TWITTER IMPROVES SEASONAL INFLUENZA PREDICTION

Harshavardhan Achrekar[1], Avinash Gandhe[2], Ross Lazarus[3], Ssu-Hsin Yu[2] and Benyuan Liu[1]
[1]*Department of Computer Science, University of Massachusetts Lowell, Massachusetts, USA*
[2]*Scientific Systems Company Inc, 500 West Cummings Park, Woburn, Massachusetts, USA*
[3]*Department of Population Medicine, Harvard Medical School, Boston, Massachusetts, USA*
{*hachreka, bliu*}*@cs.uml.edu,* {*avinash.gandhe, ssu-hsin.yu*}*@ssci.com, ross.lazarus@channing.harvard.edu*

Keywords: Flu Trends : Online Social Networks : Prediction.

Abstract: Seasonal influenza epidemics causes severe illnesses and 250,000 to 500,000 deaths worldwide each year. Other pandemics like the 1918 "Spanish Flu" may change into a devastating one. Reducing the impact of these threats is of paramount importance for health authorities, and studies have shown that effective interventions can be taken to contain the epidemics, if early detection can be made. In this paper, we introduce the Social Network Enabled Flu Trends (SNEFT), a continuous data collection framework which monitors flu related tweets and track the emergence and spread of an influenza. We show that text mining significantly enhances the correlation between the Twitter and the Influenza like Illness (ILI) rates provided by Centers for Disease Control and Prevention (CDC). For accurate prediction, we implemented an auto-regression with exogenous input (ARX) model which uses current Twitter data, and CDC ILI rates from previous weeks to predict current influenza statistics. Our results show that, while previous ILI data from CDC offer a true (but delayed) assessment of a flu epidemic, Twitter data provides a real-time assessment of the current epidemic condition and can be used to compensate for the lack of current ILI data. We observe that the Twitter data is highly correlated with the ILI rates across different regions within USA and can be used to effectively improve the accuracy of our prediction. Our age-based flu prediction analysis indicates that for most of the regions, Twitter data best fit the age groups of 5-24 and 25-49 years, correlating well with the fact that these are likely, the most active user age groups on Twitter. Therefore, Twitter data can act as supplementary indicator to gauge influenza within a population and helps discovering flu trends ahead of CDC.

1 INTRODUCTION

Seasonal influenza epidemics result in about three to five million cases of severe illness and about 250,000 to 500,000 deaths worldwide each year (Jordans, 2009). In 1918, the so-called "Spanish flu" killed an estimated 20-40 million people worldwide, and since then, human to human transmission capable influenza virus has resurfaced in a variety of particularly virulent forms much like "SARS", "H1N1" against which no prior immunity exists resulting in a devastating situation with million of casaulties. Reducing the impact of seasonal epidemics and pandemics such as the H1N1 influenza is of paramount importance for public health authorities. Studies have shown that preventive measures can be taken to contain epidemics, if an early detection is made or if we have some form of an early warning system during the germination of an epidemic (Ferguson et al., 2005; Longini et al., 2005). Therefore, it is important to be able to track and predict the emergence and spread of flu in the population.

The Center for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2009) monitors influenza-like illness (ILI) cases by collecting data from sentinel medical practices, collating reports and publishing them on a weekly basis. It is highly authoritative in the medical field but as diagnoses are made and reported by doctors, the system is almost entirely manual, resulting in a 1-2 weeks delay between the time a patient is diagnosed and the moment that data point becomes available in aggregate ILI reports. Public health authorities need to be forewarned at the earliest to ensure effective preventive intervention, and this leads to the critical requirement of more efficient and timely methods of estimating influenza incidences.

Several innovative surveillance systems have been proposed to capture the health seeking behaviour and transform them into influenza activity. Some of them include monitering call volumes to telephone triage advice lines (Espino et al., 2003), over the counter drug sales (Magruder, 2003), patients visit logs to Physicians for flu shots. Google Flu Trends uses aggregated historical log on online web search queries pertaining to influenza to build a comprehensive model that can estimate nationwide ILI activity.

In this paper, we investigate the use of novel data source, Twitter, which takes advantage of the timeliness of early detection to provide snapshot of the current epidemic condition and make influenza related predictions on what may lie ahead, on a daily or even hourly basis. We sought to develop a model which estimates the number of physician visits per week related to ILI as reported by CDC.

Our approach assumes Twitter users within United States as "sensors" and collective message exchanges showing flu symptoms like "I have Flu", "down with swine flu" as early indicators and robust predictors of influenza. We expect these posts on Twitter to be highly correlated to the number of ILI cases in the population. We analyze tweets, build prediction models and discover trends within data to study the characteristics and dynamics of disease outbreak. We validite our model by measuring how well it fits the CDC ILI rates over a course of two years from 2009 to 2011. We are interested in looking at how the seasonal flu spreads within the population across different regions of USA and among different age groups.

In this paper, we extend our preliminary analysis (Achrekar et al., 2011), and provide continuous study of tracking emergence and spread of seasonal flu in the year 2010-2011. Twitter data which demonstrated high correlation with CDC ILI rate last year, was suppressed by spurious messages and so text mining techniques were applied. We show that text mining can significantly enhance the correlation between the Twitter data and the ILI data from CDC, providing a strong base for accurate prediction of ILI rate.

For prediction, we build an auto-regression with exogenous input (ARX) model where ILI rate of previous weeks from CDC forms the autoregressive portion of the model, and the Twitter data serve as exogenous input. Our results show that while previous ILI data from CDC offer a realistic (but delayed) measure of a flu epidemic, Twitter data provides a real-time assessment of the current epidemic condition and can be used to compensate for the lack of current ILI data. We observe that the Twitter data are in fact highly correlated with the ILI data across the different regions within United States.

Our age-based flu prediction analysis indicates that for most of the regions, Twitter data best fit the age groups of 5-24 and 25-49 years, suggesting that these are likely the most active age groups using Twitter. Using fine-grained analysis on user demographics and geographical locations along with its prediction capabilities will provide public health authorities an insight into existing seasonal flu activities.

This paper is organized as follows: Section 2 describes applications that harness the collective intelligence of Online Social Network (OSN) users, to predict real-world outcomes. In Section 3, we give a brief introduction to our data collection and modelling methodolgy. In Section 4, we introduce our data filtering technique for extracting relevant information from Twitter dataset. Detailed data analysis are performed to establish correlation with CDC reports on ILI rates. Then we go one step further and introduce our influenza prediction model in Section 5. In Section 6, we perform Region-wise and Age-based analysis of flu activities in the population based on the Twitter. Finally we conclude in Section 7 and acknowledgements are provided in Section 8.

2 RELATED WORK

A number of studies have been conducted on different forms of social networks like Del.icio.us, Facebook and Wikipedia etc. Ginsberg et al. approach for estimating Flu trends suggests that relative frequency of certain search terms are good indicators of percentage of physician visits in which a patient presents influenza-like symptoms (Ginsberg et al., 2009). Culotta used a document classification component to filter misleading messages out of Twitter and showed that a small number of flu-related keywords can forecast future influenza rates (Culotta, 2010).

Twitter has been used for real-time notifications such as large-scale fire emergencies, earthquake (Sakaki et al., 2010), downtime on services provided by content providers (Motoyama et al., 2010) and live traffic updates. There have been efforts in utilizing twitter data for measuring public interest/concern about health-related events (Signorini et al., 2011), predicting national mood, forecasting box-office revenues for movies (Sitaram and Huberman, 2010), information diffusion in social media (Leskovec et al., 2009), currency tracing, performing market and risk analysis (Jansen et al., 2009) and analysing political tweets to establish the correlations between buzz on Twitter and election results (Nardelli, 2010) etc.

3 DATA COLLECTION

We describe our data collection methodology by introducing SNEFT architecture, provide description of our dataset, explore strategies for data cleaning, apply filtering techniques in order to perform quantitative spatio-temporal analysis.

3.1 SNEFT Architecture

We propose Social Network Enabled Flu Trends (SNEFT) architecture along with its crawler, predic-

tor and detector components, as our solution to predict flu activity ahead of time with certain accuracy. CDC ILI reports and other influenza related data are

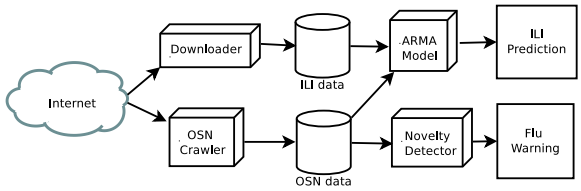

Figure 1: The system architecture of SNEFT.

downloaded into "ILI Data" database from their corresponding websites (e.g., CDC (Centers for Disease Control and Prevention, 2009)). A list of flu related keywords ("Flu" , "H1N1" and "Swine Flu") that are likely to be of significance are used by OSN Crawler as inputs into public search interfaces to retrieve publicly available posts having mention of those keywords. Relevant information about the posts are collected along with the relative keyword frequency and stored in a spatio-temporal "OSN Data" database for further data analysis.

Autoregressive Moving Average (ARMA) model is used to predict ILI incidence as a linear function of current and past OSN data and past ILI data thus providing a valuable "preview" of ILI cases well ahead of CDC reports. Novelty detection techniques can be used to continuously monitor OSN data, and detect transition in real time from a "normal" baseline situation to a pandemic using the volume and content of OSN data enabling SNEFT to provide a timely warning to public health authorities for further investigation and response.

3.2 Twitter Crawler

In this section we briefly describe the methodology for collecting our dataset. Based on the search API provided by Twitter, we develop crawlers to fetch data at regular time intervals.

The twitter search service accepts single or multiple keywords using conjunctions ("flu" OR "h1n1" OR "#swineflu") to search for relevant tweets. Search results are typically 15 tweets (maximum 50) per page up to 1,500 tweets arranged in chronologically decreasing order, obtained from a real time stream known as the public timeline. The tweet has the User Name, the Post with status id and the Timestamp attached with each post. From the twitter username, we can get the number of followers, number of friends, his/her profile creation date, location and status update count for every user. The location field helps us in tracking the current/default location of a user. Geo location codes are present in a location enabled mobile tweet. For all other purposes, we assume the location attribute within the profile page to be his/her current location and pass it as an input to Google's location based web services to fetch geo-location codes (i.e., latitude and longitude) along with the country, state, city with a certain accuracy scale. All the data extracted from posts and profile page are stored in a spatio-temporal "OSN data" Database.

We apply filters to get quantitative data within Unites States and exclude organizations and users who posts multiple times during the day on flu related activities. This data is fed into the Analysis Engine which has a detector and ARMA predictor model. The visualization tools and reporting services generate timely visual and data centric reports on the ILI situation. CDC monitors Influenza-like illness cases within USA by collecting data about number of Hospitalizations, percentages weighted ILI visits to physicians etc and publishes it online. We download the CDC data into "ILI data" database to compare our results.

4 DATA SET

In this section we briefly describe our datasets used for influenza prediction. Since Oct 18, 2009, we have searched and collected tweets and profile details of Twitter users who mentioned about flu descriptors in their tweets. The preliminary analysis for the year 2009-2010 is documented in (Achrekar et al., 2011). For 2010-2011, so far we have 4.5 million tweets from 1.9 million unique users. Twitter allows its users to set their location details to public or private from the profile page or mobile client. So far our analysis on location details of Twitter dataset suggest that 22% users on Twitter are within USA, 46% users are outside USA and 32% users have not published their location details.

Initial stage analysis for the period 2009-2010, indicated a strong correlation between CDC and Twitter data on the flu incidences (Achrekar et al., 2011). However results for the year 2010-2011 showed a significant drop in the correlation coefficient from 0.98 to 0.47. In an attempt to investigate such a drastic drop in correlation we looked at data samples and found spurious messages which suppressed the actual data. To list a few, tweets like "I got flu shot today.", "#nowplaying Vado - Slime Flu..i got one recently!" (slime flu is the name of a debut mixtape from an artist V.A.D.O. released in 2010) are false alarms of flu. In the year 2009-2010, swine flu event was so evident that the noise did not significantly affect the correlation that existed then. To mitigate this prob-

lem, we removed the spurious tweets using a filtering technique that trains a document classifier to label whether a message is indicative of flu event or not.

4.1 Text Classification

In an information retrieval scenario, text mining seeks to extract useful information from unstructured textual data. Using simple "bag-of-words" text representations technique based on vector space, our algorithm classifies tweets wherein user mentions about having acquired flu himself or having observered flu among his friends, family, relatives, etc. Accuracy of such a model is highly dependent on how well trained our model is, in terms of precision, recall and F-measure.

The set of possible labels for a given instance can be divided into two subsets, one of which are considered "relevant". To create such an annotated dataset which demands human intelligence, we use Amazon Mechanical Turks to manually classify a sample of 25,000 tweets. Every tweet is classified by exactly three Turks and the majority classified result is attached as the final class for that tweet.

The training dataset is fed as an input to different classifiers namely decision tree (J48), Support Vector Machines (SVM) and Naive Bayesian. For efficient learning, some configurations that we did incorporate within our text classification algorithm includes setting term frequency and inverse document frequency (tf-idf) weighting scheme, stemming, using stopwords list, limiting number of words to keep (feature vector set) and reordering class. Based on the results shown in Table 1, we conclude that SVM classifier with highest precision and recall rate outperforms other classifiers when it comes to text classification for our data set. Application of SVM on unclassified data originating from within USA resulted in Twitter dataset with 280K positively classified tweets from 187K unique twitter users. In order to gauge if the

Classifier	Class	Precision	Recall	F-value
J48	Yes	0.801	0.791	0.796
	No	0.813	0.704	0.755
Naive Bayesian	Yes	0.725	0.829	0.773
	No	0.813	0.704	0.755
SVM	Yes	**0.807**	**0.822**	**0.814**
	No	**0.829**	**0.814**	**0.822**

Table 1: Text Classification 10 fold cross validation results

number of unique twitter users mentioning about flu per week is a good measure of the CDC's ILI reported data, we plot (in Figure 2) the number of twitter users/week against the percentage of weighted ILI visits, which yields a high Pearson correlation coefficient of 0.8907.

Thus increase in the users tweeting about flu is accompanied by increase in percentage of weighted ILI visits reported by CDC in the same week. However the marked outlier present in Twitter data as identified in Figure 2 is coherent with Google Flu Trends data when high tweet volume were witnessed in the week starting January 2, 2011.

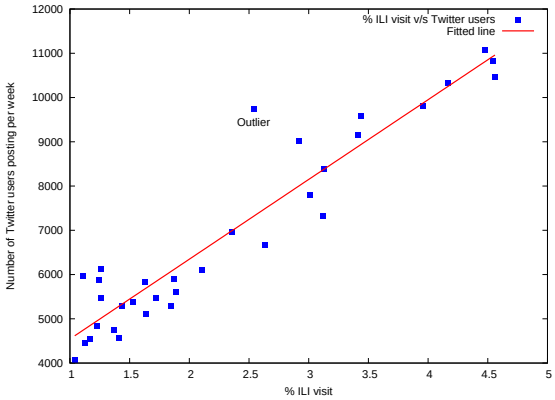

Figure 2: Number of Twitter users per week versus percentage of weighted ILI visit by CDC

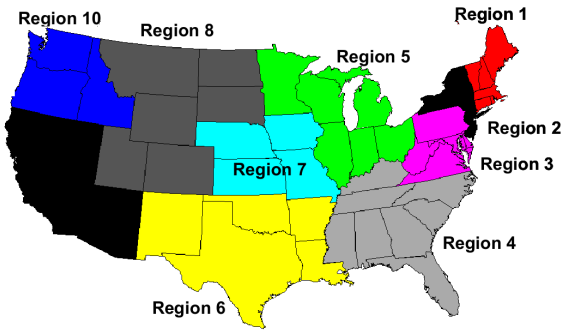

Figure 3: Regionwise Division of USA into ten Regions.

CDC has divided USA into 10 regions as shown in Figure 3. CDC publishes their weekly reports on percentage weighted ILI visits collated from its ten regions and aggregates for USA. Figure 4 compares the Twitter dataset with CDC reports with and without text classification for each of the ten regions defined by CDC and USA as a whole. We observe that the correlation coefficients have significantly improved with text classification, across all the regions and USA overall. Thus our text classification techniques plays a vital role in improving the overall detection and prediction performance.

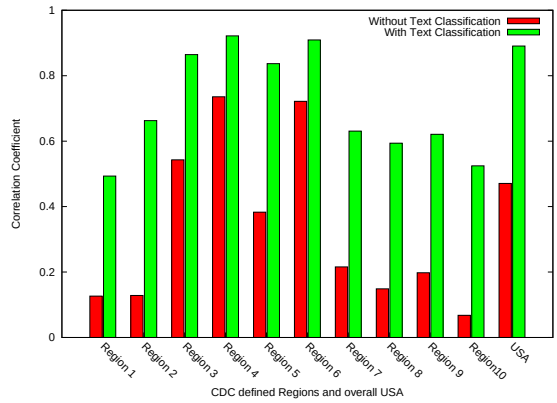

Figure 4: Classified Twitter dataset achieves higher correlation with CDC reports on Nationwide and Regional levels.

4.2 Data Cleaning

The Twitter dataset required data cleaning to discount retweets and successive posts from same users within syndrome elapsed time.

- *Retweets*: A retweet is a post originally made by one user that is forwarded by another user. For flu tracking, a retweet does not indicate a new ILI case, and thus should not be counted in the analysis. Out of 4.5 million tweets we collected, there are 541K retweets, accounting for 12% of the total number of tweets.
- *Syndrome elapsed time*: An individual patient may have multiple encounters associated with a single episode of illness (e.g., initial consultation, consultation 1–2 days later for laboratory results, and follow-up consultation a few weeks later). To avoid double counting from common pattern of ambulatory care, the first encounter for each patient within any single syndrome group is reported to CDC, but subsequent encounters with the same syndrome are not reported as new episodes until more than six weeks have elapsed since the most recent encounter in the same syndrome (Lazarus et al., 2002). We call this Syndrome Elapse time.

Hence, we created different datasets namely: Twitter dataset with No Retweets (Tweets starting with RT) and Twitter dataset without Retweets and with no tweets from same user within certain syndrome elapsed time.

When we compared different datasets mentioned in Table 2 with CDC data, we found that Twitter dataset without Retweets showed a high correlation (0.8907) with CDC Data. As opposed to a common practice in public health safety, where medical examiners within U.S. observe a syndrome elapse time period of six weeks, user behaviour on Twitter follows a

Retweets	Syndrome Elapse Time	Correlation coefficient	RMSE errors
No	**0 week**	**0.8907**	**0.3796**
No	1 week	0.8895	0.3818
No	2 week	0.8886	0.3834
No	3 week	0.886	0.3878
No	4 week	0.8814	0.3955

Table 2: Correlation between Twitter Dataset and CDC along with its Root Mean Square Errors(RMSE).

trend wherein we do not ignore successive posts from same user. Thus Twitter dataset without Retweets is our choice of dataset for all subsequent experiments.

From Figure 5, we observe that Complementary Cumulative Distribution function (CCDF) of the number of tweets posted by same individual can be fitted by a power law function of exponent -2.6429 and coefficient of determination (R-square) 0.9978 with a RMSE of 0.1076 using Maximum likelihood estimation. Most people tweet very few times (e.g., 82.5% of people only tweet once and only 6% of people tweet more than two times).

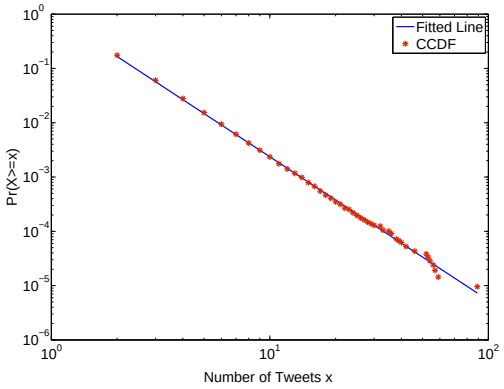

Figure 5: Complementary Cumulative Distribution function (CCDF) of the number of tweets by same users.

Most of these high-volume tweets are created by health related organization, who tweet multiple time during a day and users who subscribe to flu related RSS feeds published by these organizations. “Flu_alert”,“swine_flu_pro”, “live_h1n1”, “How_To_Tips”, “MedicalNews4U” are examples of such agencies on Twitter.

5 PREDICTION MODEL

The correlation between Twitter activity and CDC reports can change due to a number of factors. Annual or seasonal changes in flu-related trends, for instance vaccination rates that are affected by health cares, result in the need to constantly update parameters relat-

ing Twitter activity and flu activity. However, particularly at the beginning of the influenza season, when prediction is of most significance, enough data may not be available to accurately perform these updates. Additionally predicting changes in ILI rates simply due to changes in flu-related Twitter activity can be risky due to transient changes, such as changes in Twitter activity due to flu-related news.

In order to establish baseline for the ILI activity and to smooth out any undesired transients, we propose the use of Logistic Autoregression with exogenous inputs (ARX). Effectively, we attempt to predict a CDC ILI statistic during a certain week by using Twitter activity and CDC data from previous weeks. The prediction of current ILI activity using ILI activity from previous weeks forms the autoregressive portion of the model, while the Twitter data from previous weeks serve as exogenous inputs. By CDC data, we refer to the percentage of visits to a physician for ILI (also called as ILI rate).

5.1 Influenza Model Structure

Although the percentage of physician visits is between 0% and 100%, the number of Twitter users is bounded below by 0. Simple Linear ARX neglects this fact in the model structure. Therefore, we introduce a logit link function for CDC data and a logarithmic transformation of the Twitter data as follows:

Logistic ARX Model

$$\log\left(\frac{y(t)}{1-y(t)}\right) = \sum_{i=1}^{m} a_i \log\left(\frac{y(t-i)}{1-y(t-i)}\right) + \sum_{j=0}^{n-1} b_j \log(u(t-j)) + c + e(t) \quad (1)$$

where t indexes weeks, $y(t)$ denotes the percentage of physician visits due to ILI in week t, $u(t)$ represents the number of unique Twitter users with flu related tweets in week t, and $e(t)$ is a sequence of independent random variables. c is a constant term to account for offset. In our tests, the number of unique Twitter users $u(t)$ is defined as Twitter users without retweets and having no tweets from the same user within syndrome elapsed time of 0 week. The flu related tweets are defined as tweets with keywords "flu", "H1N1" and "swine flu". The rationale for the model structure in Eq. (1) is that Twitter data provides real-time assessment of flu epidemic. However, the Twitter data may be disturbed at times by events related to flu, such as news reports of flu in other parts of the world, but not necessarily to local people actually getting sick due to ILI. On the other hand, the CDC data provides a true, albeit delayed, assessment of a flu epidemic. Hence, by using the CDC data along with the Twitter data, we may be able to take advantage of the timeliness of the Twitter data while overcoming the disturbance that may be present in the Twitter data.

The objective of the model is to provide timely updates of the percentage of physician visits. To predict such percentage in week t, we assume that only the CDC data with at least 2 weeks of lag is available for the prediction, if past CDC data is present in a model. The 2-week lag is to simulate the typical delay in CDC data reporting and aggregation. For the Twitter data, we assume that the most recent data is always available, if a model includes the Twitter data terms. In other words, the most current CDC or Twitter data that can be used to predict the percentage of physician visits in week t is week t-2 for the CDC data and week t for the Twitter data.

In order to predict ILI rates in a particular week given current Twitter data and the most recent ILI data from the CDC we must estimates the coefficients, a_i , b_j and c in Eq. (1). Also, in practice, the model orders m and n are unknown and must be estimated. In our experiment, we vary m from 0 to 2 and n from 0 to 3 in Eq. (1) in order to obtain the best values of m and n to use for prediction. Intuitively, this answers the question of how many weeks of Twitter and ILI data should be used to predict the ILI activity in the current week. Within the ranges examined, m = 0 or n = 0 represent models where there are no CDC data, y, or Twitter data, u, terms present. Also, if m = 0 and n = 1, we have a linear regression between Twitter data and CDC data. If n = 0, we have standard auto-regressive (AR) models. Since the AR models utilize past CDC data, they serve as baselines to validate whether Twitter data provides additional predictive power beyond historical CDC data.

Prediction with Logistic ARX Model To predict the flu cases in week t using the Logistic ARX model in Eq. (1) based on the CDC data with 2 weeks of delay and/or the up-to-date Twitter data, we apply the following relationship:

$$\log\left(\frac{\hat{y}(t)}{1-\hat{y}(t)}\right) = a_i \log\left(\frac{\hat{y}(t-1)}{1-\hat{y}(t-1)}\right) + \sum_{i=2}^{m} a_i \log\left(\frac{y(t-i)}{1-y(t-i)}\right) + \sum_{j=0}^{n-1} b_j \log(u(t-j)) \quad (2)$$

$$\log\left(\frac{\hat{y}(t-1)}{1-\hat{y}(t-1)}\right) = \sum_{i=1}^{m} a_i \log\left(\frac{y(t-i-1)}{1-y(t-i-1)}\right) + \sum_{j=0}^{n-1} b_j \log(u(t-j-1)) \quad (3)$$

where $\hat{y}(t)$ represents predicted CDC data in week t.

It can be verified from the above equations that to predict the CDC data in week t, the most recent CDC data is from week $t-2$. If the CDC data lag is more or less than two weeks, the above equations can be easily adjusted accordingly.

5.2 Cross Validation Test Description

Based on ARX model structure in Eq. (1), we conducted tests using different combinations of m and n values. We currently have 33 weeks with both Twitter activity and CDC data available (10/3/2010–05/15/2011). Due to limited data samples, we adopted the K-fold cross validation approach to test the prediction performance of the models.

In a typical K-fold cross validation scheme, the dataset is divided into K (approximately) equally sized subsets. At each step in the scheme, one such subset is used as the test set while all other subsets are used as training samples in order to estimate the model coefficients. Therefore, in a simple case of a 30-sample dataset, 10-fold cross-validation would involve testing 3-samples in each step, while using the other 27 samples to estimate the model parameters.

In our case, the cross-validation scheme is somewhat complicated by the dependency of the sample $y(t)$ on the previous samples, $y(t-1),\ldots,y(t-m)$ and $u(t),\ldots,u(t-n+1)$ (see Eq. (1)). Therefore, the first sample that can be predicted is $y(\max(m+1,n))$ not $y(1)$. In fact, since we are predicting "two weeks ahead" of the available CDC data, the first sample that can be estimated is actually $y(\max(m+2,n+1))$. Since, prediction equations cannot be formed for $y(1),\ldots,y(\max(m+2,n+1)-1)$, those samples were not considered in any of the K subsets during our experiment to be evaluated for prediction performance. However, they were still used in the training set to estimate the values of the coefficients a_i and b_j in Eq. (1).

Considering the above constraints, our K-fold validation testing procedure is as follows:

1. For each (m,n) pair from $m=0,1,2$ and $n=0,1,2,3$, repeat the following:
   (a) Identify F, the index of first data sample that can actually be predicted. $F=max(m+1,n)$
   (b) Represent the available data indices as $t=1,\ldots,T$. Then divide the dataset into K approximately equally sized subsets $\{S_1,S_2,\ldots,S_K\}$, with each subset comprising members that have an approximately equal time interval between them. For example, the first set would be $S_1=\{y(F),y(F+K),y(F+2K),\ldots\}$, the second would be $S_2=\{y(F+1),y(F+K+1),y(F+2K+1),\ldots\}$ and so on.
   (c) For each S_k, $k=1,\ldots,K$, obtain the values of the model parameters a_i and b_j using all the other subsets with the least squares estimation technique. Based on the estimated model parameter values and the associated prediction equations in Eq. (2), predict the value of each member of S_k.
2. For each (m,n) pair, we have obtained a prediction of the CDC time-series, $y(t)$ for $t=F_{mn},\ldots,T$. Note that F still represents the first time index that can be predicted. However, we use the subscript mn to emphasize the fact that F varies depending on the values of m and n. By comparing the prediction with the true CDC data, we calculate the root mean-squared error (RMSE) as follows:

$$\varepsilon=\sqrt{\frac{1}{T-F_{\max}+1}\sum_t(y(t)-\hat{y}(t))^2} \qquad (4)$$

   The RMSE is computed over $t=F_{\max},\ldots,T$, regardless of techniques and model orders to ensure fairness in comparison.

	$n=0$	$n=1$	$n=2$	$n=3$
$m=0$		0.5355	0.4814	0.4813
$m=1$	0.6331	0.4107	0.4147	0.4314
$m=2$	0.5395	**0.3957**	0.3986	0.4256

Table 3: Root mean squared errors from 10-fold cross validation. m and n are defined in Eq. (1). The m and n values in the table specify the model that results in the RMSE in the corresponding row and column respectively. The lowest RMSE in the table is highlighted.

5.3 Cross Validation Results

According to the 10-fold cross validation results in Table 3, the model corresponding to $m=2$ and $n=1$ has the lowest RMSE. This indicates that current Twitter data and two most recent ILI data points are most useful in accurate prediction of influenza rates. In general, the addition of Twitter data improves the prediction with past CDC data alone. For the 10-fold cross validation results presented in Table 3, for example, the AR model $(m=1,n=0)$ comprising of the $y(t-2)$ term and the constant term for the prediction of $y(t)$ has a RMSE of 0.6331. For the same $m=1$, the model with additional Twitter data $u(t)$ (i.e. $n=1$) has a lower RMSE of 0.4107. We also observe that using Twitter data ($m=0$) alone is insufficient for prediction and that the past ILI rates are critical in predicting future values, as is evident from our results. The addition of Twitter data improves the prediction with past CDC data alone. Therefore, the Twitter data provides a real-time assessment of the flu

epidemic (i.e. the availability of Twitter data in week t in the prediction of physician visits also in week t as shown in Eq. (2)), while the past CDC data provides the recent ILI rates in the prediction model. As shown earlier in the paper, there is strong correlation between the Twitter data and the CDC data. Hence, the more timely Twitter data can compensate for the lack of current CDC data and help capture the current flu trend. Finally in Figure 6, we provide a sin-

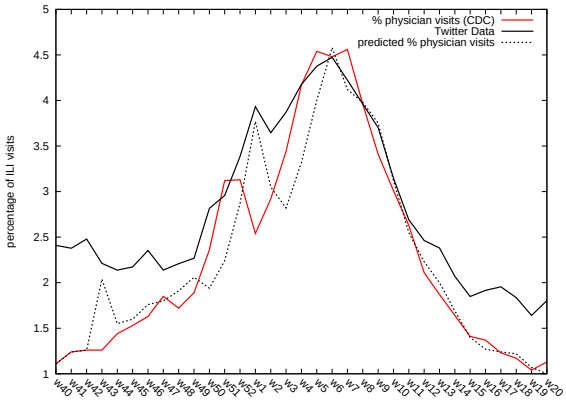

Figure 6: Weekly plot of percentage weighted ILI visits, positively classified Twitter dataset and predicted ILI rate using CDC and Twitter

gle plot for percentage weighted ILI visits, positively classified Twitter users and predicted ILI rate using CDC and Twitter for the year 2010-2011. Note that the original Twitter data alone would predict higher ILI rates for the begining and ending parts of the flu season. Using previous ILI data from CDC offers a better assessment for making flu predictions.

6 FLU PREDICTION WITHIN REGIONS AND AGE GROUPS

In this section we discuss the use of Twitter for flu predictions in specific population groups. Given the data available, we are able to study the prediction performance in specific regions of the United States. Also, with ILI rates provided in different age groups we are able to study the effectiveness of using Twitter data to predict flu trends in these age groups. The advantages of studying performance in subgroups are twofold:

- The differences in Twitter usage among different population groups and similar differences in response amongst people in different population groups to ILI-like symptoms can result in very different model parameters and prediction performance when attempting to predict flu activity among different sections of the population. It is therefore important to adapt the prediction models for different population groups.
- In our previous study, it has been shown that there exists significant correlation between Twitter reports and the percentage of ILI cases reported by CDC. However, much of our analysis is based on a limited number of data points (31 overlapping weeks for Twitter and CDC reports for the year 2009-2010 and 33 overlapping weeks for Twitter and CDC reports for the year 2010-2011) available during our period of performance evaluation, with Twitter and ILI data aggregated across the entire United States. In the year 2009-2010, only 11 out of 31 data points occurred during the weeks where the ILI rates were significant (>2%) and during this interval, the ILI rates and Twitter reports were steadily decreasing. During the period 2010-2011, 15 out of 33 data points occurred during the weeks where the ILI rates were significant (>2%) and during this interval, the ILI rates and Twitter reports were simultaneously increasing till they reached their peak in mid February 2011 and then onwards they both started decreasing.

Due to this limited time frame any claim of high correlation between the two data streams (ILI rates and Twitter reports) may be viewed with skepticism. This evaluation was performed as an experiment to see which age groups the Twitter data fit best. The results are interesting but not conclusive.

6.1 Regional Twitter and ILI rates

We analyzed the relationship between the Twitter activity and ILI rates across all geographic regions defined by the Health and Human Services (HHS) regions. For reference, the regions are shown on the USA map in Figure 3.

In studying the regional statistics, we would like to make some comparisons across regions. For instance (i) when the ILI rate peaks later in a particular region than the rest of country, do the Twitter reports also peak later, (ii) is there in relationship between the decay in ILI rates and the decay in Twitter reports.

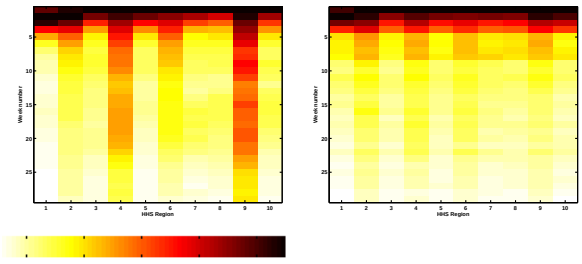

Figure 7: Heatmap of CDC's Regionwise ILI data (left) and Twitter data (right). Colormap scale included (below).

Figure 7 shows, for both ILI and Twitter data, the relative intensity across the ten Health and Human Services (HHS) regions (columns) during successive weeks (rows) in the year 2009-2010. The colormap used is a scale with white representing low intensity and black, high intensity. We are comparing "trends" among the ILI and Twitter data.

Regional analysis shows that ILI seems to peak later in the Northeast (Regions 1 and 2) than in the rest of the country by at least week. The Twitter reports also follow this trend. In Region 9, Region 4 and the Northeast, the ILI rates seem to drop off fairly slowly in the weeks immediately following the peaks. This is also reflected in the Twitter reports. Approximately 20-25 weeks after the peak ILI, the northern regions have lower levels relative to the peaks in the southern regions. This is also true of the Twitter reports. The decline in ILI rates is slowest in Region 9.

Figure 8 depicts regionwise ILI prediction performance for the year 2010-2011 using our logit model. We arbitrarily select region 1, region 6 and region 9 to represent the regions, one each from the East, South and Western U.S. and plot the true and predicted ILI values for each of these regions. We observe that

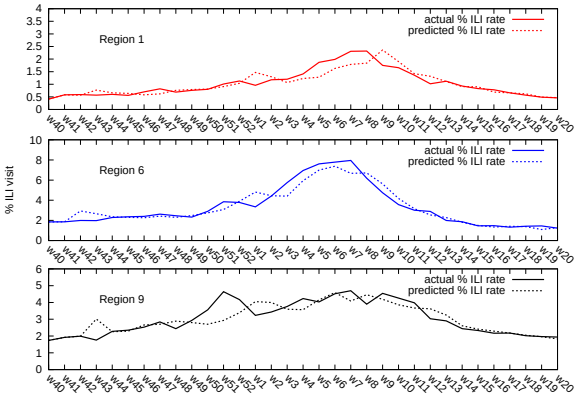

Figure 8: Comparision between Actual and Predicted regional data for Region 1, Region 6 and Region 9.

the Twitter reports and ILI rates are in fact correlated across regions and therefore corroborate our earlier findings that Twitter can improve ILI rate prediction.

6.2 Age-based Influenza Analysis

The differences in Twitter usage and susceptibility to flu among different demographics can result in very different prediction model parameters and performance when attempting to predict flu activity among different sections of the population. While any number of population groups may be defined, the CDC provides the number of ILI cases by age groups, from which we can compute the unweighted ILI rates. This then provides an opportunity to examine the prediction performance amongst different age groups when predicting ILI using Twitter data. Note that while ILI rates broken down by age group are available, we do not have Twitter activity broken down by age group. Also, it is debatable whether attempting to correlate Twitter and ILI activity within age groups is of any value; a significant percentage of Twitter activity may result from family members or friends of the affected persons. Therefore, we attempt to study the relationship between aggregate Twitter activity over all age groups with ILI rates in different age groups.

Table 4 shows the Root Relative Squared Error (RRSE) performance in different age groups for different geographical regions within USA. The RRSE normalizes the errors to the magnitude of the ground truth data (in this case the total number of ILI cases relative to total patients seen by provider) in each age group. We have highlighted the age groups with the best match between ILI rates and Twitter data within each region. In parenthesis, alongside the RRSE values are the model orders for the autoregressive and x-components of the general model, (m-n). The "best" age-group for prediction in each region is highlighted.

The results indicates that for most of the regions, Twitter data best fits the age-groups of 5-24 yrs and 25-49 yrs, which correlates well with the fact that this likely is the most active age groups using Twitter (Twitter, 2011). For Region 6 and 7, the Twitter activity best fits ILI activity amongst the 0-4 yrs age group. This is an interesting result which we currently have no specific insight into. It should be noted that for Region 6 and 7, the difference between the fits for 0-4 years and 25-49 years is marginal.

	$0-4yrs$	$5-24yrs$	$25-49yrs$	$50+yrs$
US	0.5285(0-2)	0.4261(2-2)	**0.3577(1-2)**	0.4320(1-1)
Reg1	0.5728(2-1)	0.6000(2-2)	**0.5499(1-1)**	0.7763(1-1)
Reg2	0.6954(0-3)	0.6005(2-1)	**0.4965(0-3)**	0.5171(1-3)
Reg3	0.4423(0-2)	0.3268(2-2)	**0.3066(2-3)**	0.3515(1-2)
Reg4	0.5281(0-3)	**0.3719(0-1)**	0.4792(0-1)	0.5192(0-1)
Reg5	0.6387(1-1)	0.4337(2-3)	**0.4300(0-3)**	0.5198(1-1)
Reg6	**0.3032(0-2)**	0.3407(1-2)	0.3564(0-3)	0.4469(0-3)
Reg7	**0.5426(2-3)**	0.5571(1-3)	0.5492(1-3)	0.6454(2-2)
Reg8	0.6511(1-1)	**0.6133(1-2)**	0.6649(2-2)	0.6445(2-3)
Reg9	0.7453(2-1)	**0.4229(2-1)**	0.4690(1-1)	0.6176(2-1)
Reg10	0.8548(2-1)	**0.5746(2-1)**	0.6462(2-2)	0.7347(2-1)

Table 4: Prediction performance (root relative squared error) using Twitter in different age groups for different geographical regions within the US. In parenthesis, alongside the RRSE values are the model orders, (m-n), for the autoregressive and x-components of the general model in Eq. (1) which yield the best performance.

The above results show that flu-related Twitter activity is more correlated with flu activity with certain

age-groups within the USA population and the correlation may be better in certain regions compared to others. This does indicate that training prediction models that are targeted to specific population segments is a worthwhile endeavor in a future effort.

7 CONCLUSIONS

In this paper, we have described our approach to achieve faster, near real time detection and prediction of the emergence and spread of influenza epidemic, through continuous tracking of flu related tweets originating within United States. We showed that applying text classification on the flu related tweets significantly enhances the correlation (Pearson correlation coefficient 0.8907) between the Twitter data and the ILI rates from CDC.

For prediction, we build an auto-regression with exogenous input (ARX) model where ILI rate of previous weeks from CDC formed the autoregressive portion of the model, and the Twitter data served as an exogenous input. Our results indicated that while previous ILI rates from CDC offered a realistic (but delayed) measure of a flu epidemic, Twitter data provided a real-time assessment of the current epidemic condition and can be used to compensate for the lack of current ILI data.

We observed that the Twitter data was highly correlated with the ILI rates across different HHS regions. Our age-based prediction analysis suggested that for most of the regions, Twitter data best fit the age groups of 5-24 years and 25-49 years, correlating well with the fact that these were likely the most active age group communities on Twitter. Therefore, flu trends tracking using Twitter significantly enhances public health preparedness against influenza epidemic and other large scale pandemics.

8 ACKNOWLEDGEMENTS

This research is supported in parts by the National Institutes of Health under grant 1R43LM010766-01 and National Science Foundation under grant CNS-0953620.

REFERENCES

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. (2011). Predicting flu trends using twitter data. *IEEE Infocom, 2011 workshop on on Cyber-Physical Networking Systems (CPNS) 2011*.

Centers for Disease Control and Prevention (2009). FluView, a weekly influenza surveillance report.

Culotta, A. (2010). Detecting influenza outbreaks by analyzing twitter messages. *Knowledge Discovery and Data Mining Workshop on Social Media Analytics, 2010*.

Espino, J., Hogan, W., and Wagner, M. (2003). Telephone triage: A timely data source for surveillance of influenza-like diseases. In *AMIA: Annual Symposium Proceedings*.

Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., and Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437:209–214.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Jansen, B., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power:tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(1532):2169–2188.

Jordans, F. (2009). WHO working on formulas to model swine flu spread.

Lazarus, R., Kleinman, K., Dashevsky, I., Adams, C., Kludt, P., DeMaria, A., Jr., R., and Platt (2002). Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events.

Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *International Conference on Knowledge Discovery and Data Mining, Paris, France*, 495(978).

Longini, I., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D., and Halloran, M. (2005). Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087.

Magruder, S. (2003). Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. In *Johns Hopkins University APL Technical Digest*.

Motoyama, M., Meeder, B., Levchenko, K., Voelker, G. M., and Savage, S. (2010). Measuring online service availability using twitter. *Workshop on online social networks, Boston, Massachusetts, USA*.

Nardelli, A. (2010). Tweetminister.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *19th international conference on World wide web, Raleigh, North Carolina, USA*.

Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). 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, Volume 6 — Issue 5*.

Sitaram, A. and Huberman, B. A. (2010). Predicting the future with social media. In *Social Computing Lab, HP Labs, Palo Alto, California, USA*.

Twitter (2011). Information on twitter users age-wise.