Real-Time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer

Kathy Lee Ankit Agrawal Alok Choudhary
EECS Department
Northwestern University
Evanston, IL USA
{kml649, ankitag, choudhar}@eecs.northwestern.edu

ABSTRACT

Social media is producing massive amounts of data on an unprecedented scale. Here people share their experiences and opinions on various topics, including personal health issues, symptoms, treatments, side-effects, and so on. This makes publicly available social media data an invaluable resource for mining interesting and actionable healthcare insights. In this paper, we describe a novel real-time flu and cancer surveillance system that uses spatial, temporal, and text mining on Twitter data. The real-time analysis results are reported visually in terms of US disease surveillance maps, distribution and timelines of disease types, symptoms, and treatments, in addition to overall disease activity timelines on our project website. Our surveillance system can be very useful not only for early prediction of seasonal disease outbreaks such as flu, but also for monitoring distribution of cancer patients with different cancer types and symptoms in each state and the popularity of treatments used. The resulting insights are expected to help facilitate faster response to and preparation for epidemics and also be very useful for both patients and doctors to make more informed decisions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering, Selection process*

Keywords

Influenza, Cancer, Disease Detection, Public Health, Disease Surveillance, Epidemics, Social Media, Twitter

1. INTRODUCTION

The Internet is usually the first place people turn for health information. People search for a specific disease, symptoms, and appropriate medical treatments, and often make decisions whether they should go see a doctor based on the search results. Healthcare portal sites and the social media are popular online health information resources among US Internet users [6]. Disease surveillance is the monitoring of clinical syndromes such as flu and cancer that have a significant impact on medical resource allocation and health policy. Disease surveillance plays an important role in minimizing the harm caused by the outbreaks by constantly observing the disease spread. The traditional approach employed by the Centers for Disease Control and Prevention (CDC) [5] for flu surveillance includes the collection of Influenza-like Illness (ILI) patients' data from sentinel medical practices. The main drawback of this method is the 1-2 weeks time lag between the time of medical diagnosis and the time when the data becomes available. 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. We

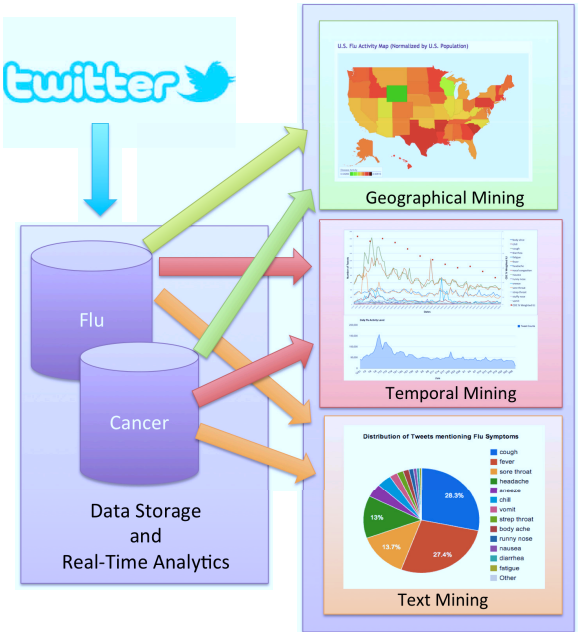

Figure 1: Real-Time Disease Surveillance System continuously downloads flu and cancer related tweets and applies geographical, temporal, and text mining. The real-time analysis data is visually reported as U.S. disease activity map, timelines, and pie charts on our project website [1][2].

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

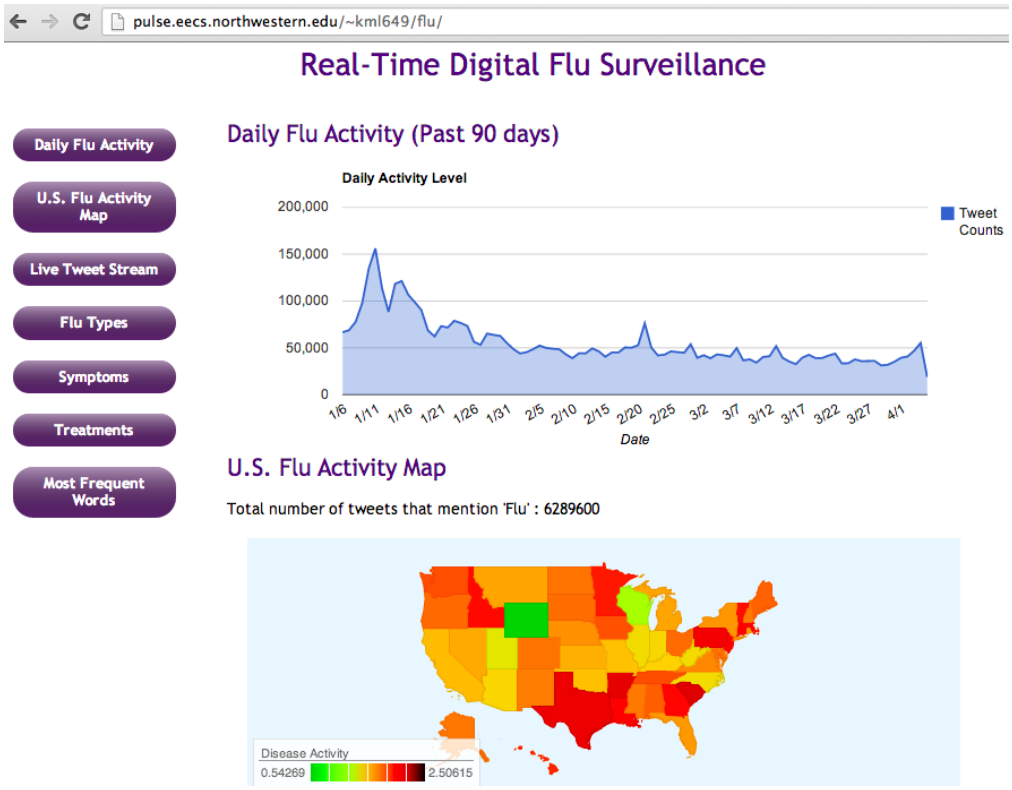

Figure 2: Screenshot of our Real-Time Digital Flu Surveillance Webpage [2]. The 'Daily Flu Activity' chart is an output of the temporal analysis and shows the volume changes of tweets mentioning the word 'flu' over time. The dramatic increase of flu tweet volume from Jan. 6 to Jan. 12 coincides with the dates when the major US newspapers reported the Boston Flu Emergency [8] and death of four children from the AH3N2 influenza outbreak [7]. The 'U.S. Flu Activity Map' is an output of the geographical analysis and shows the weighted percentage of tweet volumes mentioning 'flu' by states. Level of flu activity is differentiated by different colors for easy comparison of US regional flu epidemic.

built a novel real-time disease surveillance system that uses twitter data to track US influenza and cancer activities. Twitter [3] is a popular micro-blogging service where users can post short messages. Twitter's popularity as a medium for real-time information dissemination has been constantly increasing since its launch in 2006. The proposed system continuously downloads flu and cancer related twitter data using Twitter streaming API [4] and applies spatial, temporal, and text models on this data to discover national flu and cancer activities and popularity of disease-related terms. The output of the three models is summarized as pie charts, time-series graphs, and US disease activity maps on our project website [1][2] in real time. This demonstration is built upon and extends our previous work [10]. In this work, the text analysis on most frequently occurring terms is added. We further extended our real-time disease surveillance system to track cancer activities in addition to flu.

2. SYSTEM DESCRIPTION

Figure 1 shows the architecture of our real-time flu and cancer surveillance system. Our dataset consists of all recent tweets that mention the keywords 'flu' or 'cancer'. We have collected over 6 million flu-related tweets generated by more than 3.3 million unique users in past 5.5 months since October 16, 2012, and over 3.7 million cancer-related tweets generated by more than 1.3 million unique users in past 3 months since January 7, 2013. With continuous tweet monitoring, we plan and expect to collect bigger and bigger quantities of data in the coming months. Such big data presents a number of challenges due to its size and complexity, relating to its storage, retrieval, analysis, and visualization, especially when the whole process is required to be done in real-time, as in this work. Our system is designed to be a disease surveillance system that is (almost) always available, robust, and easily scalable for big data. Different from many other related big data projects, which perform analytics on a massive, static dataset, our system consists of a cluster of several transactional databases and high-dimensional data warehouses which are updated in real time. Three types of analytics are considered here - geographical/spatial, temporal, and textual, the results of which are suitably presented pictorially, as described next.

2.1 Geographical Analysis

The goal of geographical analysis is to track the disease spread in US states by measuring the volume of flu/cancer tweets generated in the region. For our experiments, we use user's home location in his/her Twitter profile. The dataset for geographic analysis is all users who mention 'flu'

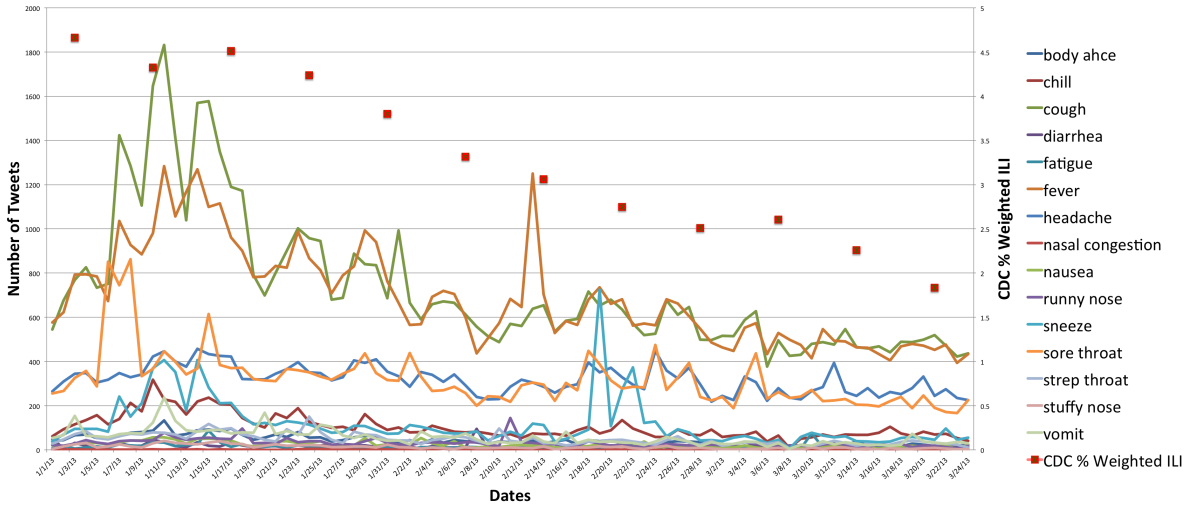

Figure 3: Flu Symptoms Timeline. The timeline displays tweet volume changes mentioning different flu symptoms from January through March 2013. 'Cough' (green line) and 'fever' (dark orange line) reach their highest levels in mid January and decrease as the actual national ILI level by CDC decreases.

or 'cancer' and have a valid US state info (e.g., 'Evanston, IL', 'somewhere in NY') in their home location field. We exclude tweets generated from outside the US (i.e., tweets from foreign location) and those with invalid location information (e.g., 'travelling', 'Wherever the wind blows me'). In our current flu dataset, there are 458,828 users with valid US state information, and in current cancer dataset, there are 193,797 users with valid US state information. The *US Flu Activity Map* is shown in Figure 2. The tweet volume mentioning 'flu' generated in each state is normalized by the population of the state.

2.2 Temporal Analysis

The goal of temporal analysis is to track the volume changes of tweets mentioning the disease and related terms over time.

2.2.1 *Disease Daily Activity Timeline*

As shown in Figure 2, *Daily Flu Activity* chart shows the tweet volume changes of flu-related tweets over three months period from January through March 2013. The data for flu/cancer timeline is created by counting the number of tweets mentioning 'flu' or 'cancer' generated daily. Our assumption is that people talk more about 'flu' when they themselves or people around them (family or friends) have flu symptoms and there are more frequent news feeds when the epidemic is wide spread. Achrekar et al. [9] reported that the volume of flu-related tweets is highly correlated with the number of reported ILI cases by the CDC. In the flu timeline, the number of flu related tweets start increasing on January 6 and reaches its peak on January 12, which coincides with the date when The Huffington Post reported the death of four children from the outbreak of AH3N2 influenza [7]. This shows how our temporal analysis effectively reflects the wide spread of the epidemics.

2.2.2 *Types, Symptoms, Treatments Timeline*

We not only track the overall flu and cancer activities, but also monitor disease types, symptoms, and treatments over time. Figure 3 shows daily tweet volume changes for various flu symptoms. From the timeline chart, we can easily tell the types and levels of flu symptoms in the general population at a specific point in time. Cough and fever are the two most dominant symptoms throughout all flu season, and headache and sore throat are the next two most common flu symptoms. The actual US national influenza activity level (percentage weighted Influenza-like Illness by the CDC) is plotted as red squares for reference. Tweet volumes mentioning flu symptoms reach their highest point around mid January and decrease as the actual flu activity level by CDC decreases.

2.3 Text Analysis

In text analysis, we reveal deep health insights by examining the content of the tweets.

2.3.1 *Disease Types, Symptoms & Treatments*

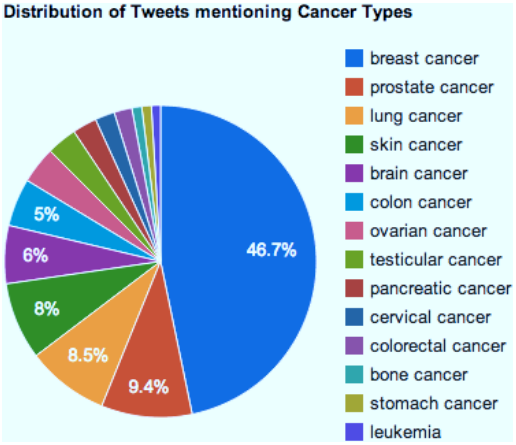

Figure 4: Cancer Types.

We are interested in investigating the popularity of terms used in three categories: (1) disease types (2) symptoms (3) treatments, and have created a keyword list for each category. For example, the keyword list for cancer types is a list of breast cancer, lung cancer, skin cancer, brain cancer, etc., the keyword list for cancer symptoms is a list of lump,

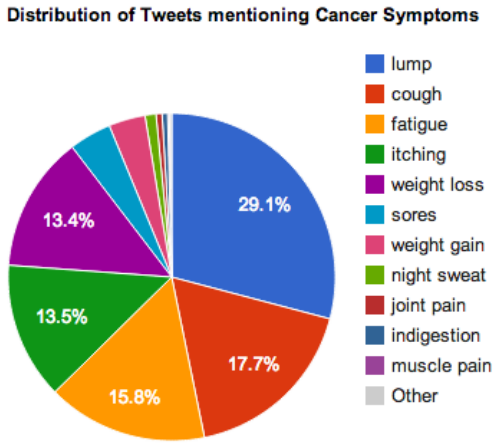

Figure 5: Cancer Symptoms.

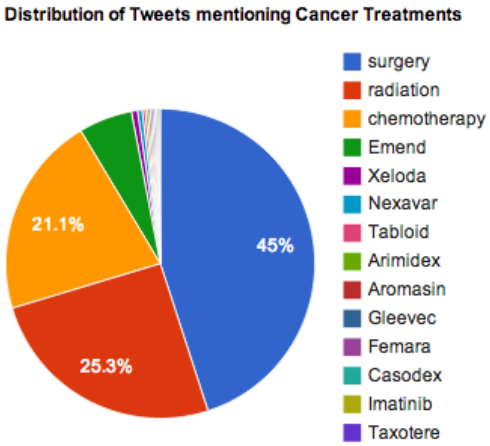

Figure 6: Cancer Treatments.

cough, fatigue, weight loss, etc, and the keyword list for cancer treatments is a list of surgery, radiation, chemotherapy, Emend, Xeloda, etc. We also have similar keyword lists for 'flu'. For 'flu', we have 9 flu types, 15 symptoms, and 31 treatments. For 'cancer', we have 58 cancer types, 21 symptoms, and 63 treatments. Figure 4, 5, and 6 show the distribution of tweets mentioning a keyword in cancer types, symptoms, and treatments keyword lists.

2.3.2 *Most Frequent Words*

We are interested in investigating which words frequently co-occur with a disease name. After tokenizing tweet texts and removing all stop words, we count the number of occurrence of each unique word. Our current flu dataset (6,097,406 tweets) consists of 83,896,915 words and 4,001,445 unique words. Figure 7 shows the top 20 most frequent words in our entire flu dataset.

3. CONCLUSION

We built a real-time disease surveillance system that uses twitter data to automatically track flu and cancer activities. The experiments show that our disease detection system can map US regional influenza and cancer activity levels near real-time, discover and compare popularity of terms related to flu/cancer types, symptoms, and treatments. The system

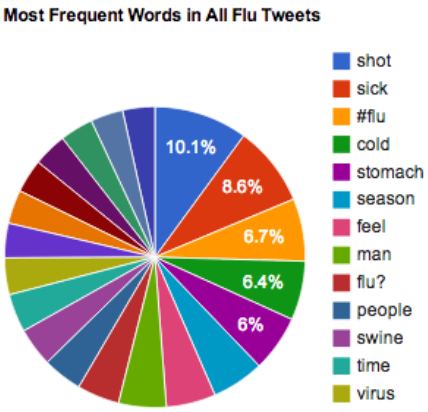

Figure 7: Most Frequent Words in Flu Tweets.

can also effectively track daily flu/cancer activities and the volume changes of tweets mentioning disease related terms over time. All of the output data is visualized as interactive maps, pie charts, and time series graphs on our project website [1][2]. Our system is highly scalable and can be easily extended to track other diseases. Because the system is completely automated and the output of analysis is updated near real time, it can detect disease outbreaks significantly faster than the traditional disease surveillance system that collects public health data from sentinel medical practices. In the future, we are interested in using the social network information (e.g., friends and followers network) to predict the disease outbreak.

4. ACKNOWLEDGMENTS

This work is supported in part by the following grants: NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, and OCI-1144061; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, and DESC0007456; AFOSR award FA9550-12-1-0458.

5. REFERENCES

[1] Real-time digital cancer surveillance. `http://pulse.eecs.northwestern.edu/~kml649/cancer/`.
[2] Real-time digital flu surveillance. `http://pulse.eecs.northwestern.edu/~kml649/flu/`.
[3] Twitter. `https://twitter.com/`.
[4] Twitter streaming api. `https://dev.twitter.com/docs/streaming-apis`.
[5] Centers for Disease Control and Prevention, seasonal influenza (flu). `http://www.cdc.gov/flu`, 2012.
[6] 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.
[7] 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.
[8] USA Today, 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.
[9] 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.
[10] K. Lee, A. Agrawal, and A. Choudhary. Real-time digital flu surveillance using twitter data. In *The 2nd Workshop on Data Mining for Medicine and Healthcare*, 2013.