HealthTvizer: Exploring Health Awareness in Twitter Data through Coordinated Multiple Views

Tommy Dang
Computer Science Department
Texas Tech University
Lubbock, TX 79409
tommy.dang@ttu.edu

Ngan V. T. Nguyen
Computer Science Department
Texas Tech University
Lubbock, TX 79409
Ngan.V.T.Nguyen@ttu.edu

Vung Pham
Computer Science Department
Texas Tech University
Lubbock, TX 79409
vung.pham@ttu.edu

***Abstract*—Analyzing public user posts and shared information on social media can assist us in measuring various population characteristics, patterns, movements, and as well as the public health conditions. In recent years, researchers have been analyzing social media (such as Facebook or Twitter feeds) to detect and predict various emerging events and market trends. Fewer attentions have been paid to the epidemic of the diseases. In this paper, we present a social media analytics tool, called *HealthTvizer*, for exploring health awareness using Twitter data through interactive and interconnected multiple views. We use topic modeling to pick the relevant and meaningful terms from more than 57 million tweets. We detect the disease name and related contents which are shared by the users of different geographical locations (mostly in the United States). We believe that the collected geolocations from the users' tweets can reveal the patterns of diseases for a given term which allows a researcher to detect, analyze, and explore information about the diseases and hence take necessary steps to improve public health awareness. We validate the effectiveness of *HealthTvizer* through an informal user study. The feedback from this study also motivates us on interesting future extensions of the tool.**

***Index Terms*—Twitter data, coordinated multiple views, social media, public health, word clouds.**

I. Introduction

Micro-blogging and social media data can help in predicting a variety of events and also able to detect the flow and evolution of the events and incidents like presidential elections [1], flu spread [2], etc. The role of social media is not only limited to predicting and analyzing the events, but it also can be useful to understand the socio-economic condition of the users of particular geographical locations, to detect behavioral patterns [3], and to assist in a single person daily lifestyle [4] for a healthy routine. With the fast growing and adaptation of social media use by the people, we are now acquiring lots of various types of information. In a single day, approximately 500 million new tweets on Twitter and 216 million posts at Facebook is added in 2016 [5]. Those data can play a significant role in addressing various problems as well as exploring solutions.

Although researchers represent a variety of techniques and approaches to use social media data based on data and text mining, we find a gap in visualizing social media data to analyze health issues and situation. Visualization can help us to explore the large text corpora in a very effective way and further to research more on specific topics or events. This idea motivates us to analyze, process, and visualize more than one hundred and thirty thousands of tweets to find out the valuable insights from those tweets in an interactive and interconnected multiple views.

Coordinated multiple views have become an important tool in visual analytics [6]. The principle is putting multiple views side by side. The data samples interactively selected from one view are instantly highlighted in all other linked views. The such views include: force directed layouts [7], scatterplot matrices [8], [9], parallel coordinates [10], [11], or time series graphs [12], [13]. Coordinated multiple views have been integrated into visual analytics platforms such as Jigsaw [14], Tableau [15], or Xmdv [16].

We develop *HealthTvizer* tool to discover the topics using multiple interactive graphs and visualization techniques. The principal aim of *HealthTvizer* is to navigate the dynamics of the topics over time and geographical location to explore the frequent topics about various diseases on Twitter. The tool provides an intuitive, interactive exploration of the disease topics and what people are writing about those diseases along with the aggregate geographic information which helps to find out the spread of a disease on the map. We show frequent users' tweets for different diseases to analyze if there is a relationship between users and a disease.

Our main contributions in this paper are:

- We present a new approach for discovering disease-related topics using topic modeling and visualization. We implement an interactive data analytic prototype, *HealthTvizer*, for detecting, visualizing, and analyzing topic abstractions and how they change over time. The interface contains multiple linked visualizations: The bar chart for top 20 users, the US map for hot diseases on Twitter, the streamgraph for top keywords, and a force-directed network showing the relationships of hot keywords and diseases.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

- We demonstrate *HealthTvizer* on more than one hundred and thirty thousand disease-related tweets (extracted from over 57 million tweets) in January 2017.
- We conduct informal user studies with domain experts to validate our approach.

The paper is structured as follows: We describe related work in the following section. Then we introduce our *HealthTvizer* interface and its components. We illustrate the use of *HealthTvizer* on Twitter data in January 2017 and present the results of our user study in Section IV. Finally, we conclude the paper and discuss future work.

II. Related Work

A. Topic Modeling

Latent Dirichlet Allocation (LDA) [17] is used to find potential topics in the text corpora using a flexible generative probabilistic model for data mining based on the words. LDA classifies the topics based on distribution across the words through a three-level hierarchical Bayesian model. In this approach, documents are interpreted as a distribution over topics and topics are formed by a certain number of correlated words.

LDA is used in the different research applications. Doyle et al. [18] implement basic LDA techniques to model the financial topics for the stock market to detect the companies which show similarity in stock price movements. Wang et al. [19] extended LDA into *Spatial Latent Dirichlet Allocation (SLDA)*, which encodes spatial structures among visual words into the same topic. For source code modeling and numerical data visualization, a tool was developed based on LDA by Zou et al. [20] named *LDAAnalyzer*. Jordan Boyd-Graber et al. develops a *Multilingual Supervised Latent Dirichlet Allocation (MLSLDA)* [21] that discovers connections across languages which can produce the underlying structure in parallel corpora, find sentiment correlated word lists in multiple languages. Chen et al. [22] propose an interactive visualization system that extracted the sentiment information from hotels reviews and classifies the sentences into topics using LDA.

SparseLDA [23] is an efficient algorithm for a faster sampling of topics which required less memory than traditional LDA-based systems. Yuening Hu et al. [24] introduce an interactive tree-based topic modeling framework which enables the users to provide feedbacks and encode those feedbacks into the model to provide better and related topics with less incertitude. This approach extends the SparseLDA to make a more efficient topic modeling algorithm with user interactions.

Jin et al. [25] implement the SEIZ model (susceptible, exposed, infected, skeptic) on the Twitter datasets to explore the related topics. Using non-linear least squares optimization this approach model the data and prove it's effectiveness by modeling rumors articles on social networks. The authors report that how SEIZ modeling can accurately catch the information propagated over different news and rumors. Topic modeling techniques have to perform some careful processing on tweets due to the noisy nature of hashtags. Twitter-LDA [26], and the behavior-topic model [3] were designed to explicitly model tweets. The behavior-topic model analyzes tweets posting behavior of each topic for user recommendation. The idea is to model the user interest topics along with the user behavioral patterns.

Topic modeling allows us to discover, organize, re-order, and summarize the topics from the large text corpora in an efficient way and hence many good visualization tools and techniques have been developed for the visualization of topics based on topic modeling. TIARA [27], a topic visualization tool developed by Wei et al. which determines time-sensitive keywords to portray the content evolution of each topic over time using stack graph metaphor. ParallelTopics [28] represent the temporal changes of topics using Parallel Coordinates view. It enables the user to examine large text corpora in a structured way to understand the correlation between the terms. In this paper, we use the disease ontology defined in the Healthcare Hashtag Project [29] in order to quickly narrow down to disease-related tweets discussed in Section III-A.

B. Social media analysis and Text visualization

Twitter, Facebook, and other social media encourage frequent user expressions of their thoughts, opinions and random details of their lives. People share various important information and events in social media as well as some asinine comments. However, the shared information can be highly useful for discovering significant patterns and valuable insights.

Barbosa et al. [30] analyze the tweets and classify those using sentiment data. Sentiment information can help to detect the flow of people opinions and thoughts in various events. It can be useful for measuring and predicting the incidents like an election result which is represented by Tumasjan et al. [1]. Lerman et al. [31] monitor the spread of news on Twitter and Digg. They analyze the dynamics of information and the flow of news and their impact on various events like different voting circumstance. Jin et al. [25] model the Twitter data for analyzing and detecting the spreading of the news and rumor over the Twitter. They show the impact and usefulness of social media for eight different events of which four are a real event, and four are rumors.

Diakopoulos et al. [32] present a visual analytics tool with multiple linked views which can assist journalist to obtain rough proxies of public response by collecting, analyzing, aggregating and visualizing content from social media about specific events. Xu et al. [33] studied and analyzed temporal topic competition over time in social media. Their tools visualize the frequent topics over the time and expose topic evolution with the various events. TextFlow [34] is a visualization tool which can depict the topic evolution and event occurrences to illustrate topic merging or splitting relationships in a text stream. Drk et al. [35] introduced a web-based system that can provide a visual summary of large-scale Twitter data streams. The authors present three interactive side by side visual system to allow a user to explore the events according to time, topics, and specific people.

C. Public health analytic and visualization

Along with the flow of news and information, people also express their situation as well as the health and disease information of their friends and family. Sometimes that information is helpful to detect the spread of the diseases.

Lampos et al. [36] show that aggregating tweets can reveal the spread and situation of a disease. They analyzed and represented the rate of influenza, and it's spread in the United Kingdom and the United States. Sadilek et al. [37] propose a model to predict disease transmission based on geo-tagged micro-blog data. By constructing a probabilistic model, they analyze and detect the likely spread of a disease based on social ties and co-location of the people. Laurent et al. [38] perform a systematic review on possibilities of using Twitter as a research tool in the health area. They classify the uses of Twitter data into six categories with their usability and some limitations. Barnwal et al. [39] discuss the effectiveness of using Twitter data in health care. They demonstrated through the experimental setup how social media like Twitter could help a researcher to find out the are top health-related topics which are being discussed over the years.

Tweeting to Health [4] is an intervention which uses Fitbits, Twitter, and gamification to analyze and support the user to provide a healthy lifestyle. In a recent study, Xu et al. [40] show that Twitter or social data analysis not only able to help in detecting disease condition and situation but also it may assist in reducing racial and ethnic disparities. The authors try to explore the differences in cancer-related tweeting by race and ethnicity and find out the problems and discrimination.

As mentioned above, there are many efforts for analyzing the social data to detect, explore, and mining social media data. However, fewer attentions have been paid to reveal patterns of diseases. In this paper, we aim to visualize the data from a lot of tweets to explore and detect valuable insights about public health information.

III. *HealthTvizer* VISUALIZATION

When dealing with public health information, domain experts concern on the following issues:

- **Q1**: What public health information can be learned from Twitter? [41]
- **Q2**: Are there any geographical areas where people are prone to specific diseases?
- **Q3**: Is there any relation of diseases with exercise or general daily behavior which we can infer from tweets? [41], [42]
- **Q4**: How will Twitter data impact the study of community health behaviors, mental health, and biosurveillance customized to specific demographic groups? [43]
- **Q5**: Which diseases are more discussed in social media and if there is a gap with respect to an actual scenario? [39]

We are not trying to answer all of these research questions. In this paper, we rather focus on a subset of these questions, specifically question 1, 2, 3, and 5. In this section, we describe various components in our prototype as depicted in Figure 1. We further validate this prototype in Section IV.

Shneiderman [44] suggests three step metaphor for information visualization: overview first, zoom and filter, and then details-on-demand. The *HealthTvizer* implements low-level analysis tasks based on this design principle [45]:

- **T1**: Provide a summary view of popular diseases and keywords over time [46]. Moreover, the spatial distributions of Twitter users can be quickly highlighted on a map to show the hot hubs of health concerns and frequent users for each disease.
- **T2**: Details on demand [44] – Displaying the actual tweets of the users on hovering the user ID in the bar chart.
- **T3**: Data filtering [47] – Updating and populating the most related terms cloud and streamgraph for the particular disease.
- **T4**: Order terms in the streamgraphs as well as the word clouds based on their frequency. Similarly, most frequent users can always be found on top of the bar chart.
- **T5**: Our *HealthTvizer* represents the relationships among the hot keywords and diseases in a force directed layout.

A. Processing Twitter data

The study described in this paper uses tweets geolocated mostly in the United States and collected from Jan 6 to Jan 31, 2017, which includes 57,934,000 tweets. We query Datasift's streaming API to collect tweets that also have meta-information including geographical coordinates, Twitter places, user profile location, among other information.

a) Geolocation identification: Firstly, we process the geographical coordinates and the self-reported location string in the user's profile metadata. In case no geolocation presents in the user's profile, we check the mentioned place in the body of tweets, and infer that as the event location. If no geolocation information is available, we will discard that piece of tweets.

b) Tweets filtering: We are interested in the disease-related tweets. According to disease ontology defined in the Healthcare Hashtag Project [29], there are 21 main disease categories such as *Cancer*, *Psych (mental disease)*, *Blood disease*, etc. Each category contains a list of keywords which can be referred to the disease. By filtering this set of keywords, we are able to get each disease's tweets. The total number of disease-related tweets we could obtain is 12,6915, accounting for 0.22% of the total number of tweets from Jan 6 to Jan 31, 2017.

B. Components

HealthTvizer contains coordinated multiple views which highlight different tweet attributes: User ID, geolocation, keywords, time stamp, and the related disease. This section provides the detailed descriptions of these components.

1) Word Cloud and Interactive Streamgraph: Figure 2(a) shows the 21 diseases obtained from the Healthcare Hashtag Project [29]. Their detailed frequency over one month period is highlighted in the streamgraph in Figure 2(b). The fluctuated

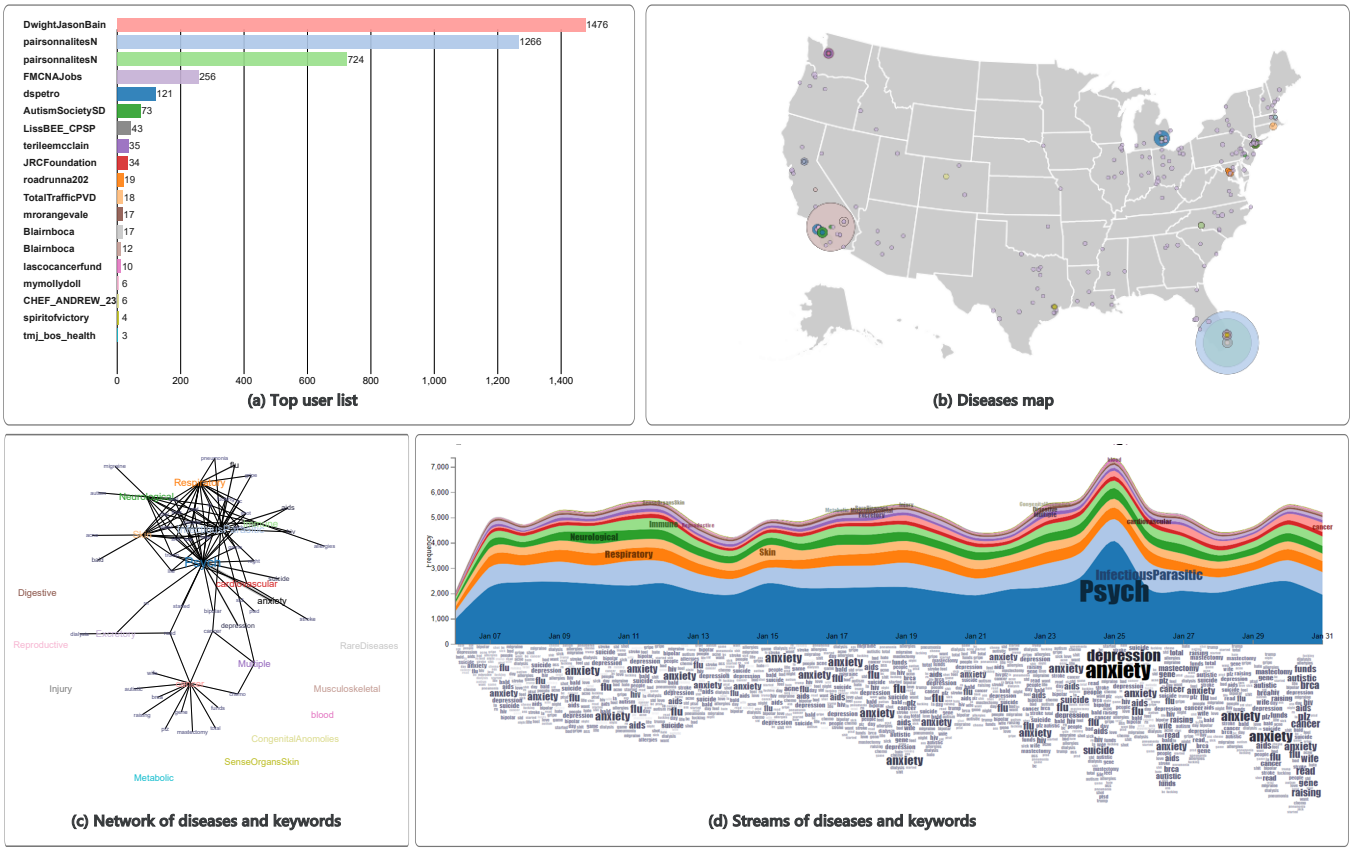


Fig. 1. The *HealthTvizer* system overview: a) Top user list based on the number of the tweets b) Disease map depicting the location of the posts c) Diseases cloud d) Interactive streamgraph representing the evolution of diseases over time.

pattern on the streamgraph is due to the fact that most people are sleeping after midnight. From disease cloud, viewers can quickly determine the diseases for ongoing discussions (which fulfills the visualization Task **T1**). The diseases are randomly assigned a color from a categorical color scale, and we use this color encoding consistently for the entire paper.

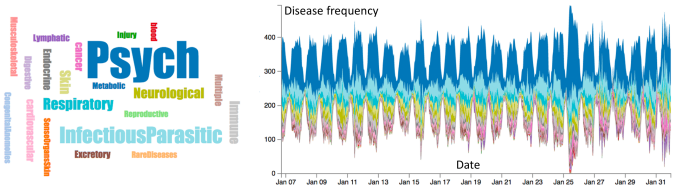


Fig. 2. Disease visualization in *HealthTvizer*: (a) Word cloud of most discussed diseases on Twitter in January 2017, (b) Streamgraph depicting the hourly frequency of discussed diseases.

Figure 3 shows the top 50 frequent keywords in the contents of tweets and their evolutions. When a user clicks on a particular disease in Figure 2, *HealthTvizer* filters the data and fetches the top 50 keywords regarding that selected disease (visualization task **T3**).

Figure 2 and Figure 3 can be combined into a single picture as shown in Figure 1(d). In particular, the upper portion shows daily discussions about diseases while in the lower portion,

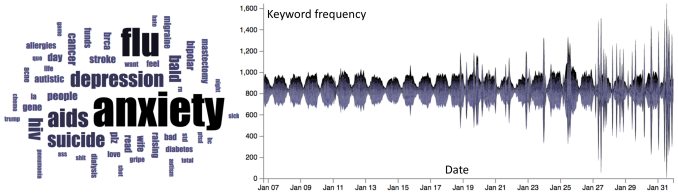
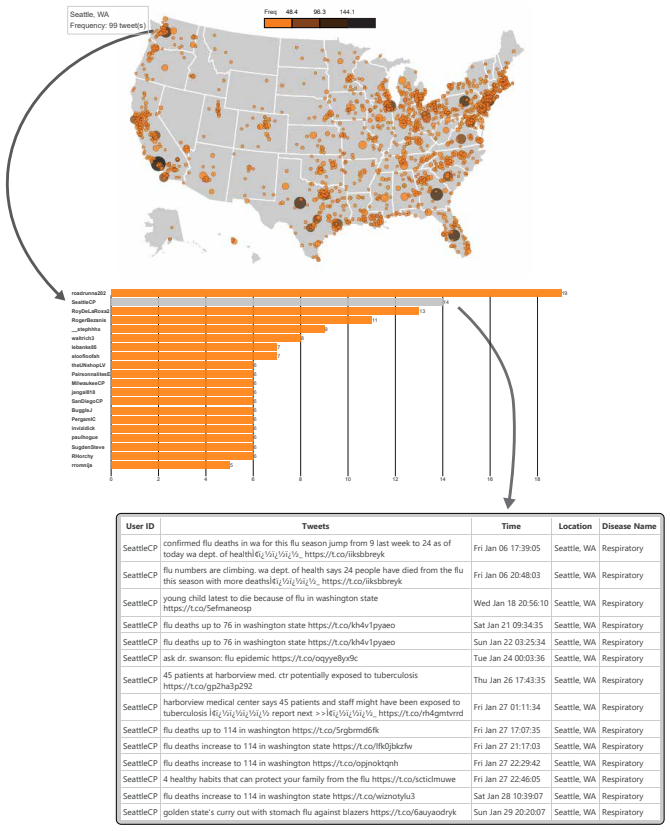


Fig. 3. Keywords visualization: (a) Most frequent words cloud about diseases in January 2017, (b) Streamgraph of hourly evolution of topics.

keywords are embedded directly into the frequency streams. We can easily notice that the peak of Psych in the upper portion is paired with highly frequent terms (*Anxiety* and despression) in the lower stream of words.

2) Disease Map: The spatial distribution of tweets of top users corresponding to 21 different diseases is visualized in a US map (Figure 1(b)) in which more popular location of health-related tweets (can be posted by different users at the same location) are highlighted in larger circles. The map also enables a viewer to detect sudden events in a geographical location as we use low-opacity circles. By selecting different diseases, users can visualize the distribution of discussions about them. For example, Figure: 4 summarizes the top 20 twitter users of "Respiratory" and their tweet geolocations.



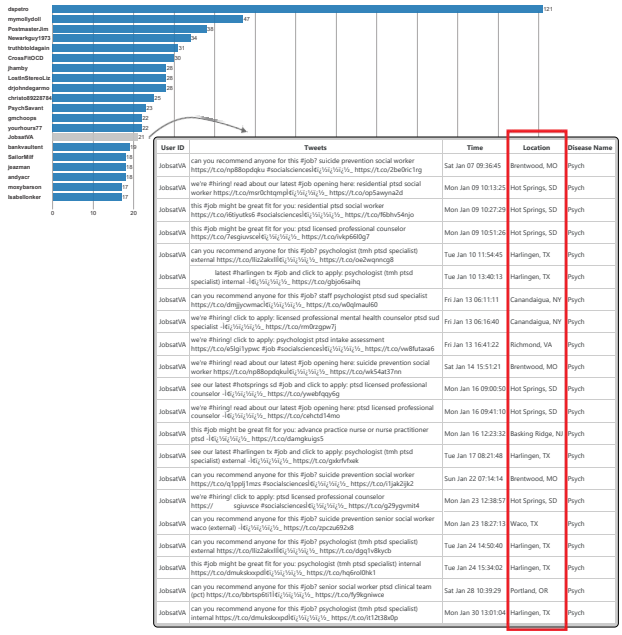
User ID	Tweets	Time	Location	Disease Name
SeattleCP	confirmed flu deaths in wa for this flu season jump from 9 last week to 24 as of today wa dept. of healthl¢ï¿½ï¿½ï¿½_ https://t.co/iiksbbreyk	Fri Jan 06 17:39:05	Seattle, WA	Respiratory
SeattleCP	flu numbers are climbing. wa dept. of health says 24 people have died from the flu this season with more deathsl¢ï¿½ï¿½ï¿½_ https://t.co/iiksbbreyk	Fri Jan 06 20:48:03	Seattle, WA	Respiratory
SeattleCP	young child latest to die because of flu in washington state https://t.co/Sefmaneosp	Wed Jan 18 20:56:10	Seattle, WA	Respiratory
SeattleCP	flu deaths up to 76 in washington state https://t.co/kh4v1pyaeo	Sat Jan 21 09:34:35	Seattle, WA	Respiratory
SeattleCP	flu deaths up to 76 in washington state https://t.co/kh4v1pyaeo	Sun Jan 22 03:25:34	Seattle, WA	Respiratory
SeattleCP	ask dr. swanson: flu epidemic https://t.co/oqyveByx9c	Tue Jan 24 00:03:36	Seattle, WA	Respiratory
SeattleCP	45 patients at harborview med. ctr potentially exposed to tuberculosis https://t.co/gp2ha3p292	Thu Jan 26 17:43:35	Seattle, WA	Respiratory
SeattleCP	harborview medical center says 45 patients and staff might have been exposed to tuberculosis l¢ï¿½ï¿½ï¿½ï¿½ report next > >l¢ï¿½ï¿½ï¿½ï¿½_ https://t.co/rh4gmtvrrd	Fri Jan 27 01:11:34	Seattle, WA	Respiratory
SeattleCP	flu deaths up to 114 in washington https://t.co/5rgbrmd6fk	Fri Jan 27 17:07:35	Seattle, WA	Respiratory
SeattleCP	flu deaths increase to 114 in washington state https://t.co/lfk0jbkzfw	Fri Jan 27 21:17:03	Seattle, WA	Respiratory
SeattleCP	flu deaths increase to 114 in washington https://t.co/opjnoktqnh	Fri Jan 27 22:29:42	Seattle, WA	Respiratory
SeattleCP	4 healthy habits that can protect your family from the flu https://t.co/scticlmuwe	Fri Jan 27 22:46:05	Seattle, WA	Respiratory
SeattleCP	flu deaths increase to 114 in washington state https://t.co/wiznotylu3	Sat Jan 28 10:39:07	Seattle, WA	Respiratory
SeattleCP	golden state's curry out with stomach flu against blazers https://t.co/6auyaodryk	Sun Jan 29 20:20:07	Seattle, WA	Respiratory

Fig. 4. Visualizing tweets about "Respiratory": (a) The tweeting locations of the top 20 users. (b) The tweets of user *SeattleCP* about "Respiratory" from Jan 06, 2017 to Jan 29, 2017.

3) User ranking chart: The top users tweeting about the specific disease are illustrated in Figure 1(a). Users are ranked by the number of tweets the users write or share for a specific disease (visualization task **T4**). When a user clicks on the user ID or the bar, *HealthTvizer* populates a table to show the actual tweets by that user including user ID, tweet content, time, location, and disease name (visualization task **T2**). An example is given in Figure 4 when we select "Respiratory" from the disease cloud. We can easily notice a larger circle in Seattle, WA regarding "Respiratory" as depicted in Figure 4(a). After examining the user ranking chart and user tweets as shown in Figure 4(b), we found that the tweets of the user SeattleCP confirm the observation about Seattle, WA location and provide details of the incident.

In another observation, we explore the tweets of the user *JobdstVA* about "Psych". As depicted in Figure 5, the different locations of the tweets indicate that *JobdstVA* moved a lot during January 2017 (research question **Q3**).

4) Network of Diseases and Keywords: Figure 6 shows the overview network of 20 diseases and top 50 popular keywords from January 6 to January 31, 2017, on Twitter. A link in this network connects a keyword with the classified disease of the tweet (visualization task **T5**). The thickness of a link indicates the co-occurrence frequency between them. Diseases are color



User ID	Tweets	Time	Location	Disease Name
JobsatVA	can you recommend anyone for this #job? suicide prevention social worker https://t.co/np88opdqku #socialsciences|¢ï¿½ï¿½ï¿½_ https://t.co/2be0ric1rg	Sat Jan 07 09:36:45	Brentwood, MO	Psych
JobsatVA	we're #hiring! read about our latest #job opening here: residential ptsd social worker https://t.co/msr0chtqmpl¢ï¿½ï¿½ï¿½_ https://t.co/op5awynw2d	Mon Jan 09 10:13:25	Hot Springs, SD	Psych
JobsatVA	this #job might be great fit for you: residential ptsd social worker https://t.co/i96lyudlo6 #socialsciences|¢ï¿½ï¿½ï¿½_ https://t.co/f82hvs54njo	Mon Jan 09 10:27:29	Hot Springs, SD	Psych
JobsatVA	this #job might be great fit for you: ptsd licensed professional counselor https://t.co/7esgiuvsce¢ï¿½ï¿½ï¿½_ https://t.co/ivkp660g7	Mon Jan 09 10:51:26	Hot Springs, SD	Psych
JobsatVA	can you recommend anyone for this #job? psychologist (tmh ptsd specialist) external https://t.co/lli2akxlll¢ï¿½ï¿½ï¿½_ https://t.co/oa2wqnncg8	Tue Jan 10 11:54:45	Harlingen, TX	Psych
JobsatVA	latest #harlingen tx #job and click to apply: psychologist (tmh ptsd specialist) internal -l¢ï¿½ï¿½ï¿½_ https://t.co/gbjo64aihq	Tue Jan 10 13:40:13	Harlingen, TX	Psych
JobsatVA	can you recommend anyone for this #job? staff psychologist ptsd sud specialist https://t.co/dmjjycwmac¢ï¿½ï¿½ï¿½_ https://t.co/w0qlmaul60	Fri Jan 13 06:11:11	Canandaigua, NY	Psych
JobsatVA	we're #hiring! click to apply: licensed professional mental health counselor ptsd sud specialist -l¢ï¿½ï¿½ï¿½_ https://t.co/rm0rzgqw7j	Fri Jan 13 06:16:40	Canandaigua, NY	Psych
JobsatVA	we're #hiring! click to apply: psychologist ptsd intake assessment https://t.co/e5lgi1ypwc #job #socialsciences|¢ï¿½ï¿½ï¿½_ https://t.co/ww8futaxa6	Fri Jan 13 16:41:22	Richmond, VA	Psych
JobsatVA	we're #hiring! read about our latest #job opening here: suicide prevention social worker https://t.co/np88opdqkul¢ï¿½ï¿½ï¿½_ https://t.co/wk54at37nn	Sat Jan 14 15:51:21	Brentwood, MO	Psych
JobsatVA	see our latest #hotsprings sd #job and click to apply: ptsd licensed professional counselor -l¢ï¿½ï¿½ï¿½_ https://t.co/ywe8dqqy6g	Mon Jan 16 09:00:50	Hot Springs, SD	Psych
JobsatVA	we're #hiring! read about our latest #job opening here: ptsd licensed professional counselor -l¢ï¿½ï¿½ï¿½_ https://t.co/cehctd14mo	Mon Jan 16 09:41:10	Hot Springs, SD	Psych
JobsatVA	this #job might be great fit for you: advance practice nurse or nurse practitioner ptsd -l¢ï¿½ï¿½ï¿½_ https://t.co/damgkuips5	Mon Jan 16 12:23:32	Basking Ridge, NJ	Psych
JobsatVA	see our latest #harlingen tx #job and click to apply: psychologist (tmh ptsd specialist) external -l¢ï¿½ï¿½ï¿½_ https://t.co/gxkrfvfwek	Tue Jan 17 08:21:48	Harlingen, TX	Psych
JobsatVA	can you recommend anyone for this #job? suicide prevention social worker https://t.co/q1ppbj1mzs #socialsciences|¢ï¿½ï¿½ï¿½_ https://t.co/i1jakZjjk2	Sun Jan 22 07:14:14	Brentwood, MO	Psych
JobsatVA	we're #hiring! click to apply: ptsd licensed professional counselor https:// sgiuvsce #socialsciences|¢ï¿½ï¿½ï¿½_ https://t.co/g29ygvmit4	Mon Jan 23 12:38:57	Hot Springs, SD	Psych
JobsatVA	can you recommend anyone for this #job? suicide prevention senior social worker waco (external) -l¢ï¿½ï¿½ï¿½_ https://t.co/zpczu692x8	Mon Jan 23 18:27:13	Waco, TX	Psych
JobsatVA	can you recommend anyone for this #job? psychologist (tmh ptsd specialist) external https://t.co/lli2akxlll¢ï¿½ï¿½ï¿½_ https://t.co/dgyv8kyzb	Tue Jan 24 14:50:40	Harlingen, TX	Psych
JobsatVA	this #job might be great fit for you: psychologist (tmh ptsd specialist) internal https://t.co/dmukskxpdl¢ï¿½ï¿½ï¿½_ https://t.co/hq6rol0hk1	Tue Jan 24 15:34:02	Harlingen, TX	Psych
JobsatVA	can you recommend anyone for this #job? senior social worker ptsd clinical team (pct) https://t.co/bbrtsp661l¢ï¿½ï¿½ï¿½_ https://t.co/fyjfkgniwca	Sat Jan 28 10:39:29	Portland, OR	Psych
JobsatVA	can you recommend anyone for this #job? psychologist (tmh ptsd specialist) internal https://t.co/dmukskxpdl¢ï¿½ï¿½ï¿½_ https://t.co/it123lb6lp	Mon Jan 30 13:01:04	Harlingen, TX	Psych

Fig. 5. Displaying the tweets of *JobdstVA* related to "Psych" from Jan 07, 2017 to Jan 28, 2017 along with the geo-location of the user at the time of tweets.

coded while keywords are in black.

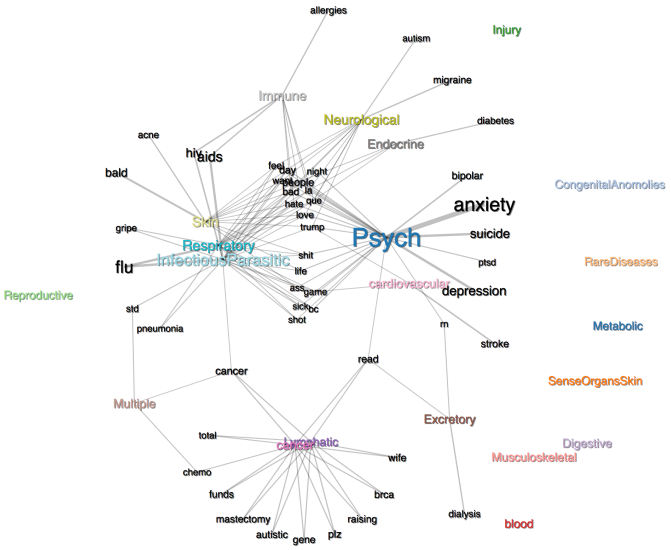


Fig. 6. Network of diseases and keywords from Jan 07 to Jan 28, 2017.

A few observations about this network:

- We can visually detect the disease clusters, such as "Respiratory", "Skin", and "Infectious" vs. ("Cancer" and "Lymphatic".
- "Psych" disease has strong connection with "Anxiety" term
- Some disease are highly connected while others are isolated and do not have any connections.
- We can see that "Psych" (mental) disease has most connections with other diseases so we can talk about

the side effects of other diseases to people with mental condition and importance of taking care of it.

C. Implementation

HealthTvizer is implemented in D3.js [48]. The online application, demo video, source code, and sample data are provided via our GitHub project repository, located at https://healthtvizer.github.io/.

IV. Evaluation and Discussion

A. Use cases

In this section, we will demonstrate some stories that we found with *HealthTvizer* on nearly 137 thousands processed tweets from Jan 07, 2017 to Jan 28, 2017.

1) Use case on "Cancer": After clicking on "Cancer" from the diseases cloud, we notice a big circle in California as depicted in Figure 7. We further investigate the incident by examining the top user *DwightJasonBain* from the user ranking chart. We come to learn that during January of 2017 the user ID *DwightJasonBain* started a fundraising organization for his wife (who is suffering from breast cancer).

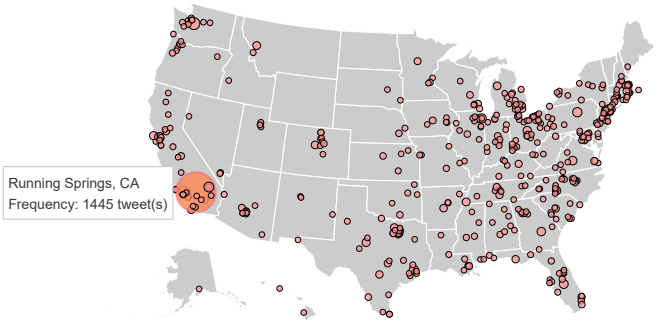


Fig. 7. Visualizing tweets about "Cancer" during January of 2017 in our *HealthTvizer*.

The contents of tweets in Figure 7 (not showing here) are identical: "Please retweet: I'm autistic raising funds, wife needs BRCA gene cancer, total mastectomy surgery/recovery" [49]. Using gofundme [50], *DwightJasonBain* tries to gather fund for the treatment of his wife. As this person lives in Running Springs, CA and people in that area retweets the appeal for raising fund, the tweets became spread out, causing a huge number of tweets in that area.

2) Use case on "InfectiousParasitic": "InfectiousParasitic" consists several diseases such as *Ebola*, *Cholera*, *Flu*, *Zika*, etc. However, the keywords cloud only shows *AIDS* for January 2017. We also found that some people were tweeting about *Flu* at that same time in California (research question **Q2**).

Our investigation discovers that on January 6, a link was shared by the Twitter users *pairsonnalitesN*, *pairsonnalitesE*, and *pairsonnalites* which are the accounts of the same organization called *Stigmabase*. *Stigmabase* works on the awareness of various diseases, especially on *AIDS*. They shared the tweets about the work of Merck [51], and it acquired plenty of retweets. This phenomenon is the reason for the big circle in Miami, FL as depicted in the lower Panel of Figure 8, as the Twitter user IDs are located there.

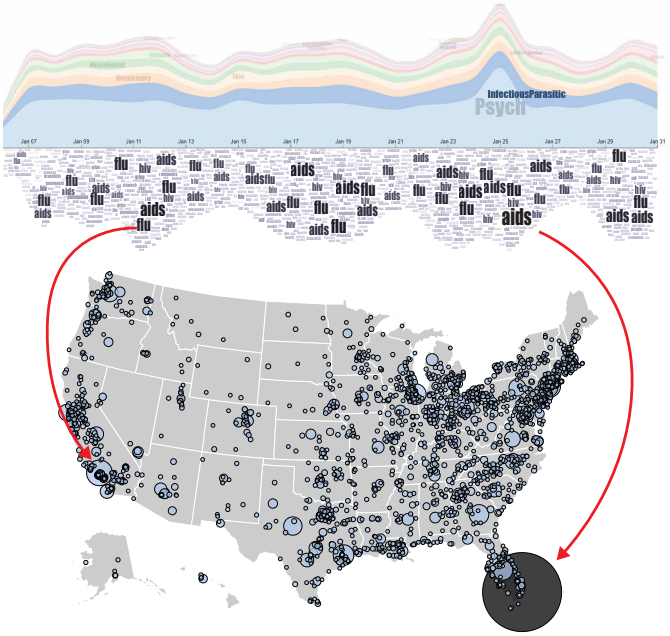


Fig. 8. The sequence of exploring "InfectiousParasitic": (top) Selection of "InfectiousParasitic" disease and its stream of keywords, (bottom) The distribution of the tweets on the United States map.

Another discovery is about the "InfectiousParasitic" and "Respiratory" sharing the same hot keyword *Flu*. On January 27th in Seattle, WA, a news published on several websites (e.g. Washington Times [52]) stating that the number of death because of *Flu* increases to 114 in Washington state. There was plenty of tweets related to this issue was tweeted especially in Seattle, WA. that we can infer that these types of reports create awareness and fear among the people. It can see by a comparatively large circle on the US map for "Respirator" and "Infectious Parasitic".

3) Use case on "Injury": The streamgraph of the terms related to the "Injury" disease on the January 17 Figure 9(b) has a significant fluctuation that shows the term *tbi* is used a lot. We can also see a bigger circle on the map as depicted in Figure 9(a). Our exploring shows us that on January 17th the Fox News Insider [53] published a news related to the list of Democrat who won't attend Trumps Inauguration. One the Twitter user with the ID *LyndaRe06304657* started to tweet about the reason. The user referred that John Lewis, the Democrat congressman, had a brain injury, so those Democrats need to take care of it. This observation the large circle (highlighted on the map) and the most common term *tbi* fluctuations in the bottom left word cloud. In this example, the user *LyndaRe06304657* tried to clarify her issues by using social media.

4) Use case on "Autism": Figure 10 shows the linked views for "Neurological" diseases. In the keyword stream,

Fig. 9. Visualizing tweets about "tbi" on January 17th 2017 in our *HealthTvizer*: (a) Geolocations of tweets about "Injury" b) The most frequent terms *tbi* that refers to some issues related to Trump's Inauguration.

"Autism" stands out on January 10. Examining the user's tweets from the user list, we found that the user *AutismSocietySD* share lots of tweets about the "7th annual swimming with autism conference" in San Marcos, CA. That explains a high frequency of "Autism" on January 10, 2017.

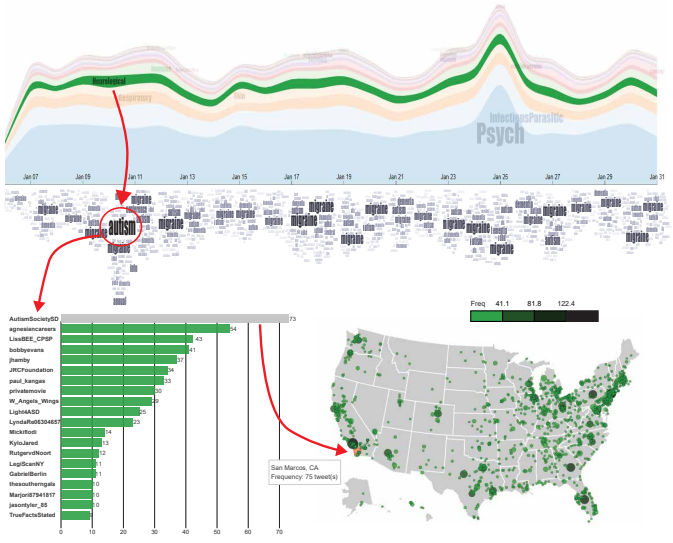


Fig. 10. Visualizing tweets about "Neurological" diseases.

B. Informal User study

We solicited qualitative responses about *HealthTvizer* from two students and one associate professor at Health Science Center at a public university. The students are the second year in medical school, and the professor has twenty years of experience in the public health domain. The informal study started with a quick description (around 10 minutes) of the main components and their functionalities to familiarize users to the main GUI of *HealthTvizer*. Then the experts are free to use *HealthTvizer* before providing feedback. All they have a unanimous idea about the US map that can serve as the starting point for public health, epidemiology, and research perspective.

- They all agree that this approach and the provided tool are interesting for them to convey when and what people are talking about for certain types of diseases. They did not think about social media could be used for this purpose before.
- One of the students was interested in inspecting the mental health and its frequency. He concluded that there is a huge need for these interactive and real-time visualizations for mental health providers in this country. He also mentioned that we could make the argument if more people are talking about something in a certain area then the government needs to devote more resources and attention to that disease or condition in that area (research question **Q5**).
- Moreover, all three domain experts agree that *HealthTvizer* could be useful for doctors who are just starting to decide where to start their careers. Specifically, when people in a specific city discuss lots a disease, they are probably in need of health care services for that condition/disease.

Besides the positive feedback, the experts also pointed out some limitations of *HealthTvizer*. For example, they had some difficulty in reading the sideways words. One of them suggested having a connection between terms of one user of different diseases together. Based on this suggestion, we have implemented a force-directed layout of top diseases and keywords which are presented in Section III-B4.

The domain experts also suggested applying multiple filters. For example, they would like to select "Cancer" and then a user ID from the top user chart, the keyword cloud, as well as the frequency of these keywords over time (and the map), should be updated accordingly. That has been included in the current prototype.

V. Conclusion

In this paper, we present a novel approach that unites coordinated multiple views to highlight various attributes of the tweets and Twitters users. We introduce an interactive data analytic prototype to help viewers to summarize the current diseases along with the specific opinions and the discussions on those diseases. The interactive analytic tool coordinates the tweets with the geolocation and depicts those on a US map which helps the users to get a quick idea about the diseases and related incidents at different locations. We demonstrate the usefulness of our tool through different use cases. We also conduct an informal study with domain experts and receive interesting feedback as well as possible future extensions.

In future work, we are planning to provide an intuitive interface to relate the different tweets of various users to achieve a better picture of diseases and current hot topics [54]. In this paper, we demonstrate the *HealthTvizer* only for the tweets on January 2017. We want to apply the tools for more data of different period and explore the evolution and relationship of the diseases with respect to the geographical location over time.

REFERENCES

[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, no. 1, pp. 178–185, 2010.

[2] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on twitter," *Machine Learning and Knowledge Discovery in Databases*, pp. 599–602, 2010.

[3] M. Qiu, F. Zhu, and J. Jiang, "It is not just what we say, but how we say them: Lda-based behavior-topic model," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 794–802.

[4] A. E. Chung, A. C. Skinner, S. E. Hasty, and E. M. Perrin, "Tweeting to health: a novel mhealth intervention using fitbits and twitter to foster healthy lifestyles," *Clinical Pediatrics*, vol. 56, no. 1, pp. 26–32, 2017.

[5] L. Lowe, "Socialpilot," https://socialpilot.co/blog/125-amazing-social-media-statistics-know-2016/.

[6] J. C. Roberts, "State of the art: Coordinated multiple views in exploratory visualization," in *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*, July 2007, pp. 61–71.

[7] T. N. Dang and L. Wilkinson, "Scagexplorer: Exploring scatterplots by their scagnostics," in *2014 IEEE Pacific Visualization Symposium*, March 2014, pp. 73–80.

[8] L. Wilkinson, A. Anand, and R. Grossman, "High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1363–1372, 2006.

[9] N. Elmqvist, P. Dragicevic, and J. D. Fekete, "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1539–1148, Nov 2008.

[10] A. Dasgupta and R. Kosara, "Pargnostics: Screen-space metrics for parallel coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 1017–2626, 2010.

[11] X. Zhao and A. Kaufman, "Structure revealing techniques based on parallel coordinates plot," *Vis. Comput.*, vol. 28, no. 6-8, pp. 541–551, Jun. 2012. [Online]. Available: http://dx.doi.org/10.1007/s00371-012-0713-0

[12] M. Dork, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1129–1138, 2010.

[13] T. N. Dang, A. Anand, and L. Wilkinson, "TimeSeer: Scagnostics for high-dimensional time series," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 3, pp. 470–483, March 2013.

[14] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: Supporting investigative analysis through interactive visualization," *Information Visualization*, vol. 7, no. 2, pp. 118–132, Apr. 2008. [Online]. Available: http://dx.doi.org/10.1145/1466620.1466622

[15] Tableau Software, "Tableau," http://www.tableausoftware.com.

[16] J. Y. E. A. Rundensteiner, M. O. Ward and P. R. Doshi, "Xmdv-Tool: visual interactive data exploration and trend discovery of high dimensional data sets," in *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*. ACM, 2002, pp. 631–631.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[18] G. Doyle and C. Elkan, "Financial topic models," in *Working Notes of the NIPS-2009 Workshop on Applications for Topic Models: Text and Beyond Workshop*, 2009.

[19] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *Advances in neural information processing systems*, 2008, pp. 1577–1584.

[20] C. Zou and D. Hou, "Lda analyzer: A tool for exploring topic models," in *Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on*. IEEE, 2014, pp. 593–596.

[21] J. Boyd-Graber and P. Resnik, "Holistic sentiment analysis across languages: Multilingual supervised latent dirichlet allocation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 45–55.

[22] Y.-S. Chen, L.-H. Chen, and Y. Takama, "Proposal of lda-based sentiment visualization of hotel reviews," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, 2015, pp. 687–693.

[23] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 937–946.

[24] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine learning*, vol. 95, no. 3, pp. 423–469, 2014.

[25] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on twitter," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013, p. 8.

[26] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European Conference on Information Retrieval*. Springer, 2011, pp. 338–349.

[27] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang, "Tiara: a visual exploratory text analytic system," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 153–162.

[28] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*. IEEE, 2011, pp. 231–240.

[29] "Healthcare Hashtag Project, howpublished = https://www.com/healthcare-social-media-research, note = Accessed: 2017-05."

[30] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 2010, pp. 36–44.

[31] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks." *ICWSM*, vol. 10, pp. 90–97, 2010.

[32] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 2010, pp. 115–122.

[33] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. Zhu, and H. Qu, "Visual analysis of topic competition on social media," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2012–2021, 2013.

[34] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.

[35] M. Dork, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *IEEE transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1129–1138, 2010.

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

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

[38] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: A systematic review," *American journal of public health*, vol. 107, no. 1, pp. e1–e8, 2017.

[39] A. K. Bamwal, G. K. Choudhary, R. Swamim, A. Kedia, S. Goswami, and A. K. Das, "Application of twitter in health care sector for india," in *Recent Advances in Information Technology (RAIT), 2016 3rd International Conference on*. IEEE, 2016, pp. 172–176.

[40] S. Xu, C. Markson, K. L. Costello, C. Y. Xing, K. Demissie, and A. A. Llanos, "Leveraging social media to promote public health knowledge: example of cancer awareness via twitter," *JMIR public health and surveillance*, vol. 2, no. 1, 2016.
[41] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health." *Icwsm*, vol. 20, pp. 265–272, 2011.
[42] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media." in *ICWSM*, 2013, p. 2.
[43] M. Dredze, "How social media will change public health," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 81–84, 2012.
[44] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of the 1996 IEEE Symposium on Visual Languages*, ser. VL '96. Washington, DC, USA: IEEE Computer Society, 1996, pp. 336–. [Online]. Available: http://dl.acm.org/citation.cfm?id=832277.834354
[45] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *Proc. of the IEEE Symposium on Information Visualization*, 2005, pp. 15–24.
[46] D. A. Keim, C. Panse, and M. Sips, "Information visualization : Scope, techniques and opportunities for geovisualization," in *Exploring Geovisualization*, J. Dykes, Ed. Oxford: Elsevier, 2004, pp. 1–17.
[47] N. Andrienko, G. Andrienko, and P. Gatalsky, "Exploratory spatio-temporal visualization: an analytical review," *Journal of Visual Languages & Computing*, vol. 14, no. 6, pp. 503–541, 2003.
[48] M. Bostock, V. Ogievetsky, and J. Heer, "D3 data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.
[49] D. J. Bain, "Fund raising tweet," https://twitter.com/DwightJasonBain/status/822992475488821248.
[50] ——, "Go fund me," https://www.gofundme.com/brca-oophorectomy-surgery-help.
[51] "Merck," http://dlvr.it/N3D7c8.
[52] "Washington times," http://www.washingtontimes.com/news/2017/jan/27/flu-deaths-increase-to-114-in-washington-state/.
[53] "Fox news insider," http://insider.foxnews.com/2017/01/17/heres-list-democrats-who-wont-attend-trumps-inauguration.
[54] T. N. Dang, N. Pendar, and A. G. Forbes, "TimeArcs: Visualizing Fluctuations in Dynamic Networks," *Computer Graphics Forum*, 2016.