Online Social Networks Flu Trend Tracker: A Novel Sensory Approach to Predict Flu Trends

Harshavardhan Achrekar\textsuperscript{1}, Avinash Gandhe\textsuperscript{2}, Ross Lazarus\textsuperscript{3}, Ssu-Hsin Yu\textsuperscript{2}, and Benyuan Liu\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, University of Massachusetts Lowell, Massachusetts, U.S.A.
\textsuperscript{2}Scientific Systems Company Inc, 500 West Cummings Park, Woburn, Massachusetts, U.S.A.
\textsuperscript{3}Department of Population Medicine, Harvard Medical School, Boston, Massachusetts, U.S.A.

Abstract. Seasonal influenza epidemics cause several million cases of illnesses cases and about 250,000 to 500,000 deaths worldwide each year. Other pandemics like the 1918 “Spanish Flu” may change into devastating event. 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 Social Network Enabled Flu Trends (SNEFT), a continuous data collection framework which monitors flu related messages on online social networks such as Twitter and Facebook and track the emergence and spread of an influenza. We show that text mining significantly enhances the correlation between online social network(OSN) data 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 OSN data and CDC ILI rates from previous weeks to predict current influenza statistics. Our results show that, while previous ILI data from the CDC offer a true (but delayed) assessment of a flu epidemic, OSN 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 OSN 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. Therefore, OSN data can act as supplementary indicator to gauge influenza within a population and helps to discover 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 [11]. 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” and “H1N1” against which no prior immunity exists, resulting in a devastating situation with severe casualties. 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\[7\,14\]. 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)\[3\] 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 need 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. These include monitoring call volumes to telephone triage advice lines\[6\], over the counter drug sales\[15\], and patients visit logs on 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\[9\].

In this paper, we investigate the use of a novel data source, OSN data, which takes advantage of the timeliness of early detection to provide a snapshot of the current epidemic conditions and makes 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 treats OSN users within United States as “sensors” and collective message exchanges showing flu symptoms like “I have Flu”, “down with swine flu”, etc. - as early indicators and robust predictors of influenza. We expect these posts on OSN’s to be highly correlated to the number of ILI cases in the population. We analyze messages, build prediction models and discover trends within data to study the characteristics and dynamics of disease outbreak. We validate our model by measuring how well it fits the CDC ILI rates over the 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\[1,2\], and provide a continuing study of using OSN’s to track the emergence and spread of seasonal flu in the year 2010-2011. OSN data which demonstrated high correlation with CDC ILI rate for the year 2009-2010, was affected by spurious messages and so text mining techniques were applied. We show that text mining can significantly enhance the correlation between the OSN 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 rates of previous weeks from CDC form the autoregressive component of the model, and the OSN 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, OSN 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 OSN data are in fact highly correlated with the ILI data across the different regions within United
States. Using fine-grained analysis on user demographics and geographical locations along with prediction capabilities will provide public health authorities an insight into current 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 modeling framework. In Section 4, we introduce our data filtering technique for extracting relevant information from the Twitter and Facebook datasets. Detailed data analysis is 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 analysis of flu activities in the population based on the Twitter and Facebook. Finally we conclude in Section 7 and acknowledgements are provided in Section 8.

2 Related Work

A number of measurement related studies have been conducted on different forms of social networks like Del.icio.us, Facebook and Wikipedia etc [8,22]. Sitaram et al. demonstrated how social media content like chatter from Twitter can be used to predict real-world outcomes of forecasting box-office revenues for movies [21]. Sakaki et al. used a probabilistic spatio-temporal model to build an autonomous earthquake reporting system in Japan using twitter users as sensors and applying Kalman filtering and particle filtering for location estimation [19]. Meme Tracking in news cycles as explained by Leskovec et al. was an attempt to model information diffusion in social media like blogs and tracking handoff from professional news media to social networks [13].

Ginsberg et al. in his paper discussing his approach for estimating Flu trends proposed that the relative frequency of certain search terms are good indicators of the percentage of physician visits and established a linear correlation to weekly published ILI percentages between 2003 and 2007 for all nine regions identified by CDC [9]. 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 [5].

OSN data has been used for real-time notifications such as large-scale fire emergencies, downtime on services provided by content providers [17] and live traffic updates. There have been efforts in utilizing twitter data for measuring public interest/concern about health-related events [18,20], predicting national mood [16], currency tracing and performing market and risk analysis [10]. Tweetminster, a media utility tool design to make UK politics open and social, analyses political tweets, to establish the correlations between buzz on Twitter and election results. In June 2010, we introduced the SNEFT architecture as a continuous data collection engine which combines the detection and prediction capability on social networks in discovering real world flu trends [1,2,4].

3 Data Collection

In this section we describe our data collection methodology by introducing the SNEFT architecture, provide a description of our dataset, explore strategies for data cleaning, and apply filtering techniques in order to perform quantitative spatio-temporal analysis.
3.1 SNEFT Architecture

We propose the Social Network Enabled Flu Trends (SNEFT) architecture along with its crawler, predictor and detector components, as our solution to predict flu activity ahead of time with a certain accuracy.

CDC ILI reports and other influenza related data are downloaded into the “ILI Data” database from their corresponding websites (e.g., CDC [3]). A list of flu related keywords (“Flu”, “H1N1” and “Swine Flu”) that are likely to be of significance are used by the OSN Crawler as inputs into public search interfaces to retrieve publicly available posts mentioning those keywords. Relevant information about the posts such as time, location and other demographic information is collected along with the relative keyword frequency and stored in a spatio-temporal “OSN Data” database for further data analysis.

An Autoregressive with Exogenous input (ARX) 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 OSN Crawler

Based on the search API provided by Twitter and Facebook, we have developed 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 “Post by everyone” option allows us to search public posts for given keywords in Facebook. All results that show up are available to the public for a limited time period. We are interested in getting useful information (profile ID, time stamp of the post, and the post content) out of posts. Given a profile ID, we will retrieve the detailed
information of the profile, which typically includes, among other things, name, gen-
der, age, affiliations (school, work, region), birthday, location, education history, and
friends.

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/post. For all other pur-
poses, 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 organiza-
tions and users who posts multiple times during a certain period of time on flu related
activities. This data is fed into the Analysis Engine which has a detector and ARX pre-
dictor model. The visualization tools and reporting services generate timely visual and
data centric reports on the ILI situation. The 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 with our results.

4 Data Set

In this section we briefly describe our datasets used for influenza prediction. OSN has
emerged as a primary source of user interactions on daily events, health status up-
dates, entertainment, etc. At any given time, tens of millions of users are logged onto
OSN’s, with each user spending an average of tens of minutes daily. Since Oct 18,
2009, we have searched and collected tweets and profile details of Twitter users who
mentioned flu descriptors in their tweets. Facebook opened their Search functionality
in early February 2010 and since then we have been fetching status updates and wall
posts of Facebook users with mention of flu descriptors. The preliminary Twitter anal-
ysis for the year 2009-2010 is documente d in [1]. For 2010-2011, we have 4.5 million
tweets from 1.9 million unique users and 2.0 million facebook posts from 1.5 million
unique facebook 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 the 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. Analysis on
location details of Facebook dataset suggest that 22% users are within USA, 17% users
are outside USA and 61% users have not published their location details.

Initial analysis for the period 2009-2010 indicated a strong correlation between CDC
and Twitter data on the flu incidences [1]. 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. In the year
2009-2010, the Swine Flu event was so evident that the noise did not significantly affect
the correlation that existed then. To mitigate this problem, we removed the spurious
tweets using a filtering technique that trains a document classifier to label whether a
message is indicative of a 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 a simple “bag-of-words” text representations
technique based on a vector space, our algorithm classifies messages wherein user men-
tions having contracted the flu himself or has observed the 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 is 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 and 10,000 status updates. Every message is classified by exactly three
Turks and the majority classified result is attached as the final class for that message.

Table 1. Twitter Text Classification 10 fold cross validation results (left) followed by Facebook’s
10 fold cross validation results (right)

| Classifier | Class | Twitter | | Facebook | |
|------------|-------|---------|----------|---------|----------|
|            |       | Precision | Recall | F-value | Precision | Recall | F-value |
| J48        | Yes   | 0.801    | 0.791   | 0.796   | 0.684     | 0.785   | 0.731   |
|            | No    | 0.813    | 0.704   | 0.755   | 0.629     | 0.501   | 0.557   |
| Naive Bayesian | Yes   | 0.725    | 0.829   | 0.773   | 0.688     | 0.847   | 0.759   |
|            | No    | 0.813    | 0.704   | 0.755   | 0.69      | 0.47    | 0.559   |
| SVM        | Yes   | **0.807**| **0.822**| **0.814**| **0.696** | **0.857**| **0.768**|
|            | No    | **0.829**| **0.814**| **0.822**| **0.71**  | **0.485**| **0.576**|

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 incorporated within our text classification algorithm in-
clude setting term frequency and inverse document frequency (tf-idf) weighting, stem-
ing, using a stopwords list, limiting the 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 the United States resulted in a Twitter dataset with 280K pos-
itively classified tweets from 187K unique twitter users and 185K positively classified
facebook posts from 164K unique Facebook users. In order to gauge if the number of
unique twitter users mentioning the 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 per-
centage of weighted ILI visits, which yields a high Pearson correlation coefficient of
0.8907. A similar plot was generated for the number of unique Facebook users men-
tioning about flu per week against the percentage of weighted ILI visits resulting in
Pearson correlation coefficient of 0.8728.
This increase in the number of users posting about the flu is accompanied by an increase in the percentage of weighted ILI visits reported by CDC in the same week. The marked outlier present in the Twitter data as identified in Figure 2 is consistent with Google Flu Trends data when high tweet volumes were witnessed in the week starting January 2, 2011. The CDC has divided the United States into 10 regions as shown in Figure 3. The CDC publishes their weekly reports on percentage weighted ILI visits collated from its ten regions and aggregates then for United States. Figure 4 compares the OSN dataset with CDC reports with and without text classification for each of the ten regions defined by the CDC and for the entire United States 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 play a vital role in improving the overall prediction performance.

4.2 Data Cleaning

The OSN dataset required data cleaning to discount retweets and successive posts from the same users within a certain period of time.

– Retweets: A retweet in Twitter 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 [12]. We call this the 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. For Facebook we create dataset namely Facebook dataset with no posts from same user within certain syndrome elapsed time.

When we compared the 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. Similarly Facebook data with Syndrome elapse time of zero showed a high correlation of 0.8728. 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 [12], user behaviour on Twitter and Facebook follows a 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. Similarly Facebook data within same week becomes our choice of dataset for all subsequent analysis.

From Figure 5 we observe that the Complementary Cumulative Distribution Function (CCDF) of the number of tweets posted by same individual on Twitter 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). However, we do not observe the power-law behavior in the CCDF of number of posts per user on Facebook, as shown in plot on the right hand side of Figure 5.
Table 2. Correlation between OSN Datasets and CDC along with its Root Mean Square Errors (RMSE)

| Syndrome Elapse Time | Twitter          | Facebook         |
|----------------------|------------------|------------------|
|                      | Retweets Correlation coefficient | RMSE errors | Correlation coefficient | RMSE errors |
| 0 week               | No 0.8907 0.3796 | 0.8728 0.4287 |
| 1 week               | No 0.8895 0.3818 | 0.8709 0.4314 |
| 2 week               | No 0.8886 0.3834 | 0.8698 0.4332 |
| 3 week               | No 0.886 0.3878 | 0.8689 0.4346 |
| 4 week               | No 0.8814 0.3955 | 0.8681 0.4357 |

Fig. 5. Complementary Cumulative Distribution function (CCDF) of the number of tweets/posts on Twitter/Facebook by same users

Most of these high-volume tweets in Twitter are created by health-related organizations, who tweet multiple times 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. Similarly, one can identify agencies like “Flu Trackers”, “Influenza Flu” and specific users that actively post on Facebook.

5 Prediction Model

The correlation between OSN 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 relating OSN 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 OSN activity can be risky due to transient changes, such as changes in OSN activity due to flu-related news.

In order to establish a 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 current and past OSN activity, and CDC data from previous weeks. The prediction of current ILI activity using ILI activity from previous weeks forms the autoregressive component of the model, while the OSN data from previous weeks serve as exogenous inputs. By CDC data, we refer to the percentage of visits to a physician for Influenza-Like Illness (also called ILI rate).

5.1 Influenza Model Structure

Although the percentage of physician visits is between 0% and 100%, the number of OSN 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 OSN data as follows:

Logistic ARX Model

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

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/Facebook 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 OSN 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 or Facebook users having no posts from the same user within syndrome elapsed time of 0 week. The flu related messages are defined as posts with keywords “flu”, “H1N1” and “swine flu”. The rationale for the model structure in Eq. (1) is that OSN data provides real-time assessment of the flu epidemic. However, the OSN 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 OSN data, we may be able to take advantage of the timeliness of the OSN data while overcoming the disturbance that may be present in the OSN 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 OSN data, we assume that the most recent data is always available, if a model includes the OSN data terms. In other words, the most current CDC or OSN 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 OSN data.

In order to predict ILI rates in a particular week given current OSN 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 OSN 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 OSN data, \( u \), terms present. Also, if \( m = 0 \) and \( n = 1 \), we have a linear regression between OSN 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 OSN 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 OSN data, we apply the following relationship:

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

(2)

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

(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:

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

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

### 5.3 Cross Validation Results

We fit our model with Twitter data, Facebook data, and the combination of Twitter and Facebook data. According to the cross validation results in Table 3 the models corresponding to $m = 2$ and $n = 0$ have the lowest RMSE for both Twitter and Facebook. This indicates that two most recent data points are required to perform accurate prediction of influenza rates using Twitter or Facebook data. However the model corresponding to $m = 1$ and $n = 2$ for the combination of Twitter and Facebook data has the lowest RMSE among all models. Thus the model corresponding to $m = 1$ and $n = 2$ is used for accurate prediction of influenza rates and it uses most recent CDC ILI data, in addition to the two most recent OSN data points. In general, the addition of OSN 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 $(m = 1, n = 0)$

---

1 Cross Validation Results presented for Twitter dataset differs from our previous work [2] as we disregard the scaling effect caused by creation of new Twitter accounts over time.
Table 3. Root mean squared errors from 10-fold cross validation applied to Twitter Dataset, Facebook Dataset and combination of Twitter and Facebook Dataset. 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.

|          | TWITTER |          |          | FACEBOOK |          |          | TWITTER + FACEBOOK |
|----------|----------|----------|----------|----------|----------|----------|-------------------|
|          | $n = 0$  | $n = 1$  | $n = 2$  | $n = 3$  | $n = 0$  | $n = 1$  | $n = 2$  | $n = 3$  |
| $m = 0$  | 0.3491   | 0.3555   | 0.3765   | 0.4077   | 0.3651   | 0.3812   | 0.5190  | 0.3449  | 0.4297 |
| $m = 1$  | 0.6465   | 0.3708   | 0.3884   | 0.4175   | 0.6465   | 0.4088   | 0.4111  | 0.4061  | 0.6465  | 0.3553 | 0.3108 | 0.3398 |
| $m = 2$  | 0.5527   | 0.3532   | 0.3665   | 0.4016   | 0.5527   | 0.3976   | 0.4015  | 0.4101  | 0.5527  | 0.4121 | 0.3675 | 0.3608 |

comprising of the $y(t - 2)$ term and the constant term for the prediction of $y(t)$ has a RMSE of 0.6465. For the same $m = 1$, the model with additional Twitter data (i.e. $n = 1$) has a lower RMSE of 0.3708, Facebook data (i.e. $n = 1$) has a lower RMSE of 0.4088 and combination of Twitter and Facebook data (i.e. $n = 1$) has a lower RMSE of 0.3553. We observe that the combination of Twitter and Facebook data provides further improvement for prediction accuracy over Twitter and Facebook alone. In this model, using OSN data ($m = 0$) alone is insufficient for prediction and the past ILI rates are critical in predicting future values, as is evident from our results. Therefore, the OSN 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 OSN data and the CDC data. Hence, the more timely OSN data can compensate for the lack of current CDC data and help capture the current flu trend.

Finally in Figure 6 we provide the plots for percentage weighted ILI visits, positively classified Twitter (left) and Facebook (right) users and predicted ILI rate using CDC and Twitter and Facebook for the year 2010-2011.

Fig. 6. Weekly plot of percentage weighted ILI visits, positively classified OSN (Twitter (Left) and Facebook (Right)) datasets and predicted ILI rate using CDC and OSN.
6 Flu Prediction within Regions

We analyzed the relationship between the OSN 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 shows, for both ILI (left) and Twitter (right), the relative intensity across the ten Health and Human Services (HHS) regions (columns) during successive weeks (rows) in the year 2009-2010 during which the H1N1/Swine Flu was evident.

| HHS Region | Week number |
|------------|-------------|
|            | 1           |
|            | 2           |
|            | 3           |
|            | 4           |
|            | 5           |
|            | 6           |
|            | 7           |
|            | 8           |
|            | 9           |
|            | 10          |

Fig. 7. Heatmap of CDC’s Regionwise ILI data (left) and Twitter data (right). Colormap scale included (below).

Figure 8 shows the comparison between actual and predicted ILI rates for Region 1, Region 6 and Region 9.
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 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 actual and predicted ILI values for each of these regions using Twitter data, Facebook data, and the combination of Twitter and Facebook data. We observe that the OSN reports and ILI rates are in fact correlated across regions and therefore corroborate our earlier findings that OSN can improve ILI rate prediction.

7 Conclusions

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

For prediction, we build an auto-regression with exogenous input (ARX) model where the ILI rate of previous weeks from CDC formed the autoregressive portion of the model, and the OSN 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, OSN 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 OSN data was highly correlated with the ILI rates across different HHS regions. Therefore, flu trend tracking using OSN’s significantly enhances public health preparedness against the influenza epidemic and other large scale pandemics.

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

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B.: Predicting flu trends using twitter data. In: International Workshop on Cyber-Physical Networking Systems (June 2011)
2. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B.: Twitter improves seasonal influenza prediction. In: Fifth Annual International Conference on Health Informatics (February 2012)
3. Centers for Disease Control and Prevention: FluView, a weekly influenza surveillance report (2009), [http://www.cdc.gov/flu/weekly](http://www.cdc.gov/flu/weekly)
4. Chen, L., Achrekar, H., Liu, B., Lazarus, R.: Vision: towards real time epidemic vigilance through online social networks: introducing sneft – social network enabled flu trends. In: ACM Mobile Cloud Computing and Services, San Francisco, California (June 2010)
5. Culotta, A.: Detecting influenza outbreaks by analyzing twitter messages. In: Knowledge Discovery and Data Mining Workshop on Social Media Analytics (2010)
6. Espino, J., Hogan, W., Wagner, M.: Telephone triage: A timely data source for surveillance of influenza-like diseases. In: AMIA: Annual Symposium Proceedings (2003)
7. Ferguson, N.M., Cummings, D.A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iam-sirithaworn, S., Burke, D.S.: Strategies for containing an emerging influenza pandemic in southeast asia. Nature 437, 209–214 (2005)
8. Gauvin, W., Ribeiro, B., Towsley, D., Liu, B., Wang, J.: Measurement and gender-specific analysis of user publishing characteristics on myspace. IEEE Networks (September 2010)
9. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2009)
10. Jansen, B., Zhang, M., Sobel, K., Chowdury, A.: Twitter power:tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(1532), 2169–2188 (2009)
11. Jordans, F.: WHO working on formulas to model swine flu spread (2009), http://www.physorg.com/news165686771.html
12. Lazarus, R., Kleinman, K., Dashovsky, I., Adams, C., Kludt, P., DeMaria Jr., A., Platt, R.: Use of automated ambulatory-care encounter records for detection of acute illness clusters, including potential bioterrorism events (2002), http://www.cdc.gov/ncidod/EID/vol8no8/02-0239.html
13. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: International Conference on Knowledge Discovery and Data Mining, Paris, France, vol. 495(978) (2009)
14. Longini, I., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D., Halvoran, M.: Containing pandemic influenza at the source. Science 309(5737), 1083–1087 (2005)
15. Magruder, S.: Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. Johns Hopkins University APL Technical Digest (2003)
16. Mislove, A.: Pulse of the nation: U.S. mood throughout the day inferred from twitter (2010), http://www.ccs.neu.edu/home/amislove/twittermood/
17. Motoyama, M., Meeder, B., Levchenko, K., Voelker, G.M., Savage, S.: Measuring online service availability using twitter. In: Workshop on Online Social Networks, Boston, Massachusetts, USA (2010)
18. Paul, M., Dredze, M.: You are what you tweet:analyzing twitter for public health. Association for the Advancement of Artificial Intelligence (2011)
19. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: 19th International Conference on World Wide Web, Raleigh, North Carolina, USA (2010)
20. Signorini, A., Segre, A.M., 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) (May 2011)
21. Sitaram, A., Huberman, B.A.: Predicting the future with social media. Social Computing Lab, HP Labs, Palo Alto, California, USA (2010)
22. Webb, S., Caverlee, J.: A large-scale study of myspace: Observations and implications for online social networks. Association for the Advancement for Artificial Intelligence (2008)