a topic-sentiment model and compare the model with the raw

tweets-based sentiment analysis and DOC model discussed previously. The topic-sentiment analysis will refine the DOC model, as the DOC model takes into account all words in tweets.

We use a three-step approach to extract the health tweets topics and topic-sentiments (Figure 4). The first step is loading and filtering. In this step, the tweets are extracted and stored into text files labeled by date, and each text file is sent to the topic model as an instance for processing. The traditional topic model uses all words as inputs, but not all words contribute equally to the topic modeling. To overcome this shortcoming, in our method, a TF-IDF method is used to filter the words before training with LDA topics. The intuition is to "learn" the informative words from the tweets, besides filtering out stop words from a pre-defined corpus. Secondly, the LDA model will be applied to the filtered tweets to extract N topics. Each topic is represented by a set of words and is associated with a weight value from 0 to 1. Thirdly, sentiments of words in each topic will be analyzed and quantified. By this three-step approach, the trend of topics and topic-sentiments over a period of time can be measured.

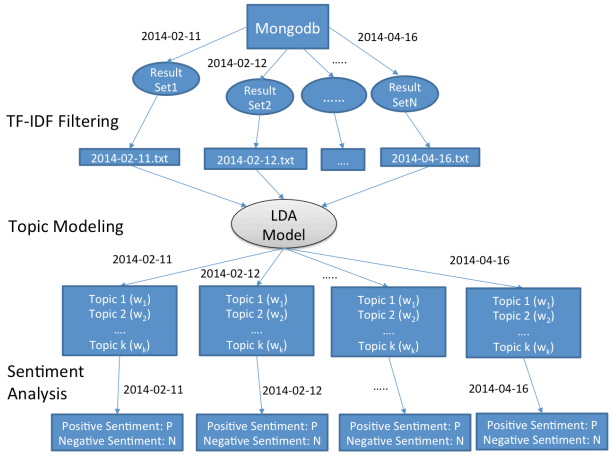

Figure 4. Method of extracting topics and topic-sentiments

Formally, suppose we have n days of tweets. For topic modeling, the input $D=\{D_1, D_2, ..., D_k, ... D_n\}$, where D_k is a set containing all TF-IDF filtered tweets in day k. The output is $T=\{T_1, T_2, ... T_k, ... T_n\}$, where T_k is the topic distribution of tweets in day k is defined as $T_k = \{<t_1, w_1, k>, <t_2, w_2, k>, ... <t_N, w_N, k>\}$ where t_i is i-th topic in day k, w_i is the weight of topic i and $w_1+w_2+ ... + w_N = 1$. For topic sentiment, the input is the output of topic modeling: $T=\{T_1, T_2, ... T_k, ... T_n\}$, and the output $S = \{S_1, S_2, ..., S_n\}$, where S_k is the topic sentiment of day k. $S_k = <P_k, N_k>$ where P_k is the positive sentiment of day k, and N_k is the negative sentiment of day k. $P_k + N_k = 1$.

The MPQA subjectivity lexicon dictionary [26], AFINN [27], and LIWC [28] can be used to detect polarity of emotional signals for topics. In LDA, every topic consists of N keywords, and each keyword is associated with a weight value. When sentiment analysis techniques are applied to the weighted topic keywords, each topic will be measured in the positive sentiment dimension and the negative sentiment dimension. Since the proportion of each topic changes with time, the weighted sentiment of every topic changes as well.

Our initial experiment shows that the sentiments mined from the topics is more "stable" than the sentiments from the raw tweets.

4. PREDICTION OF HEALTH CONDITIONS

Research has shown that some conditions are correlated with each other ("co-morbidities"). If comorbidity relationships can be predicted by a model, the potential conditions could be discovered more quickly, and the treatments could be more effective at the early signs of a disease. The research goal is to develop a technique, extending recommender systems, to predict the most probable conditions a patient will experience in the future, given a set of conditions that he already had in the past.

4.1 Related Work

A recommender system (RS) is a system that predicts the ratings or preferences that users would assign to an item that they have not seen or heard yet. A utility value will be calculated for each unknown item, and usually the system will generate a ranked list of items that users might be interested in, according to the utility value of each item on the list.

Davis et al. [29] proposed CARE, which is the first system that applied recommender system techniques to patient disease prediction. They used ICD-9 to represent the diseases. Similarities of each pair of patients are computed by vector similarity, and the vector similarity is adjusted by inverse frequency, which gives high weights to rare diseases.

As indicated by previous research, personal health records (PHRs) were utilized as the dataset to create the recommender system. Analogous with PHRs, the patients' profiles on social network can also be used as the dataset based on which future diseases are predicted. With social data, researchers do not need to worry about the privacy of the PHRs.

4.2 Collaborative Health Prediction Model

Analogous to collaborative recommender systems, the patients are coded as users, diagnosed conditions as items, and presence or absence of conditions as a rating with a binary value. The technique is able to predict what the most probable conditions a patient will exhibit in the future are, based on the conditions that the patient was already diagnosed with in the past.

There are two scenarios of prediction, the first one is to predict a new user's future conditions, and the second one is to predict each existing user's conditions after a "cut point" given the conditions before the cut point. The predicted risks of conditions can help users with preventive healthcare, available treatment options, related social consequences and expected side effects.

5. CONCLUSIONS AND WORKING DIRECTIONS

We presented different components of the Social Data Integration and Analytics framework, addressing research issues, related works, our approaches and prototype applications with preliminary results. Each individual component is as follows:

(1) For the Data Integration component, we proposed a semantic integration data model. The SPARQL query-based semantic search operations are employed, but we plan to utilize the more appropriate triple store. Currently data collection is automatic but not in real time, so it is desirable to expand the data collection process into a batch procedure or a real-time process.

(2) For the Population Analytics component, the challenge is real-time surveillance of Epidemics. The goal is to build a batch-processing job, which will act as a pipeline, starting with real-time data preprocessing, continuing to real-time analytics, such as sentiment classification and topic extraction, and ending with real-time rendering of analysis results through web pages.

(3) For the Individual Disease Prediction component, temporal and non-temporal similarity measures between patients will be applied and the results will be compared. The *neighborhood pickup function*, which will incorporate a threshold and a top-N selection method, will be developed to restrict the selection of neighborhoods. Different parameters, such as neighborhood threshold and head size offset will be tuned for achieving better prediction results. Medical experts will evaluate the results.

6. ACKNOWLEDGMENTS

The work is partially supported by Intelligent Automation Inc. for their summer internship training and by CUNY-PSC grant.

7. REFERENCES

[1] Ji, X., Chun, S. A. and Geller, J. Social infobuttons: integrating open health data with social data using semantic technology. In *Proceedings of the Fifth Workshop on Semantic Web Information Management*, New York, 2013.

[2] *Disease Control Priorities Project*. http://www.dcp2.org/file/153/dcpp-surveillance.pdf

[3] Ji, X., Chun, S. A. and Geller, J. *Monitoring Public Health Concerns Using Twitter Sentiment Classifications*. In *Proceedings of International Conference on Health Informatics,* Philadelphia, PA, 2013.

[4] Angold, A., Costello, E.J. and Erkanli, A. Comorbidity. *J. Child Psychology and Psychiatry*, 40, 1 1999, 57-87.

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

[6] Chun, S. A. and MacKellar, B. Social health data integration using semantic Web. In *Proceedings of the Proceedings of the 27th Annual ACM Symposium on Applied Computing*, Trento, Italy, 2012.

[7] Tofferi, J. K., Jackson, J. L. and O'Malley, P. G. Treatment of fibromyalgia with cyclobenzaprine: A meta-analysis. *Arthritis and rheumatism*, 51, 1 (Feb 15 2004), 9-13.

[8] Cimino, J. J., Li, J., Bakken, S. and Patel, V. L. Theoretical, empirical and practical approaches to resolving the unmet information needs of clinical information system users. In *Proceedings of the American Medical Informatics Association Annual Symposium, 2002*, 170-174.

[9] Del Fiol, G., Huser, V., Strasberg, H. R., Maviglia, S. M., Curtis, C. and Cimino, J.J. Implementations of the HL7 Context-Aware Knowledge Retrieval ("Infobutton") Standard: challenges, strengths, limitations, and uptake. *Journal of biomedical informatics*, 45, 4 (2012), 726-735.

[10] Fernandez-Luque, L., Karlsen, R. and Bonander, J. Review of extracting information from the Social Web for health personalization. *Journal of Medical Internet Research*, 13(1):e15, 2011.

[11] *PHP Simple HTML DOM Parser*. http://simplehtmldom.sourceforge.net.

[12] Collins, S. A., Currie, L. M., Bakken, S. and Cimino, J. J. Information needs, Infobutton Manager use, and satisfaction by clinician type: a case study, Journal of the American Medical Informatics Association, 16(1): 140-142, 2009.

[13] Pandey, V. and Iyer, C. V. K. *Sentiment Analysis of Microblogs*. Stanford University, 2009.

[14] Barbosa, L. and Feng, J. Robust sentiment detection on Twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 2010.

[15] Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W., Mazaitis, M., Felix, V., Feng, G. and Kibbe, W. A. Disease Ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40, Database issue (Jan 2012), D940-946.

[16] Wiebe, J. and Riloff, E. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th international conference on Computational Linguistics and Intelligent Text Processing*, Mexico City, Mexico, 2005.

[17] Ji, X., Chun, S. A. and Geller, J. Epidemic outbreak and spread detection system based on twitter data. In *Proceedings of the First international conference on Health Information Science*, Beijing, China, 2012.

[18] *WebMD Health News*. http://www.webmd.com/food-recipes/food-poisoning/news/20111104/listeria-cantaloupe-death-toll-sets-record.

[19] *DHSS Press Release*. http://www.dhss.delaware.gov/dhss/pressreleases/2012/measlescaseidentified-013012.html.

[20] *CBS News*. http://www.cbsnews.com/8301-504763_162-57373848-10391704/measles-patient-attended-super-bowl-village-health-officials-warn/.

[21] *CBC News*. http://www.cbc.ca/news/canada/newbrunswick/story/2011/10/25/nb-tb-unbsj.html.

[22] *Fox News*. http://www.foxnews.com/health/2011/11/11/occupy-atlanta-shelter-tests-positive-for-tuberculosis/.

[23] *Communicable Disease Control and Prevention, http://www.sfcdcp.com/measles.html.*

[24] Blei, D. M., Ng, A. Y. and Jordan, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research 3 (2003)*, 993-1022.

[25] Ramage, D., Dumais, S. and Liebling, D. *Characterizing microblogs with topic models*. In *Proceedings of the Fourth International Conference on Weblogs and Social Media,* Washington DC, 2010.

[26] Wilson, T., Wiebe, J. and Hoffmann, P. *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing,* Stroudsburg, PA, 2005.

[27] Nielsen, F. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs, arXiv:1103.2903 .

[28] *Linguistic Inquiry and Word Count (LIWC)*. http://www.liwc.net/.

[29] Davis, D.A., Chawla, N. V., Christakis, N. A. and Barabasi, A. L. Time to CARE: a collaborative engine for practical disease prediction. *Data Min. Knowl. Discov.*, 20, 3 2010, 388-415.