J Med Syst (2016) 40: 189
DOI 10.1007/s10916-016-0545-y

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Regional Level Influenza Study with Geo-Tagged Twitter Data

Feng Wang[1] · Haiyan Wang[1] · Kuai Xu[1] · Ross Raymond[1] · Jaime Chon[1] · Shaun Fuller[1] · Anton Debruyn[1]

Received: 17 November 2015 / Accepted: 10 June 2016 / Published online: 2 July 2016

Abstract The rich data generated and read by millions of users on social media tells what is happening in the real world in a rapid and accurate fashion. In recent years many researchers have explored real-time streaming data from Twitter for a broad range of applications, including predicting stock markets and public health trend. In this paper we design, implement, and evaluate a prototype system to collect and analyze influenza statuses over different geographical locations with real-time tweet streams. We investigate the correlation between the Twitter flu counts and the official statistics from the Center for Disease Control and Prevention (CDC) and discover that real-time tweet streams capture the dynamics of influenza cases at both national and regional level and could potentially serve as an early warning system of influenza epidemics. Furthermore, we propose a dynamic mathematical model which can forecast Twitter flu counts with high accuracy.

Keywords Influenza · Regional level · Partial differential equation modeling · Geo-tagged twitter stream

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Feng Wang
fwang25@asu.edu

Haiyan Wang
Haiyan.Wang@asu.edu

Kuai Xu
Kuai.Xu@asu.edu

[1] School of Mathematical and Natural Sciences, New College of Interdisciplinary Arts and Sciences, Arizona State University, Glendale, Arizona USA

Introduction

Accurate flu reporting and flu forecasting is important for improving public health since it can lead to accurate planning for hospital personnel and vaccine. Although government agencies such as the CDC regularly report official statistics on the trends of influenza or outbreaks such as SARS and Ebola, these statistics often fail to reflect the latest development and progress since there is delay caused by manual data collections and complicated reporting processes. Researchers have explored the possibility of consulting big data for public health purpose. For example, Google Flu Trend (GFT) is a popular analytics tool aiming at predicting the location and severity of flu outbreaks [1]. However, GFT modeling has recently been criticized due to the lack of transparency of Google flu data and the overestimation of influenza prevalence by conflating signals of influenza awareness (such as media attention) with signals of actual infection [2].

The rising popularity of social media has led people to share their flu statuses and symptoms online, thus allowing an alternative channel to collect, analyze and monitor the latest trends of influenza development. Recent research [3] argues that Twitter provides open data collection and the interests in flu and the number of real flu cases are separable in Twitter flu data, which makes it a valuable and trustworthy source of flu analysis and modeling. Broniatowski et al. [4] investigate Twitter flu data during 2012 and 2013 from both national and local level. To be more specific, they collaborated with New York City's Department of Health and Hygiene to discover the strong correlation between Twitter indicative flu cases in NYC and the municipal data. Achrekar et al. [5] also discovered the volume of flu related tweets is highly correlated with the number of Influenza Like Illness (ILI) cases reported by CDC based

on data collected during 2009 and 2010 at the national level.

Towards building a flu-surveillance system and studying whether Twitter data can be used as a robust indicator of influenza at both *national* and *regional* level, this paper designs, implements, and evaluates a prototype system which automatically collects, analyzes and models geo-tagged flu tweets from real-time Twitter data streams. Specifically, we extract flu tweets from real-time data streams and tag each tweet with geographical locations based on three information sources: i) the geographical location in the profile of the user who tweeted the message, ii) the physical location where the user sent the tweet and enabled their geographical location tracking in the Twitter App, or iii) the geographical location mentioned in the content of the tweets. The empirical analysis reveals strong correlation between the flu tweet counts and the CDC-reported ILI cases at both national level and regional level. To be more specific, eight out of the ten CDC-defined regions demonstrate a correlation coefficient of around 0.9. Furthermore, we propose a partial differential equation based mathematical model to predict the Twitter flu tweet counts in a certain region during a certain week. We use the tweet flu data collected over 13 weeks from January to March 2014 to fit to the model. Our experimental results show the model achieves a prediction accuracy of 83.88 % if the first 11-week data is used as a training set to predict the flu tweet count of week 12. The surveillance system allows us to characterize and model flu trends at different geographical locations in real-time, which can serve as early warning signals before CDC releases official statistics, typically with a couple of weeks delay and possible revisions.

In summary, the contributions of this paper are three-fold as follows:

- We develop a prototype framework which automates flu tweet collection, flu tweet processing such as geo-tagging Twitter flu tweets, separating tweets about flu cases and tweets about flu awareness, etc, and mathematical modeling.
- We discover strong correlation between Twitter indicative flu cases and CDC released ILI cases at both national and regional levels. To the best of our knowledge, this is the first paper which studies the correlation at regional level.
- We propose a PDE-based mathematical model to predict the Twitter flu trend from both temporal and spatial dimension, i.e., the model can predict the Twitter flu cases in a certain region at a certain week, while most existing models only focus on temporal dimension.

The remainder of this paper is organized as follows. Section "System framework for tweet data collection, processing and analysis, and mathematical modeling" describes the architecture of the real-time system we have developed for collecting, processing, and modeling influenza statuses with Twitter data streams. In Section "Statistics of collected flu tweets and correlation analysis", we present empirical results with real-world tweets and characterize the correlation between flu tweet counts and reported flu cases at CDC at both national and regional level. Section "Mathematical modeling" introduces the mathematical model and presents the prediction accuracy of the model. Section "Related works" describes related work of modeling influenza with social media data, while Section "Conclusions and future work" concludes this paper and outlines our future work.

System framework for tweet data collection, processing and analysis, and mathematical modeling

Figure 1 illustrates the framework to collect, process, and model Twitter data for flu surveillance.

Twitter data collection

Twitter provides two mechanisms for programmatic access to their data encoded in JSON format: REST API and Stream API. The REST API provides access to user profile and follower data. This data allows the generation of a user

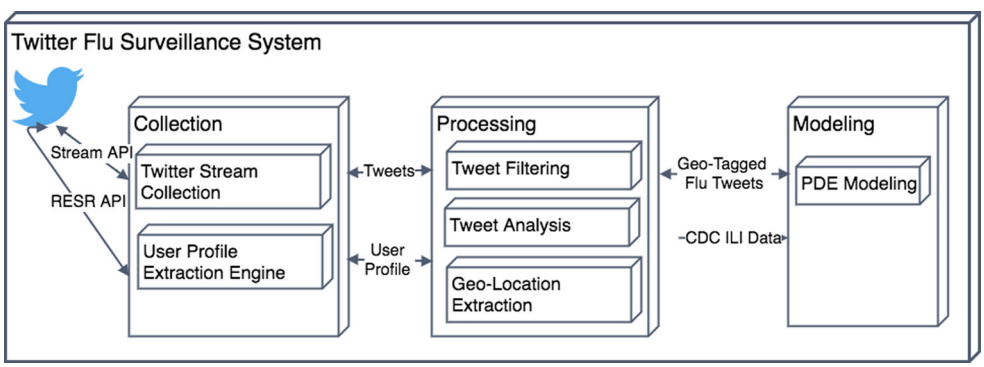


Fig. 1 Twitter Flu Surveillance System

following topology, which captures who follows whom in Twitter and is critical to study the influence between Twitter users. The Streaming API continuously delivers real-time stream of tweets matching a given search query over a persistent HTTP connection. The stream API offers two endpoints: Sample and Filter. The Sample endpoint delivers a small random sample (typically 1 %) of all public statuses (tweets). The Filter endpoint delivers all tweets that match a given query, which can include keywords, locations, and users, up to 1 % of all public Tweets. The aggregate Tweet information can be used as an indicator of what happens in real life.

We collect both Twitter user and tweets information. **Twitter Stream Data Collection** module handles establishing and maintaining the connection with Twitter servers to retrieve tweets based on chosen keywords and/or tweets with a particular user as a source. Tweets are saved in JSON format in flat files with collection dates as the file name. JSON format is chosen to eliminate encoding/decoding errors, efficiency, and flexibility. Flat file is chosen to remove the overhead of a traditional database.

User profile extraction module handles the collection of user data, including user profile and user's follower ids for every user observed in the Tweet data stream. The fetched data is stored in a MongoDB database, which is a document-based non-SQL database which provides fast and scalable storage. MongoDB is chosen due to its ability to handle the complex random reads needed by our user processing algorithms.

Twitter data processing

Tweet filtering module carries out the cleaning of the raw tweet stream from Twitter. The main functions of the module are: filtering and handling of messages (notifications), removing extraneous fields from tweets and user profiles, and reordering user's tweets). The raw Tweet stream contains not only tweets matching the given query, but additional messages (notifications) from Twitter. These messages need to be filtered from the Tweet Stream. Two examples of notifications are: a tweet matching the query has been deleted, or the backlog for the stream is filling up. The latter occurs when the tweet processing algorithm is not fast enough to process the tweets as fast as Twitter sends them; 2) We also remove all extraneous fields for each tweet to conserve disk space. This filter can be safely applied since a tweet contains a number of fields not relevant to our processing needs. The removed fields are only relevant in the case of displaying a tweet to a user. For example, the tweet may contain a url for a CSS file containing custom formatting information for a particular tweet). The last component of this module reorders the tweets on disk to ensure tweets are in order of their timestamp. This needs to be done because Twitter does not guarantee tweets will arrive in order, users are in different time zones, and most importantly all of the tweet processing algorithms depend on processing tweets in order for the simplicity/efficiency of the algorithms.

Geo-location extraction module Based on [6], 84.2 % of twitter users have specified location in their twitter profiles and 10.3 % of twitter users have geo-location enabled Tweet. However, there are still challenges as follows: only a very small percentage of Twitter users add GPS information to their tweets,; a significant number of users attempt to thwart automated systems by using bogus locations in their profile or by using valid locations with non-standard spelling or characters ("a" would be replace by "@"); all geocoding services have api limits that would be easily reached, and currently all geocoding services rely on input being as close to a location as possible and not on random text that may contain a location. We implement our module based on the "carmen" library [7] for geocoding tweets. This library provides us with a framework to resolve Tweet locations, and a small database of known locations.The included database contains names of states, abbreviations, cities, and common misspellings. We have expanded the database to include more of the previously mentioned entities, as well as zip codes, airport codes, monuments, etc. We add enhancement to the Carmen library by adding a new resolver to process the tweet text. The four fields we use to geo-tag a tweet are coordinates, place, profile.location, and text.

Tweet analysis module Besides generating statistics of collected tweets, a major functionality of this module is to discover the retweet relationship, which is a mapping from a source (original) tweet to all its recursive retweets/replies. Discovering which tweets are retweet, reply, or contain the same content as the source tweet is important since when the flu tweet count is calculated, only source tweets instead of all tweets mentioning flu are considered. This is because for source tweets with tens of thousands of retweets/replies, most of the retweets are just simple "take care" or "get well" which does not reflect whether the person who retweets the source tweet has flu symptoms or not. The flu cases are majorly captured in the source tweets.

There are three ways that a tweet can be retweeted. User clicks the "retweet" button, or "reply" button, or directly retweets a tweet by typing RT at the beginning of a Tweet to indicate they are re-posting someone else's content.

To identify the source tweets and count the number of retweets, we go through each tweet and check if it is a reply (checks if the "in_reply_to_status_id" has a value) or a retweet (checks the "retweeted" field and if the pattern "RT @" occurs within the text), then increments a counter if the

tweet belongs to an already identified source tweet. If the tweet is not categorized as reply or a retweet, the tweet id and text is stored as a source tweet.

In the case where the "retweeted" field is missing and the tweet contains the "RT @user name" pattern in the text, the algorithm will employ various techniques to compare the text to the text of already discovered tweets. One key item to note is that the retweet pattern can occur several times. The algorithm will loop through each "RT @" pattern from the outermost to the innermost and check if the user name belongs to any observed tweet. If this check passes, the algorithm will attempt to match the text to the already observed texts. This means taking into account truncated text, as well as user added messages in the beginning and in the end of the text.

Mathematical modeling

After data is collected, cleaned, and analyzed, the last component in our prototype is the mathematical modeling. PDE models are used to describe temporal and spatial diffusion of flu related topics and predict flu trend in real life.

Statistics of collected flu tweets and correlation analysis

Statistics of collected flu tweets

We have collected raw tweets that contain the keyword *flu*. This is one of the most common illnesses that CDC tracks closely. Table 1 gives a brief summary of the scale of the collected data.

Table 1 Flu tweet data

Category	Amount
Total Size of Tweets	3.5GB
Total Number of Tweets	2,945,941
Total Number of Source Tweets	1,439,678
Total Number of Unique Users	1,592,460
Data Collection Start	January 3, 2014
Data Collection End	March 24, 2014

Correlation between Twitter flu tweet trend and CDC reported ILI case trend

To verify the relevance of flu trends modeled by our system, we correlate geo-tagged flu tweets with the reported flu cases released from CDC official statistics. We adopt the flu data collected between January 3 and March 26, 2014, which is a subset of all the collected flu data that align with the flu season.

Figure 2 shows the number of weekly new flu tweets in Twitter and the number of weekly reported ILI cases provided by CDC. It shows strong linear correlation between the lines. In order to measure the linear correlation, we adopt Pearson's product-moment correlation coefficient $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where $cov(X, Y)$ is the covariance between variable X and Y, and σ_X is the standard deviation of X. The

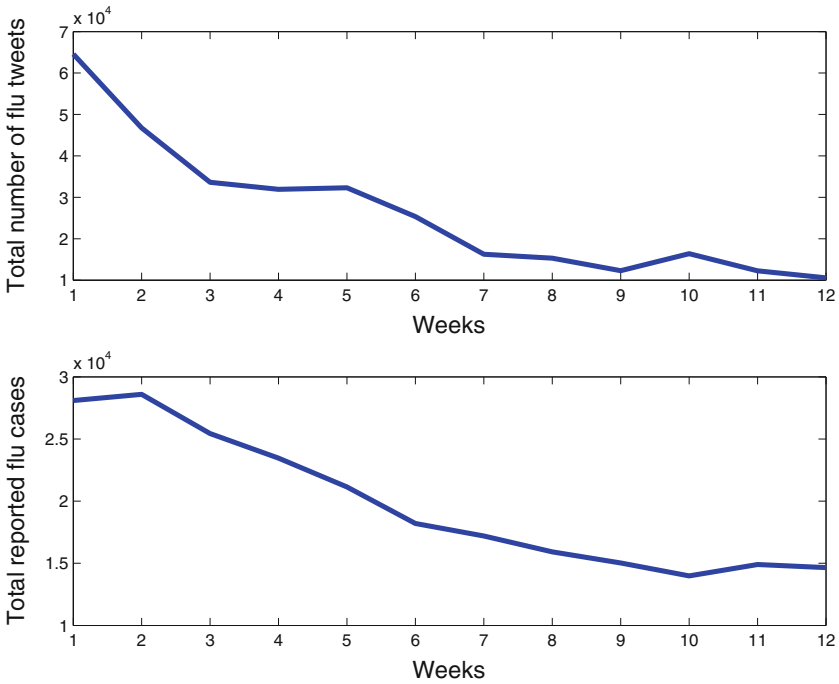


Fig. 2 Twitter flu tweet trend vs. CDC reported ILI case trend at National Level

result shows that the correlation coefficient between Twitter weekly new flu tweets count and the newly reported ILI is as high as 0.9297.

We further divide the flu tweet counts by regions and investigate the correlation between regional tweet counts and CDC regional ILI cases to investigate whether Twitter data can be used to indicate the flu trend at regional level. Figure 3a illustrates the 10 regions defined by CDC [8]. Figure 3b shows the Pearson's correlation coefficient between CDC ILI cases and Twitter flu tweet counts for each region. As we can see, except for Region 1 and Region 2, eight CDC-defined regions show strong correlation. For example, region 6 has correlation coefficient as 92.82 %, and region 10 has correlation coefficient as 97.54 %. The low correlation of Region 1 and Region 2 may be caused by noises in the tweet text and needs further investigation.

In summary, the empirical results reveal a strong temporal correlation between flu tweet counts and CDC ILI cases in both national and regional levels. The strong correlation demonstrates the potential application of our system for providing early prediction and warning of flu trends in finer granularity. In next section, we present our work to study the regional level Twitter flu cases prediction.

Mathematical modeling

PDE-based mathematical modeling

In this section, we present a dynamic spatial-temporal mathematical model. Our model is different from existing statistical regression models and ordinal differential equation models in that we can give prediction in both spatial and temporal dimension, that is, we can predict the number of flu cases (called user density in the description of the model) in region x and at week t, while most existing models can only predict the number of flu cases at a certain time t. Following is the description of the PDE model.

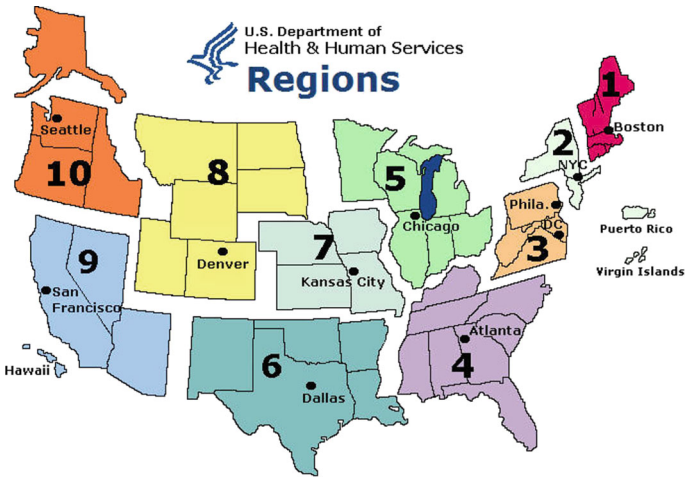
$$\frac{\partial u}{\partial t} = \frac{\partial (ae^{-bx} \frac{\partial u}{\partial x})}{\partial x} + r(t)\, u \left[h(x) - \frac{u}{K}\right] \quad (1)$$

where

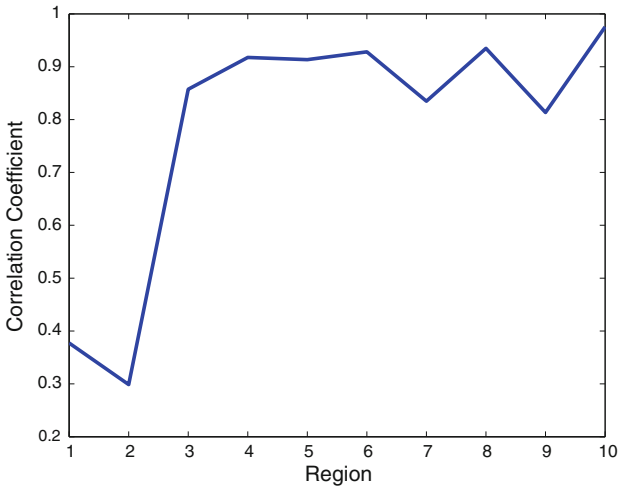
- $u = u(x, t)$ is the user density in region x at time t
- $\frac{\partial u}{\partial t}$ corresponds to the rate of change of user density as time progresses
- $ae^{-bx}, a, b > 0$ is an exponential decay term for the number of interactions between different regions
- $\frac{\partial u}{\partial x}$ corresponds to the rate of change of user density across regions
- $r(t)$ represents the decay or growth of user density with respect to time t
- $h(x)$ represents the heterogeneity of growth rate in region x
- K is the carrying capacity for the system

Equation 1 has **three primary components**.

1. $\frac{\partial u}{\partial t}$ is a mathematical description of how user density in each region changes over time
2. $\frac{\partial (ae^{-bx} \frac{\partial u}{\partial x})}{\partial x}$ corresponds to the flu spreading within a certain region
3. $r(t)\, u \left[h(x) - \frac{u}{K}\right]$ corresponds to flu spreading across regions



(a) CDC-defined region map

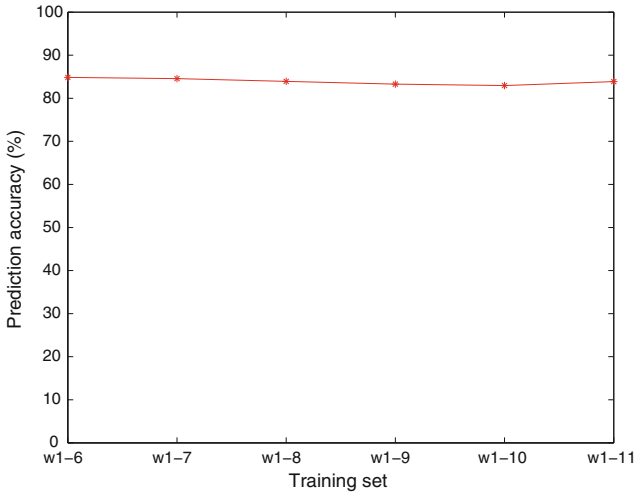


(b) Correlation between Twitter flu tweet counts and CDC reported regional ILI cases

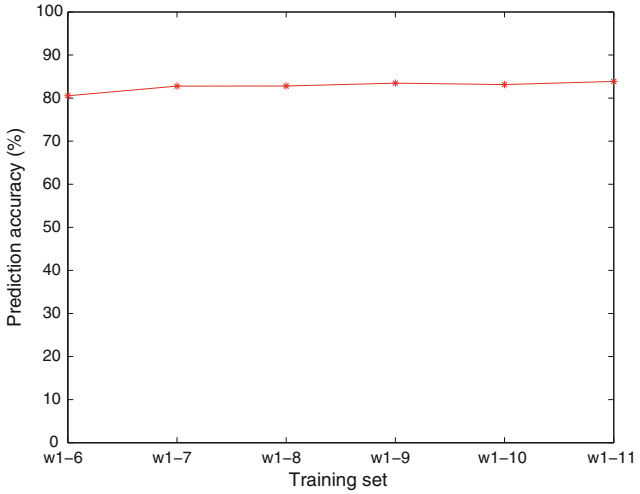
Fig. 3 Regional Twitter flu tweet trend vs. CDC reported regional ILI case trend

Prediction Accuracy

In this section, we present the prediction accuracy of our proposed PDE model. We carried out two sets of evaluations: single week prediction and multiple week prediction. For single week prediction, we vary the period of the training data set from the 6 weeks to 11 weeks and predict the twitter flu cases for the following week, i.e., we use week 1-6, 1-7, ... 1-11 as training data, and predict for the following week, 7, 8, ... 12 correspondingly and record the average prediction accuracy of all ten regions as shown in Fig. 4a); For multiple weeks prediction, we vary the period of the training data set from 6 weeks to 11 weeks and predict the twitter flu cases for the remaining weeks, i.e., use week 1-6, 1-7, ... 1-11 as training data and predict for week 7-12, 8-12, ... 12 correspondingly and record the average prediction accuracy of all ten regions as shown in Fig. 4b). Our experimental results also show that the model can predict the flu tweet counts for all ten regions with similar accuracy. It is clear that our model can achieve stable and relatively high prediction accuracy across all ten regions for both single week prediction and multiple weeks prediction. For example, Fig. 4a) shows that the model achieves a prediction accuracy of 84.57 % if the first 7-week data is used as training set to predict the flu tweet count of week 8. The model achieves a prediction accuracy of 83.88 % if the first 11-week data is used as training set to predict the flu tweet count of week 12. Figure 4b) shows that the model achieves a prediction accuracy of 82.70 % if the first 7-week data is used as training set to predict the flu tweet count of week 8 to week 12. This indicates that our model can be used as an analysis tool to predict the twitter flu cases in real time in the granularity of regions. Since there exists strong correlation between twitter flu cases and CDC reported ILI cases for eight out of ten CDC defined regions, we argue that our framework can be used to analyze and predict the flu cases in real life.



(a) Single Week Prediction Accuracy



(b) Multiple Weeks Prediction Accuracy

Fig. 4 Prediction Accuracy for Single Week and Multiple Week Prediction

Related works

Many research studies have explored the possibility of consulting big data for public health purpose. For example, Google Flu Trend (GFT) is a popular analytics tool aiming at predicting the location and severity of flu outbreaks [1]. However, GFT modeling has recently been criticized due to the lack of transparency of Google flu data and the overestimation of influenza prevalence by conflating signals of influenza awareness (such as media attention) with signals of actual infection [2].

The rising popularity of social media has led people to share their flu statuses and symptoms online, thus allowing an alternative channel to collect, analyze and monitor the latest trends of influenza development. Recent research [3] argues that Twitter provides open data collection. In addition, the interest in flu and the number of real flu cases are separable in Twitter flu data, which makes it a valuable and trustworthy source of flu analysis and modeling.

Some research investigated the role Twitter stream data playing in the scope of general flu trend description and prediction, that is, they collected flu tweets from Twitter stream, correlated Twitter flu data with the CDC ILI flu cases, and predict the ILI cases based on Twitter flu data and historical CDC ILI cases. Broniatowski et al. [4] investigated the Twitter flu data during 2012 and 2013 from both national and local level. To be more specific, they collaborate with New York City's Department of Health and Hygiene to discover the strong correlation between Twitter indicative flu cases in NYC and the municipal data. Achrekar et al. [5] also discovered the volume of flu related tweets is highly correlated with the number of ILI cases

reported by CDC based on data collected during 2009 and 2010 at the national level. It also devises an auto-regression model to predict the ILI activity level in a population by utilizing both historical CDC data and Twitter flu tweet count. In addition, [10, 11] presented a disease surveillance system which captures and visualizes the Twitter flu information including geographical tweet distribution, tweet text analysis, and temporal volume change of flu tweets. Our work is difference from the above work since we reveal the correlation between Twitter flu cases and CDC ILI cases at both national and regional level and propose a PDE-based model to predict Twitter flu cases. We adopt the same idea of identifying source tweet as [5] by filtering the retweet/reply of source tweets and also remove tweets from the same user within a certain syndrome elapsed time since they do not indicate new ILI cases.

There are also research focusing on the outbreak of a specific disease. Chunara et al. [13] investigated the role Twitter played during the outbreak of Haitian cholera. It showed strong correlation between Twitter stream data and Health Map data. In addition, it showed a good correlation between Twitter and MSPP (Haitian Ministry of Public Health) data in the initial period of the outbreak. Signorini et al. [12] studied using Twitter data to describe the public interests and concerns with respect to H1N1 and make real-time estimates of ILI activity at national and regional level using support-vector regression. For regional level ILI estimation, it fitted geolocated tweets to CDC region ILI readings from nine of the ten CDC regions to construct a model then used the model to estimate ILI values for the remaining CDC region, Region 2. Our paper is different from this work since: 1) we are interested in general flu cases, not only H1N1, 2) we carry out empirical analysis to correlate Twitter flu counts and CDC ILI cases in regional level, 3) we propose a PDE-based model to predict the Twitter flu counts for all ten regions.

Some research effort focuses on how to filter the noise in Twitter stream data since tweets related to flu but do not report an infection can add noise to the data. Lamb et al. [9] talks about separating flu case signals from flu awareness and concerns and media report through a deeper analysis of the tweet content. Aramaki et al. [14] proposed a support vector machine which is a machine learning based classifier to filter out negative influenza tweets and only extract positive influenza tweets from Twitter stream data. They evaluated their classifier using influenza reports provided by the Japanese Infection Disease Surveillance Center.

Conclusions and future work

Twitter offers an alternative channel to continuously monitor, collect and model influenza trends via flu tweets from real-time Twitter data streams. In this paper, we develop a prototype system to automatically collect, analyze and model flu trends via Twitter data streams. More importantly, we explore the geographical locations from user profiles, tweet location feature that attaches the current user location to a tweet, and the geographical information in the content of the tweets to tag flu tweets with coarse-grained and fine-grained locations. These geo-tagged flu tweets provide an accurate view of the latest flu trends at different regions. With empirical experiments, we correlate geo-tagged flu tweets with CDC statistics on the reported flu cases at both national and regional levels. We also propose PDE-based model to use historical flu tweets to predict future Twitter indicative flu cases. Our future work lies in developing more accurate geo-tagging mechanism and flu case extraction mechanism, extending our framework to implement real-time application which can provide reliable early prediction and warning of flu trends.

Acknowledgements This project is supported by NSF grant CNS #1218212.

References

1. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L., Detecting influenza epidemics using search engine query data. *Nature* 457:1012–1014, 2009.
2. Lazer, D., Kennedy, R., King, G., and Vespignani, A., The parable of google flu: Traps in big data analysis. *Science* 343:1203–1205, 2014.
3. Broniatowski, D.A., Paul, M.J., and Dredze, M., Twitter: Big data opportunities. *Science* 345:148, 2014.
4. Broniatowski, D.A., Paul, M.J., and Dredze, M., National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PLoS one* 8, 2013.
5. Achrekar, H.A., Gandhe, R., Lazarus, S. H.Y. u., and Liu, B.: Predicting Flu Trends using Twitter data, IEEE Conference on Computer Communications Workshop on Cyber-Physical Networking Systems, 2011.
6. http://www.beevolve.com.
7. Dredze, M., Paul, M., Bergsma, S., and Tran, H., Carmen: A Twitter Geolocation System with Applications to Public Health. In: AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI), 2012.
8. http://www.cdc.gov/flu/weekly/regions2008-2009/hhssenusmap.htm.
9. Lamb, A., Paul, M.J., and Dredze, M.: Separating fact from fear: Tracking flu infections on twitter. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2013.
10. Lee, K., Agrawal, A., and Choudhary, A.: Real-Time Disease Surveillance using Twitter Data: Demonstration on Flu and Cancer. Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2013.
11. Lee, K., Agrawal, A., and Choudhary, A.: Real-Time Digital Flu Surveillance using Twitter Data, The 2nd Workshop on Data Mining for Medicine and Healthcare, 2013.

12. Signorini, A., Segre, A.M., and Polgreen, P.M., 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), 2011.
13. Chunara, R., Andrews, J. R., and Brownstein, J. S., Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Am. J. Trop. Med. Hyg.* 86:39–45, 2012.
14. Aramaki, E., Maskawa, S., and Morita, M. *Twitter catches the flu: detecting influenza epidemics using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics*, 2011.
15. Lampos, V., and Cristianini, N. *Tracking the flu pandemic by monitoring the social web. Proceedings of the 2nd International Workshop on Cognitive Information Processing (CIP)*, 2010.