Real-Time Digital Flu Surveillance using Twitter Data

Kathy Lee* Ankit Agrawal* Alok Choudhary*

Abstract

Social media is producing massive amounts of data at an unprecedented scale, where people share their experiences and opinions on a variety of different things, including healthcare-related topics, like health conditions, their symptoms, treatments, side-effects, and so on. This makes the publicly available social media data an invaluable resource for mining such data to discover interesting and actionable healthcare insights. In this paper, we describe an online resource for real-time surveillance of flu that we have developed using spatial, temporal, and text mining on Twitter data. The real-time analysis results are subsequently reported visually in terms of a US flu surveillance map, distribution and timelines of flu types, flu symptoms, and flu treatments, in addition to overall flu activity timeline. Such a surveillance system can be very useful for early prediction of flu outbreaks, which in turn can facilitate faster and better response preparation. Further, the resulting insights are also expected to be very useful for both patients and doctors to make informed decisions.

1 Introduction

The Internet is usually the first place people turn for health information. People most often search for a specific disease or medical problems, and appropriate medical treatments or procedures. Health-care portal sites and the social media are two of the most popular online health information resources among the U.S. Internet users [9].

Disease surveillance is the monitoring of clinical syndromes such as flu, cancer, allergies, diabetes and many others that have a significant impact on medical resource allocation, health policy and education. The main role of disease surveillance is not only to minimize the harm caused by the outbreaks by observing and predicting disease spread, but also to understand what factors contribute to such circumstances.

The traditional approach employed by the Centers for Disease Control and Prevention (CDC) [7] for flu surveillance includes the collection of Influenza-like Illness (ILI) patients' data from sentinel medical practices. However, it takes time to collect and process data, and there is usually 1-2 weeks time lag before the data becomes available. Furthermore, the flu estimates are updated only once a week. Early detection of a disease outbreak is critical because it would allow faster communication between health agencies and the public and provide more time to prepare a response.

Twitter[1] is a popular micro-blogging service where users can post short messages limited to 140 characters. Twitter has been used as a medium for real-time information dissemination and it has been used in various brand campaigns, elections, and as a news media. Because Twitter data can be collected in real-time, it has been used to predict real world outcomes. Since its launch in 2006, the popularity of its use has increased dramatically. One website [16] reported that, as of November 2012, Twitter has approximately 500 million registered users. Although a very high volume of twitter stream contains general chatter, it does contain enough health-related information to track disease spread.

In this paper, we built a novel flu surveillance system that uses twitter data to track U.S. influenza activities in real-time. The proposed system consists of four stages: Data Collection, Data Preprocessing, Data Modeling, and Data Visualization. The data collecting module continuously downloads flu-related public twitter data using Twitter streaming API [6]. The preprocessor module extracts tweet texts, time stamps, and user locations and stores them in a database for further analysis. There are three models in the data modeling stage: geographical model, text model, and temporal model. In geographical modeling, we detect U.S. regional flu activity levels. In text modeling, we track and compare popularity of different flu types, symptoms, and treatments. Our temporal model tracks daily flu activity level and popularity of flu-related terms over time. In the final data visualization stage, the flu activity level of each U.S. state (the output of geographical model) is presented as an interactive map with different shades of colors, where darker colors indicate higher level of flu activity. The popularity of different flu types, symptoms, and treatments (the output of the text model) are presented as bar charts for easy visualization and

*Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL USA. E-mail: {kml649, ankitag, choudhar}.eecs.northwestern.edu. This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456.

[1] https://twitter.com/

comparison. Lastly, the daily tweet volume changes of flu-related terms (the output of the temporal model) are visualized as time series graphs. All data processing steps from collecting, preprocessing, modeling, to visualizing the data, are automated in our proposed system.

2 Related Work

Many researchers have used Twitter data to predict various real world outcomes, to monitor earthquakes in Japan [17], forecast box-office revenues for movies [12], and infer U.S. mood throughout the day [5].

There are several existing influenza surveillance platforms. FluTrackers.com [3] has been tracking infectious diseases since 2006. It is a nonprofit corporation composed of citizens from many countries who document emerging diseases. Flusurvey [2] is an online system that measures flu trends in the UK. In contrast to CDC, they collect data directly from the general public, rather than via hospitals. Each week, registered users report any flu-like symptoms they have experienced.

A number of recent papers have addressed the use of online data for disease detection (digital disease detection). Google Flu Trends [8] uses google online search queries related to flu-like symptoms to track influenza activities. It often accurately predicts the flu cases weeks in advance of CDC records. Pervaiz et a. [15] built an early epidemic detection system using the Google Flu Trends data. Achrekar et al. [11] devised an auto-regression model for prediction and uses Twitter data to predict Flu trends and showed that their model can accurately predict flu epidemics when past CDC data is used together with Twitter data. Signorini et al. [18] applied support vector regression method on Twitter data to track swine flu activities and examined disease transmission in particular social contexts, disease counter measures, and consumer concerns about pork consumption. Paul and Dredze [14] applied their Ailment Topic Aspect Model to Twitter data to track public health over time, measuring behavioral risk factors, localizing illness by geographic region, and analyzing symptoms and medication usage. Our work is different from all prior works in that we not only track the U.S. regional and temporal flu activity, but also track the popularity of the major flu-related terms such as flu types, symptoms, and treatments over time. Furthermore, every data processing step in our system is automated and the most current data is updated on our project website in near real-time. Real time spatio-temporal tracking of diseases is critical to prevent the epidemic.

3 Data and Methods

As shown in Figure 1, the proposed system consists of four stages: Data Collection, Data Preprocessing, Data Modeling, and Data Visualization. In the data collection stage, the data collector module continuously downloads flu-related public twitter data using Twitter streaming API [6]. The data preprocessor module extracts tweet texts, time stamps, and user locations and stores them in a database for further analysis. In the modeling stage, we inspect the data by looking at three different aspects of the tweet data: (1) Geographical/Spatial Modeling, (2) Text Modeling, and (3) Temporal Modeling. In the Data Visualization stage, the output data from the Geographical Model is presented as U.S. Flu Activity Map, the output data from the Text Model is presented as Flu Types, Flu Symptoms, and Flu Treatments bar charts, and the output data from the Temporal Model is presented as a Flu Activity Timeline.

3.1 Data Collection. The data collector module continuously downloads the tweet data using Twitter Streaming API and stores the raw data into the MySQL database. Twitter Streaming API [6] allows high-throughput near real-time access to global stream of public tweets that matches pre-specified filter predicates. Multiple parameters may be specified in a single connection to the streaming API to determine what tweets will be delivered on the stream.

Our dataset consists of over 2 million public tweets that contain the keyword 'flu' (called 'flu tweets' hereafter) generated by over 1.3 million unique users since October 16, 2012.

3.2 Data Preprocessing. A tweet has tweet text, user name, time-stamp, and location. The data preprocessor module reads the raw tweet data from the database, extracts the tweet text, tweet timestamp and user location from the tweet JSON, and stores the data back into the database.

Tweet text is a short text message limited by 140 characters in length posted by users on Twitter. People often post messages about interesting news, their daily activities, thoughts, feelings, as well as health conditions. Patients suffering from various diseases talk on Twitter about their true health conditions, their feelings about the symptoms, and treatments they take to relieve the symptoms, some of which they may not discuss even with their medical doctors. Twitter is a great data resource to monitor Influenza-like Illness activity and flu symptoms because most patients with mild flu symptoms tend to take over-the-counter flu medicines and do not bother to go see their doctors. Table 1 shows examples of tweets mentioning keyword 'flu'. Many users describe their flu symptoms (sore throat, cough, headache, can't even sleep, high temp) and feelings ("just want to

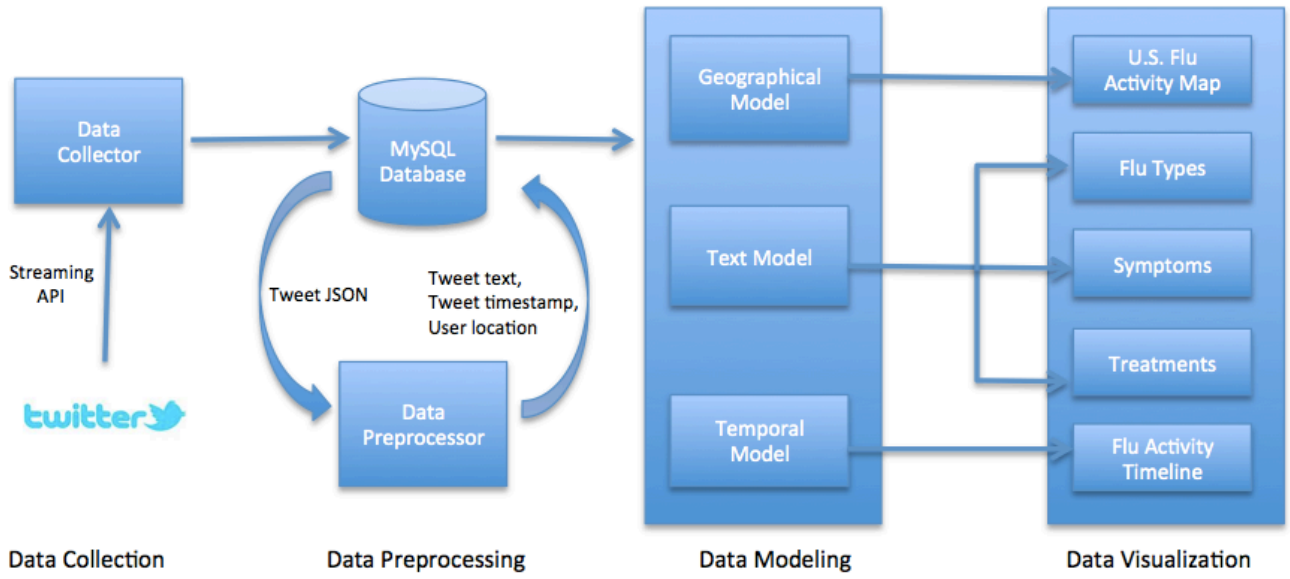

Figure 1: Real-Time Flu Surveillance System Architecture. The system consists of four stages: (1) Data Collection stage - continuously downloads public tweets mentioning 'flu'; (2) Data Preprocessing stage - extracts tweet text, tweet timestamp and user location; (3) Data Modeling stage - apply spatial, temporal and text mining; (4) Data Visualization stage - visually report real-time analysis results.

get better") in their tweets. One user reports having fever from flu even after having had the flu shot. Another user expresses his/her frustration ("I don't even know what to do with myself anymore").

Table 1: Example of flu related tweets

Flu + sore throat -_-
Pretty sure I have the flu so no school for me
sore throat + flu + cough + headache = can I have kris for tonight to take care of me
Im so sick that I dont even know what to do with myself anymore, I can't even sleep. :(I just want to get better #cough #fever #flu
And yet, the flu, and headache. So feeling holidays spent in bed. - -.
My high temp was 103.6 and I too had the flu shot!!

Tweet timestamp is a string that indicates the time of tweet generation (e.g., 'Sat Dec 29 01:37:25 +0000 2012') and is tagged to each tweet. '+0000' indicates that the tweet time is in Greenwich Mean Time (GMT).

User location is a home location a user enters into his/her Twitter profile. Note that we use the user location information only to identify the flu related tweet volume for each U.S. state and it is not used for any other purposes.

3.3 Data Modeling. In the data modeling stage, there are three models: a geographical model, a text model, and a temporal model.

3.3.1 Geographical Model. Our goal for the geographical modeling is to track the spread of Influenza-like Illness by geographic region by measuring the volume of flu tweets generated in each U.S. state.

A tweet or a user can have two types of location, a text-based user profile location or a sensor-based geo-location. User profile location is a user-entered random text string he/she declares as his/her home location. Sensor-based tweet location is a actual geo-location (with longitude and latitude values) of a user provided with a tweet. This information is available only if a twitter user selects his/her location to be visible to public in his/her twitter profile. Because there are very few tweets that have actual geo-location information, we chose to use the location information in users' twitter profile for our experiments.

For all users who generated flu tweets in our dataset, we identify the tweets with valid U.S. state names in the user profile location strings, and tag the tweet location with two character U.S. state codes. Because we are interested in tracking flu tweet volumes within U.S., we ignore the ones generated from outside of the states (tweets with foreign location) and the ones with invalid

location information.

We were able to identify 182,724 users with valid US state information which is approximately 14.1% of the total 1.3 million users in our dataset. 29.5% of the users did not have any textual information in their profile location at all. The rest of the users were either from foreign countries or did not have U.S. state information in the location field.

Table 2 shows examples of user locations with valid U.S. state names. The second column indicates the state codes that were tagged for the locations in the first column. Table 3 shows random texts, often humorous, in user locations. One user has 'Infinity & beyond', a well known phrase from the animation movie 'Toy Story', and another user has 'Home Tweet Home' as their home location.

The high percentage of users with no location information or random text could be an indication of users' concerns with the privacy issues of identifying home locations in their public OSN(Online Social Networks) profile.

Table 2: Examples of Valid User Location

Valid Location	State
Riverside, CA	CA
somewhere in NY	NY
Evanston, IL	IL
Miami, FL	FL
TCU The Fort, Texasssssss	TX
gainesville, florida	FL

Table 3: Examples of Invalid User Location

Invalid Location
On my way to Paradise
On A Never Ending Paper Chase
JESUS HEART
traveling
Where The Wild Things Are
Home Tweet Home
His Heart
Infinity & Beyond
Somewhere over the rainbow
bottom of a wine glass
Floating with the stars!
On my way to success...
Wherever the wind blows me....

3.3.2 Text Model. People not only talk about news and daily activities but also freely express their thoughts and feelings on social networks. With its sheer volume and real-time message propagation, Twitter data has become an extremely rich source of information.

The goal of text modeling is to discover useful health information from the tweet texts. We are interested in investigating the following three categories : (1) Flu types (2) Flu Symptoms (3) Flu Treatments. For each category, we created a keyword list. For example, the keyword list for flu types is a list of swine flu, bat flu, H1N1, H5N1, etc., the keyword list flu symptoms is a list of fever, cough, sore throat, headache, etc., and the keyword list for flu treatments is a list of tamiflu, theraflu, tylenol, advil, vitamic c, etc. These keyword lists are used to track the popularity (frequency) of the words in the flu tweets. For each word in three keyword lists, we count the number of tweets that contain both the keyword and the word 'flu'. These numbers are used to create correlation bar charts in the data visualization stage.

3.3.3 Temporal Model. The goal of temporal modeling is to track the volume changes of the flu tweets over time. Our assumption is that people talk more about 'flu' when people around them (friends, family, co-worker, etc.) or they themselves have the disease. The word 'flu' will appear more frequently on Twitter news feeds when the outbreak is wide spread than when it is sporadic. Achrekar et al. [11] reported that Twitter data provides real-time assessment of flu activities and the volume of flu related tweets is highly correlated with the number of reported ILI cases by the CDC. Therefore the volume change of keyword 'flu' over time is a good reflection of the flu activity level change over time. For the temporal model, we count the number of flu related tweets generated daily. This data is used to create the flu activity level timeline in the data visualization stage.

We also track the daily tweet volume of each keyword in the three keyword lists (flu types, flu symptoms, flu treatments) described in section 3.3.2.

3.4 Data Visualization. The goal of Data Visualization stage is to convey the output of our Geographical, Text, and Temporal Models to the end users in very easy-to-understand graphic formats. We deployed a project website [1] to present our near real-time flu detection from each modeling stage. All data shown on the website is updated near real-time.

The output of the geographical model described in section 3.3.1 is mapped onto an interactive U.S. map. The volume of tweets generated from each state is differentiated by different intensity of shades. The darker regions indicate higher volume of tweets, and the lighter regions present lower volume of tweets from that state. If a user mouses over a state on the map, he/she

can view the state name and the percentage of the flu tweets generated from that state. All charts (U.S. flu activity map, bar charts, and the timeline charts) on our project website [1] are created using Google Chart Tools [4].

4 Results

In this section, we present and discuss the output of our geographical, text, and temporal models. The figures presented in section 4 are available in real-time on our *Digital Flu Surveillance* project website [1].

4.1 Geographical Analysis. The dataset for geographical analysis is all flu related tweets that have valid U.S. state names in the user location as described in section 3.3.1.

Figure 2 displays the percentage of the flu activity level for each U.S. state. Darker regions indicate that higher number of tweets mentioning the keyword 'flu' were generated from the state. New York state shows the highest percentage of flu tweets (9.6%), followed by California (8.48%) and Texas (8.15%). Wyoming had the least percentage of flu tweets (0.11%). The percentage of state s is calculated by dividing the number of tweets that has state s in user location by the total number of tweets that has an identifiable U.S. state in the user profile.

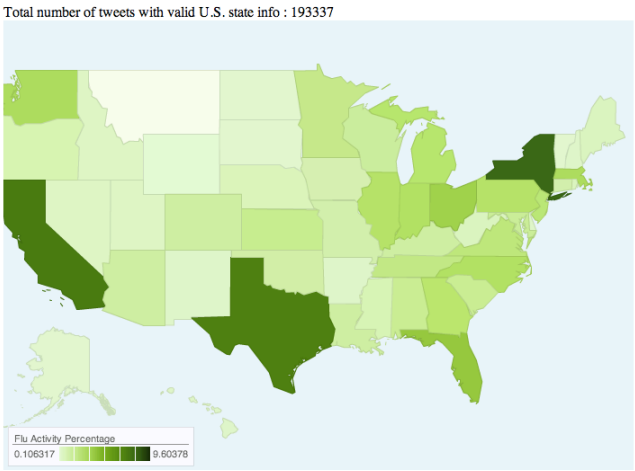

Figure 2: U.S. Flu Surveillance Map. The darker regions indicate that higher number of flu tweets were generated in the region. Users can see the state name and the percentage of flu tweets in the given state compared to the total tweets in the country by mousing over a state on the map.

4.2 Text Analysis

4.2.1 Flu Types. Figure 3 shows the distribution of tweets mentioning different types of flu. There are 62,730 tweets in total that mention at least one of the flus in the flu types list. 'Swine flu' is by far the most talked about flu (47.7%), followed by 'bat flu' (17.5%), 'bird flu' (12.8%), and 'H1N1' (11.4%). 'Bat flu' and 'H1N1' have noticeably increasing tweet volume in January 2013.

4.2.2 Symptoms. Figure 5 shows the distribution of tweets mentioning symptoms of flu. There are 98,052 tweets in total that mention at least one of the symptoms in the flu symptoms list. Among the tweets that mention at least one of the flu symptoms in our keyword list, 'cough' (27.9%) and 'fever' (26.1%) are the two most frequently mentioned symptoms, followed by 'sore throat' (15.8%), 'headache' (12.0%), 'chill' (4.2%), 'sneeze' (4/0%), 'vomit' (2.4%), 'body ache' (1.7%), 'strep throat' (1.6%), 'runny nose' (1.3%), 'diarrhea' (0.97%), 'nausea' (0.95%), 'stuffy nose' (0.48%), 'fatigue' (0.47%), and 'nasal congestion' (0.08%).

4.2.3 Treatments. Figure 7 shows the distribution of tweets mentioning various treatments used for flu. There are 13,505 tweets in total that mention at least one of the keywords in our flu treatments list. Interestingly, the most popular flu treatment is 'vitamin C' (31.1%), followed by 'tamiflu' (25.8%), 'theraflu' (16.6%), 'tylenol' (11.5%), 'vitamin D' (5.5%), and 'advil' (3.5%).

4.3 Temporal Analysis. First we show the overall flu related tweet volume changes over time, then show the volume changes of tweets mentioning each keyword in our three categories (flu types, symptoms, and treatments).

4.3.1 Flu Tweet Stream Timeline. Figure 9 shows the change of tweet volume mentioning the word 'flu' over past 10 days. The number of flu related tweets starts increasing dramatically starting January 9, 2013 and reaches peak on January 12, 2013. The daily flu tweet volume doubles in 3 days. January 9th, the starting point of tweet activity level increase on our Flu Tweet Stream Timeline, coincides with the date when USA Today, one of the largest national newspaper, released a news article titled *700 cases of flu prompt Boston to declare emergency* [19]. On January 12, The Huffington Post, another large newspaper, reported the death of four children from the outbreak of AH3N2 influenza [10]. This shows how our real-time temporal analysis of flu related tweets reflects the wide spread of current influenza outbreak.

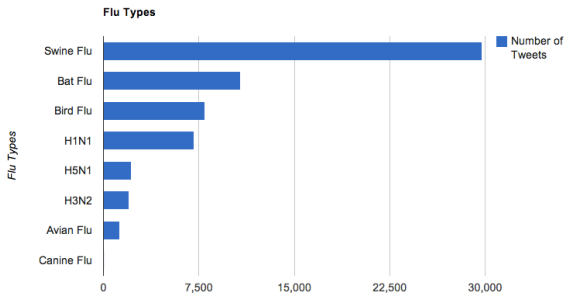

Figure 3: Flu Types Bar Chart shows distribution of tweets mentioning different types of flu.

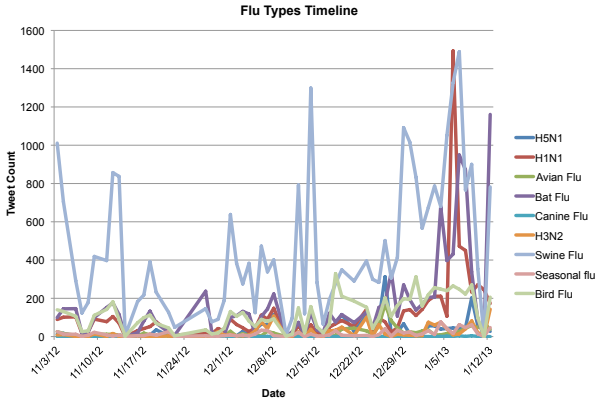

Figure 4: Flu Types Timeline shows daily tweet volume changes of different flu types.

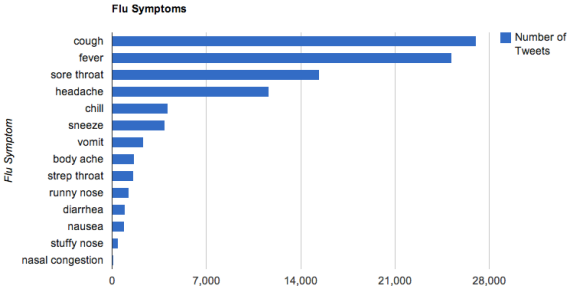

Figure 5: Flu Symptoms Bar Chart shows distribution of tweets mentioning different flu symptoms.

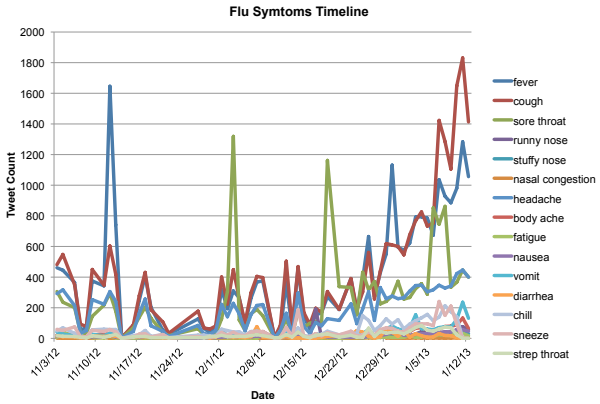

Figure 6: Flu Symptoms Timeline shows daily tweet volume changes of different flu symptoms.

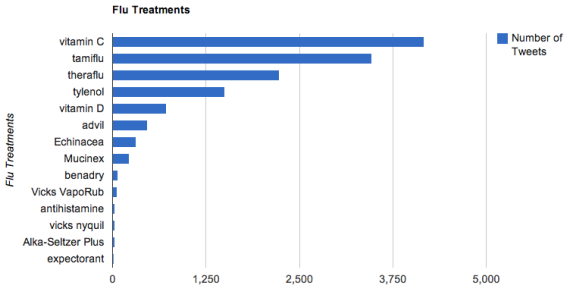

Figure 7: Flu Treatments Bar Chart shows distribution of tweets mentioning different flu treatments.

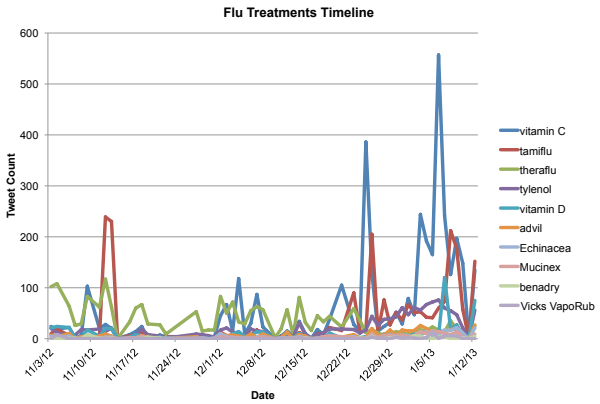

Figure 8: Flu Treatments Timeline shows daily tweet volume changes of different flu treatments.

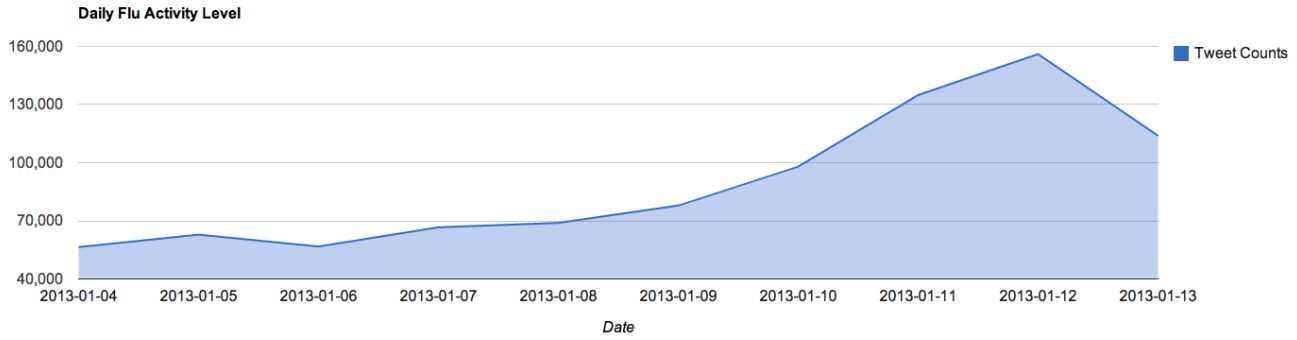

Figure 9: Flu Activity Timeline. The graph shows the change of tweet volume mentioning the word 'flu'. The dramatic increase of flu tweet volume from Jan. 9 to Jan. 12, coincides with the dates when the major US newspapers reported the Boston Flu Emergency and death of four children from the AH3N2 influenza outbreak.

4.3.2 Flu Types Timeline. Figure 4 shows the daily tweet volume changes of different types of flu (e.g., swine flu, bat flu, H1N1, H5N1, etc) over the past 70 days.

4.3.3 Symptoms Timeline. Figure 6 shows the daily tweet volume changes mentioning various symptoms of flu (e.g., fever, cough, sore throat, runny nose, chill, strep throat, etc) over time. The number of tweets mentioning 'cough' and 'fever', the most common influenza symptoms, starts to prominently rise starting mid-december and continues to increase. 'Cough' dominates over 'fever' during the second week of 2013. Other flu symptoms such as 'headache', 'body ache', 'vomit' as well as 'cough' reaches their peak points on January 12, where as the words 'sneeze' and 'sore throat' peaks 3-4 days before the other flu symptoms, indicating that 'sneeze' and 'sore throat' are the early signs of influenza. There are a couple of points in the timeline where 'fever' (dark blue line) and 'sore throat' (green line) gets unusually high tweet volumes compared to other symptom words ('fever' on Nov. 12 and Dec. 30, and 'sore throat' on Dec. 3 and Dec.19).

4.3.4 Treatments Timeline. Figure 8 shows the timeline of daily tweet volume changes of flu treatments (e.g., tamiflu, theraflu, vitamin C, vitamin D, tylenol, advil, etc) over past 70 days.

The popularity of the keyword 'theraflu' (green line) constantly dominates over the popularity of the keyword 'tamiflu' (red line) until December 21, 2012 except for the period of November 11, 2012 through November 15, 2012. In contrast, the use of the word 'tamiflu' suddenly rises around December 21, 2012 and stays high above the use of the word 'theraflu'. 'Theraflu' is an over-the-counter cold and flu medicine that anyone with mild flu symptoms can purchase at a nearby drug store while 'tamiflu' is an anti-viral drug that is used to slow the spread of influenza virus between cells in the body and can only be purchased with a doctor's prescription. This trend is very interesting because this shows that the increased need for prescription influenza drug confirming that the influenza spread across the country reached high levels since the end of year 2012, weeks before the usual time of late January in past years. Another interesting trend is that use of the keyword 'vitamin C' peaks a few days prior to the dates of the increased use of the flu medicine 'tamiflu'. This could be due to people trying to relieve their early flu-like symptoms by taking vitamin C. This fact conforms with the study by Gorton and Jarvis [13] that vitamin C administered before or after the appearance of cold of flu symptoms help relieve and prevent the symptoms.

4.4 Most Frequent Words (Word Cloud). Figure 10 is an example of high frequency words appearing in flu related tweets. We randomly picked 200 tweets that mention 'flu' and created a word cloud of all words in those 200 tweets. 'Flu' and 'RT', a symbol for retweet, are the two most frequently mentioned words. Retweet is a re-posting someone else's tweet and used to to quickly share impotant or interesting posts with his/her followers. 'Sick', 'shot', 'symptoms', 'stomach', 'catch', 'fever' and 'cold' are flu related words that frequently co-occur with 'flu'.

5 Conclusion

We designed and implemented a novel flu surveillance system that uses twitter data to automatically track flu activities in real time. Flu-related public twitter data is continuously downloaded using Twitter streaming API. Then the preprocessor module extracts tweet texts,

Figure 10: Word Cloud of flu related tweet texts.

time stamps, and user locations and stores them in a database, which is used for further analysis in the geographical, text, and temporal modeling modules.

The experiments show that our flu detection system can map U.S. regional influenza activity levels near real-time, discover and compare popularity of terms related to flu types, symptoms, and treatments. The system can also effectively track daily flu activities and the popularity of flu-related terms over time. All the output data is visualized as an interactive map, bar charts, and time series graphs on our project website [1]

Our system is scalable and can be easily extended to track other diseases. Because the system is completely automated, it will be a very low-cost alternative to replace the traditional high-cost disease surveillance system that collects public health data from sentinel medical practices.

In the future, we are interested in tracking the disease spread by looking at how disease-related tweets propagate on social media networks. We are also interested in using the social network information (e.g., friends and followers network) to predict the disease outbreak. We plan to gather additional online data (e.g., Facebook, blogs data, news feeds) for real-time disease surveillance. Another possible research direction is to apply our system for real-time surveillance of other infectious diseases.

References

[1] Digital flu surveillance. `http://pulse.eecs.northwestern.edu/~kml649/flu/index.php`.

[2] flusurvey. `http://flusurvey.org.uk`.

[3] Flutrackers.com. `http://www.flutrackers.com/forum/index.php`.

[4] Google chart tools. `https://developers.google.com/chart/interactive/docs/gallery`.

[5] Pulse of the nation. `http://www.ccs.neu.edu/home/amislove/twittermood`.

[6] Twitter streaming api. `https://dev.twitter.com/docs/streaming-apis`.

[7] Centers for disease control and prevention, seasonal influenza (flu). `http://www.cdc.gov/flu`, 2012.

[8] Google flu trends. `http://www.google.org/flutrends`, 2012.

[9] World of DTC marketing.com, web first place people go for health information. but you knew that already didn't you. `http://worldofdtcmarketing.com`, 2012.

[10] The Huffington Post, michigan flu season 2013: Four children die in influenza outbreak of ah3n2. `http://www.huffingtonpost.com/2013/01/12/michigan-flu-season-2013-ah3n2_n_2458916.html`, 2013.

[11] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, 2011.

[12] S. Asur and B. A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[13] H. C. Gorton and K. Jarvis. The effectiveness of vitamin c in preventing and relieving the symptoms of virus-induced respiratory infections. *J Manipulative Physiol Ther*, 22, 1999.

[14] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*. The AAAI Press, 2011.

[15] F. Pervaiz, M. Pervaiz, N. Abdur Rehman, and U. Saif. Flubreaks: Early epidemic detection from google flu trends. *J Med Internet Res*, 14(5):e125, Oct 2012.

[16] D. M. Ramblings. [how many people use the top social media? `http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media/`, 2012.

[17] T. Sakaki, M. Okazaki, and Y. 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, New York, NY, USA, 2010. ACM.

[18] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. *PLoS ONE*, 6(5):e19467, 05 2011.

[19] USAToday. 700 cases of flu prompt boston to declare emergency. `http://www.usatoday.com/story/news/nation/2013/01/09/boston-declares-flu-emergency/1820975`, 2013.