Forecasting Influenza Levels using Real-Time Social Media Streams

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

***Abstract*—Seasonal influenza is a contagious respiratory illness that can cause various complications, worsen chronic illnesses, and sometimes lead to deaths. During 2009 H1N1 flu pandemic, up to 203,000 deaths occurred worldwide. Early detection and prediction of disease outbreak is critical because it can provide more time to prepare a response and significantly reduce the impact caused by a pandemic. The traditional influenza surveillance system by Centers for Disease Control and Prevention (CDC) collects U.S. Influenza-Like Illness related physicians visits data from sentinel practices and provides a retrospective analysis delayed by two weeks. Google Flu Trends proposed a method that uses online search queries data to estimate current (real-time) influenza activity. Here we present a system that (1) predicts future influenza activities, (2) provides more accurate real-time assessment than before, and (3) combines real-time big social media data streams and CDC historical datasets for predictive models to accomplish accurate predictions. Although retrospective analysis and observations are important, prediction of future flu levels can represent a big leap because such predictions provide actionable insights for public health that can be used for planning, resource allocation, treatments and prevention. Thus, compared to previous work, our work represents an advancement in accuracy of assessments, prediction of future flu activity accurately and an ability to combine big social data and observed CDC data to build predictive models.**

I. Introduction

Seasonal influenza is an acute viral infection that can cause severe illnesses and complication. For instance, the annual epidemics cause about 250,000 to 500,000 deaths worldwide. Centers for Disease Control and Prevention (CDC) reported 105 pediatric deaths due to influenza during 2012-2013 flu season[1]. Monitoring of disease activity enables an early detection of disease outbreaks, which will facilitate faster communication between health agencies and the public, thereby providing more time to prepare a response. Disease surveillance helps minimize an impact from a pandemic and make better resource allocation. The traditional influenza surveillance system by CDC reports weekly national and regional Influenza-Like Illness (ILI) physicians visit data collected from sentinel medical practices[2]. This data is updated once a week and there is typically a two weeks time lag before the data is published. Furthermore, the published data is updated for several more weeks as more clinical data is gathered.

For an early detection of influenza activity, Ginsberg et al.[12] proposed a method that uses flu-related online search engine query data to estimate the current flu activity with one day reporting lag, 1-2 weeks ahead of CDC, and its estimation has been known to be reasonably accurate for most parts. However, in February 2013, an article titled "When Google got flu wrong" [5] reported Google Flu Trends's over-estimation of peak of U.S. flu activity, which was almost double that of CDC's observations.

During the last decade, the number of internet and social networking site users have dramatically increased. People share ideas, events, interests and their life stories over the internet. Twitter[3] is a popular micro-blogging service where users can post short messages up to 140 characters in length. As of January 2017, Twitter has 100 million daily active users and 5 million tweets are generated per day[4]. Experiences and opinions on various topics including personal health concerns, symptoms and treatments are shared on Twitter. Mining such publicly available health related data potentially provides valuable healthcare insights. Furthermore, increasing number of users that access social media platforms on their mobile devices makes social media data an invaluable source of real-time information.

In this paper, we propose a model that (1) predicts future influenza activities, (2) provides more accurate real-time assessment than before, and (3) combines real-time social media data streams and CDC historical datasets for predictive models to accomplish accurate predictions. The results show that our model using multilayer perceptron with back propagation on a large-scale Twitter data can forecast current and future flu activities with high accuracy. The goal of our work is to predict expected influenza activity for the future, a week or more ahead of time so that it can be used for planning, intervention, resource allocation and prevention. Furthermore, we aim to exploit social media communication for the prediction.

The remainder of this paper is organized as follows. We introduce related work on flu surveillance in section II. We describe our method in section III and results for our current and future flu forecast work in section IV. Finally, we conclude in section V.

II. Related Work

For an early detection of disease outbreaks, researcher have used different statistical and machine learning algorithms on

[1] http://www.cdc.gov/flu/spotlights/children-flu-deaths.htm
[2] http://www.cdc.gov/flu
[3] https://twitter.com
[4] https://www.omnicoreagency.com/twitter-statistics/

difference source of data. Over-the-counter pharmaceutical sales data [21] and telephone triage [10] have been used for surveillance of ILI. Christakis et al. [9] studied whether monitoring of social friends could provide early detection of flu outbreaks. Web search queries data has also been used for influenza surveillance [11], [13], [24], [29], [12], [26], [23]. Ginsberg et al. [12] used flu-related google search queries data to estimate current flu activity and the near real-time estimation is reported on Google Flu Trends (GFT) website[5]. Researchers have used GFT data to build an early detection system for flu epidemics [23], [26]. Shaman et al. [26] used GFT data and WHO/NERVSS collaborating laboratories data to to estimate flu activity. The estimated data is then recursively used to optimize a population-based mathematical model that predicts flu activity. Pervaiz et al. [23] developed FluBreaks[6], an early warning system for flu epidemics using Google Flu Trends.

The use of social networking sites for public health surveillance has been steadily increasing in the past few years [6]. Most diseases surveillance works using social media data are focused on Twitter. A very unique feature of Twitter is that messages propagate in real time. Researchers have used Twitter data to predict various real world outcomes [25], [3], [4].

For current estimation of influenza activity, Signorini et al. [27] applied support vector regression algorithm to Twitter stream generated during the influenza A H1N1 pandemic to public sentiment, and Achrekar et al. [1] used auto-regression with exogenous inputs (ARX) model on Twitter data. Lee et al. built a real-time disease surveillance website that tracks U.S. regional and temporal flu activities including popularity of terms related to flu types, symptoms, and treatments [18], [19]. Aramaki et al. [2] proposed a Twitter-based influenza epidemics detection method that used Natural Language Processing (NLP) to filter out negative influenza tweets. Chew et al. [8] analyzed content and sentiment of tweets generated during the 2009 H1N1 outbreak and showed the potential and feasibility of using social media to conduct infodemiology studies for public health. Paul and Dredze [22] applied Ailment Topic Aspect Model to track illnesses over times (syndromic surveillance), measure behavioral risk factors, localize illnesses by geographic region, analyze symptoms and medication usage, and showed the broad applicability of Twitter data for public health research. Li [20] proposed Flu Markov Network (Flu-MN), a spatio-temporal unsupervised Bayesian algorithm based on a 4 phase Markov Network for flu activity prediction. Lampos et al. [17] proposed an automated tool that tracks ILI in the United Kingdom using a regression model and Bolasso, the bootstrapped version of LASSO, for features extraction of Twitter data. Lamb et al. [16] classified tweets into different categories to distinguish those that report infection versus those that express concerns about flu, tweets about authors versus tweets about others in an attempt to improve performance of influenza surveillance. Researchers have studied the diversity of tweets [14], ran real-time spatio-temporal analysis of West Nile virus using Twitter data [15]. Sugumaran and Voss advised to integrate existing epidemic systems, those that uses crowd-sourcing, news media (e.g., GPHIN, MedISys), mobile/sensor network, and real-time social media intelligence, for an improved early disease outbreak system [28]. Chakraborty et al. [7] combined social indicators and physical indicators and used a matrix factorization-based regression approach using neighborhood embedding to predict ILI incidences in 15 Latin American countries.

Retrospective analysis and current estimates are important as they can describe the observed trends. However, further prediction of future flu levels can represent a big leap because such predictions provide actionable insights for public health that can be used for planning, resource allocation, treatments and prevention. In contrast to other approaches, we propose a system that not only estimates current flu activity more accurately, but also forecasts future influenza activities a week in advance beyond the current week using aggregated ILI data by CDC and real-time Twitter data. The results show that our proposed model using multilayer perceptron with back-propagation algorithm can forecast both current and future influenza activities with high accuracy.

III. Method

The data collection and modeling process is illustrated in Figure 1.

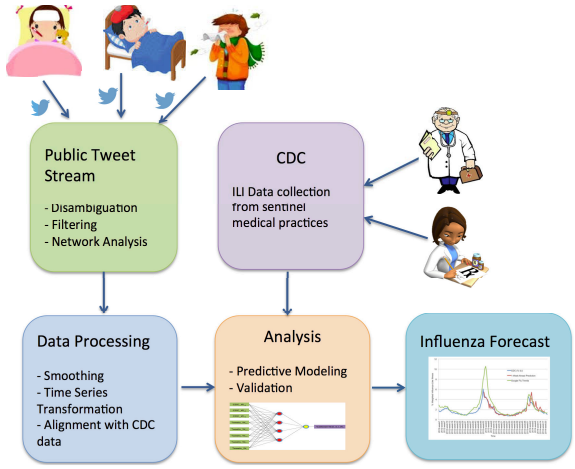


Fig. 1: Data collection and modeling process. Disambiguation, filtering and network analysis are performed on continuously downloaded flu-related tweets. Weekly time-series flu-related tweet counts are computed after data is smoothed out to align with CDC data. Current and 1-week ahead flu prediction models are built.

[5]http://www.google.org/flutrends
[6]http://www.newt.itu.edu.pk/flubreaks

TABLE I: Examples of flu-related tweets.

Category	Tweet
user	I've got the worst flu ever... already D:
user	After a week sick in bed with the flu, look what I just woke up to!
user	trying to get over this flu... I had completely forgot how much harder it is to deal with it during pregnancy.. feeling like death :"c
user, symptom	This flu and cough is killing me T.T
user, symptom	Coding OAuth2 filters with a flu and fever... I look better with a mask on!
someone else	@friend feel better! The flu is nooo fun! Huggs!!
someone else	My roommate has the flu and I get sick really fast I am packing my stuff and won't be returning
someone else	please pray for my mom she's caught the flu and is extremely ill at this moment
symptom	Sore throat, fever, flu, headache, cough. Uhuk uhuk
symptom	sick with flu, sore throat, and slight fever.

A. Dataset

We continuously download publicly available tweets that mention 'flu' using Twitter Streaming API[7]. The dataset used in this paper consists of 20 million tweets generated between December 2012 and May 2014. 71 weeks' data (from week 1, 2013 until week 19, 2014) were used to build the model. Disambiguation of tweets was performed using text analysis techniques to understand if a tweet was about a person talking about his/her own flu or about someone else's or if there were any mentions of common symptoms. Table I lists examples of flu-related tweets. In the category column, *user* indicates that the tweet is about the twitter user being sick with flu, *someone else* indicates that the tweet is about someone else (friends, family, etc.) being sick with flu, and *symptom* indicates that the tweet describes one's flu symptoms. Data was filtered to remove tweets that may contain product advertisements (or links to websites) and using network analysis repeated tweets by the same persons were filtered.

B. Data Preprocessing

The following data preprocessing steps were taken on Twitter data.

- **Smoothing**: We take 7-day moving average of daily tweet volume to identify the long-term flu activity trend by smoothing out the fluctuations and noise in short-term data. Moving average is a popular technique for analyzing time-series data that is often used in financial data analysis such as stock prices.
- **Weekly counts and alignment**: Weekly Twitter data is then computed by summing smoothed daily tweet volumes from Sunday through Saturday. The dates for weekly Twitter data were aligned with dates in CDC weekly surveillance reports so that analysis and predictions can be validated with CDC reports.

[7]https://dev.twitter.com/docs/streaming-apis

- **Normalization**: Weekly data is normalized by dividing each weekly data by the maximum of 72 weekly data points.

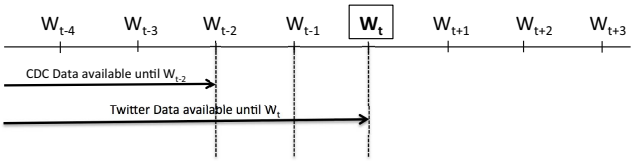


Fig. 2: Data available at current week t. At the end of week t, all flu-related Twitter data collected during current week t and prior are available. At time t, past two weeks (W_{t-1} and W_t)' CDC data is not available as CDC's collection, retrospective analysis and reports take two weeks.

TABLE II: Features.

Notation	Description
CDC-4-3-2	CDC ILI Data for W_{t-4}, W_{t-3}, W_{t-2}
CDC-3-2	CDC ILI Data for W_{t-3}, W_{t-2}
CDC-2	CDC ILI Data for W_{t-2}
Twitter-4-3-2-1-0	Twitter Data for W_{t-4}, W_{t-3}, W_{t-2}, W_{t-1}, W_t
Twitter-3-2-1-0	Twitter Data for W_{t-3}, W_{t-2}, W_{t-1}, W_t
Twitter-2-1-0	Twitter Data for W_{t-2}, W_{t-1}, W_t
Twitter-1-0	Twitter Data for W_{t-1}, W_t
Twitter-0	Twitter Data for W_t

C. Feature Selection

In order to perform predictive modeling, features from the data were defined and extracted as described below. Figure 2 depicts the data available at the end of week t. W_t denotes the current week and any time window beyond this represents the future. W_{t-n} denotes n week(s) prior to current week, and W_{t+n} denotes n week(s) after current week. Each week starts on Sunday and ends on Saturday to align with CDC weekly data. CDC data for current week, W_t, and the week before, W_{t-1}, is not available due to the time it takes to collect patients data from the sentinel practices. The latest available CDC data is weekly data for W_{t-2}. Since we are able to download publicly available tweets in real time, we have all Twitter data generated during W_t. We used the most recent 5 weeks' data for both CDC and Twitter in our experiments.

TABLE III: Twitter data improves prediction performance.

Current Forecast		
Feature	Correlation Coefficient	Improvement
CDC-4-3-2_Twitter-4-3-2-1-0	0.9525	+2.93%
CDC-4-3-2	0.9232	
1-Week Ahead Forecast		
Feature	Correlation Coefficient	Improvement
CDC-3-2_Twitter-4-3-2-1-0	0.9268	+6.37%
CDC-3-2	0.8631	

We experimented with different combinations of CDC and Twitter data shown in table II as features of our predictive model to find the best features for influenza prediction. The model was trained and validated using 10-fold cross validation on 71 weeks data. As shown in table III, the best feature for the current flu level forecast model was feature CDC-4-3-2_Twitter-4-3-2-1-0 (latest 3 weeks' CDC plus latest 5

TABLE IV: Comparison of current flu forecast model performance using different learning rate and varying number of hidden layers and hidden units. The highest correlation of 0.9559 was obtained using learning rate $\lambda = 0.2$ and one hidden layer with 4 activation units.

	Number of activation units in first and second hidden layers									
Learning Rate	2-0	3-0	4-0	5-0	2-2	3-2	4-2	5-2	2-3	3-3
$\lambda = 0.1$	0.9517	0.9496	0.9501	0.946	0.7359	0.8843	0.8976	0.9008	0.8973	0.9143
$\lambda = 0.2$	0.9548	0.954	**0.9559**	0.9527	0.9482	0.9481	0.9469	0.946	0.9498	0.9485
$\lambda = 0.3$	0.953	0.9548	0.9532	0.9499	0.9509	0.9511	0.95	0.9495	0.9518	0.9512

	Number of activation units in first and second hidden layers									
Learning Rate	4-3	5-3	2-4	3-4	4-4	5-4	2-5	3-5	4-5	5-5
$\lambda = 0.1$	0.9038	0.9115	0.915	0.9117	0.9182	0.9134	0.9168	0.9176	0.9256	0.9224
$\lambda = 0.2$	0.9465	0.9457	0.9501	0.948	0.9472	0.9455	0.9502	0.9483	0.9472	0.9466
$\lambda = 0.3$	0.9495	0.9492	0.9521	0.9506	0.9504	0.9491	0.9523	0.951	0.9504	0.9496

TABLE V: Comparison of 1-week ahead flu forecast model performance using different learning rate λ and varying number of hidden layers and hidden units. The highest correlation of 0.929 was obtained using learning rate $\lambda = 0.2$ and one hidden layer with 4 activation units.

	Number of activation units in first and second hidden layers									
Learning Rate	2-0	3-0	4-0	5-0	2-2	3-2	4-2	5-2	2-3	3-3
$\lambda = 0.1$	0.9115	0.9176	0.9064	0.9018	0.8919	0.894	0.8907	0.8908	0.8984	0.8947
$\lambda = 0.2$	0.8996	0.904	**0.929**	0.9268	0.88	0.8843	0.8792	0.8768	0.8917	0.883
$\lambda = 0.3$	0.8491	0.8845	0.9268	0.8944	0.8831	0.878	0.8788	0.8775	0.887	0.8799

	Number of activation units in first and second hidden layers									
Learning Rate	4-3	5-3	2-4	3-4	4-4	5-4	2-5	3-5	4-5	5-5
$\lambda = 0.1$	0.8937	0.8931	0.8958	0.8981	0.8961	0.895	0.8957	0.8979	0.8981	0.8969
$\lambda = 0.2$	0.8806	0.8804	0.8948	0.8957	0.8877	0.8833	0.8965	0.8939	0.8916	0.8869
$\lambda = 0.3$	0.8759	0.8775	0.8893	0.8846	0.9023	0.8767	0.8902	0.9055	0.881	0.8824

week's Twitter data) with correlation coefficient of 0.9525, with +2.93% performance improvement over feature CDC-4-3-2 (latest 3 weeks' CDC data). The best feature for 1-week ahead prediction model was CDC-3-2_Twitter-4-3-2-1-0, which resulted in correlation coefficient of 0.9268, with +6.37% improvement over CDC-3-2. This clearly shows that adding Twitter data significantly improves the performance of both current and future flu level forecasts compared to that using only past CDC data.

D. Predictive Modeling

The proposed model has two parts. The first estimates current flu activity in terms of percentage of ILI-related physicians visit (2 weeks ahead of CDC data). The second part is forecasting future influenza activity a week into the future (3 weeks ahead of CDC data). We use multilayer perceptrons (MLP) with back propagation as it had the best performance among many learning and predictive modeling algorithms we experimented with in forecasting both current and future influenza activities. In our experiments, we used 3-layer MLP with 4 activation units in the hidden layer. The network structure for our current flu activity forecast model is shown in figure 3.

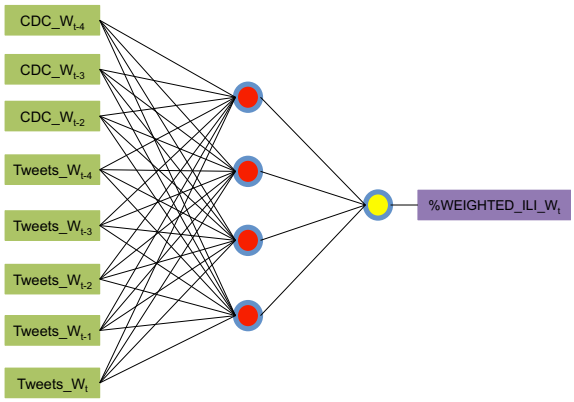


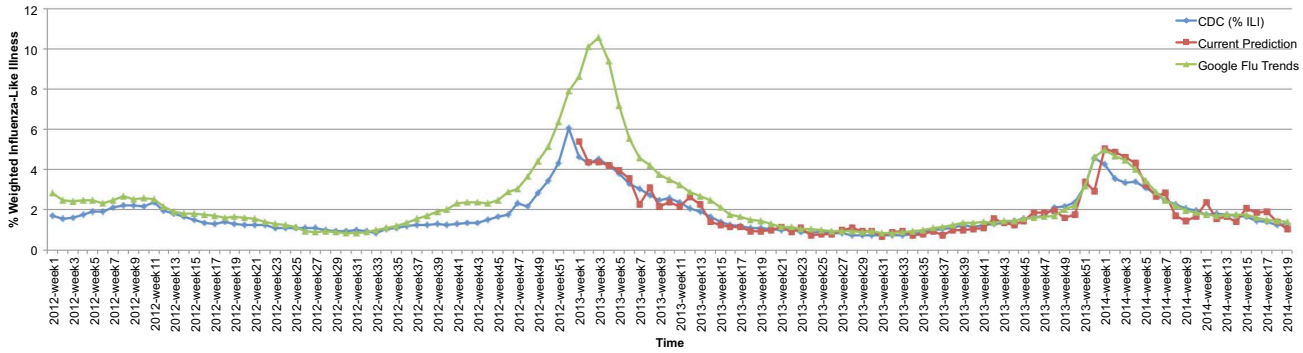
Fig. 3: Structure of multilayer perceptron used in our influenza activity forecast model.

IV. Results

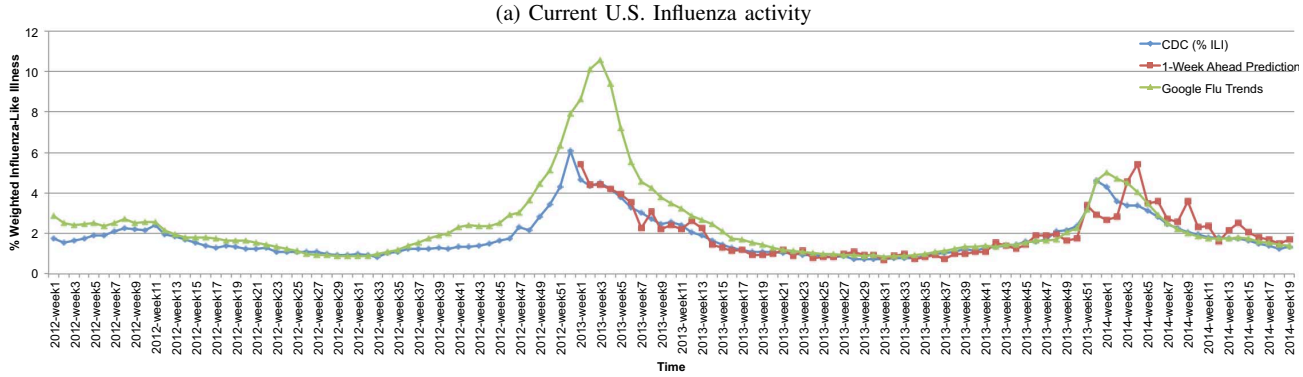
Table IV and V show how the performance of current and 1-week ahead forecast model changes with different value of learning rate and varying number of hidden layers and units in each hidden layer respectively. In notation "A-B", A indicates the number of activation units in first hidden layer (layer 2) and B indicates the number of activation units in second hidden layer (layer 3). Both the current and the 1-week ahead forecast models achieved the best performance using learning rate λ = 0.2 and 3-layer multilayer perceptron structure (input layer, 1 hidden layer, output layer) with 4 activation units in the hidden layer as shown in Figure 3.

Current Influenza Activity Estimation

Our current flu forecast model uses CDC-4-3-2-Twitter-4-3-2-1-0 (i.e., all currently available CDC and Twitter data generated in recent 5 weeks) as features because it gave the highest correlation of 0.9525 when the model was trained and validated using 10-fold cross validation on 71 weeks



(a) Current U.S. Influenza activity



(b) 1-week ahead U.S. Influenza activity

Fig. 4: Comparison of our current and 1-week ahead U.S. influenza activity forecast results against CDC and Google Flu Trends data. For current week prediction, a correlation coefficient of 0.9522 over 52 training data and a correlation coefficient of 0.929 over 19 held-out test data points were obtained. For 1-week ahead forecast, a correlation coefficient of 0.895 over 52 training data and a correlation coefficient of 0.71 over 19 previously unseen test data points were obtained.

data. Although our Twitter dataset has been collected for 1.5 years, each weekly data makes only one data point for the weekly flu activity forecast model. To best utilize the number of available data points, we built the initial model using the first one year data (52 data points for year 2013) with 10-fold cross validation. Then each week we incrementally built a new model with all available data points. For example, a new model is trained using 52 data points (week 1, 2013 – week 52, 2013) to make current flu level prediction for week 1, 2014. Then a newer model is built again using 53 data points (week 1, 2013 – week1, 2014) to make current prediction for week 2, 2014. As we continue to collect more Twitter data, the model will be trained on a larger data set and therefore be more robust. Figure 4 is a time-series graph that compares our flu activity prediction (red line) against the actual CDC %ILI (blue line) and Google Flu Trends data (GFT) [12] (green line). The earliest prediction by our model is for the first week of 2013 because we started collecting flu-related Twitter data in late 2012. Both our prediction (Fig. 4(a)) and GFT data are available two weeks earlier than official CDC ILI report. Our model was fitted on 52 weeks data (week 1, 2013 – week 52, 2013) with a correlation of 0.9522 and a mean absolute error (MAE) of 0.2383, and was further validated on 19 previously unseen weekly data (week 1, 2014 – week 19, 2014) with a correlation of 0.929 and MAE of 0.493. As can be seen, our prediction does as well or better than the GFT data at most data points, and aligns very well with CDC ILI data. Furthermore, our prediction performs significantly better than GFT during January 2013 when GFT's algorithm significantly overestimated peak flu levels [5].

Future Influenza Activity Forecast

Our 1-week ahead flu forecast model uses CDC-3-2-Twitter-4-3-2-1-0 as features. This feature set provided the highest correlation of 0.9268 on the model trained and validated using 10-fold cross validation on 71 weeks data, which is higher than the correlation of 0.8952 obtained by using only CDC-3-2. Here also adding Twitter data improved the model performance. An initial model was built using the first one-year data and a newer model was incrementally rebuilt in the following weeks (in a similar manner our current flu forecast model was built). Our 1-week ahead forecast data (Fig. 4(b)) is available 3 weeks ahead of the official CDC ILI report and 1 week ahead of GFT data. The model was fitted using 52 data points (week 1, 2013 - week 52, 2013) and incrementally rebuilt using all available data (including the new weekly data

collected during the current week) thereafter. The final model was validated by measuring a correlation between the CDC weekly percentage weighted ILI and that predicted by our model on 19 additional previously unseen weekly data points (week 1, 2014 through week 19, 2014). A correlation of 0.895 and MAE of 0.3846 were obtained on the training data and a correlation of 0.71 and MAE of 0.662 were obtained on the previously unseen test data. These results are very good considering our forecast data is available 3 weeks faster than the official CDC data.

V. Conclusion and Future Work

We presented a model that predicts weekly percentage of U.S. population with Influenza-Like Illness using multilayer perceptron with back propagation algorithm on a large-scale social media stream. Adding recent flu-related Twitter data as features improved the model's performance for both current and future forecast. Our proposed model can predict current and future influenza activities with high accuracy 2-3 weeks faster than the traditional flu surveillance system can. The performance for the current prediction is comparable to or better (in January 2013) than GFT. We expect the model's performance to improve as we continuously collect more Twitter data. We believe these results present a very important step in not only accurately forecasting flu activity for the future, prevention, resource planning, but also demonstrating a technique that can combine social media, unstructured communication data, with observational data for prediction. For future work, we would like to investigate how mentions of different symptoms is related with the actual flu levels and whether it can be used to improve the influenza activity forecast. We also want to improve flu forecast accuracy by classifying tweets into multiple categories (health, news, ads, etc.) and by applying varying weights on different types of tweets because the number of posts talking about one flu incidence can vary depending on the category of the tweet.

Acknowledgment

This work is supported in part by the following grants: NSF award CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330, and Northwestern Data Science Initiative.

References

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting Flu Trends using Twitter data. In *Proceedings of the IEEE Conference on Computer Communications Workshops*, 2011.

[2] E. Aramaki, S. Maskawa, and M. Morita. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

[3] S. Asur and B. A. Huberman. Predicting the Future with Social Media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, pages 492–499, 2010.

[4] J. Bollen and H. Mao. Twitter mood as a stock market predictor. *Computer*, 44(10):91–94, 2011.

[5] D. Butler. When Google got flu wrong. *Nature*, 494(7436):155–156, Feb. 2013.

[6] D. Capurro, K. Cole, I. M. Echavarría, J. Joe, T. Neogi, and M. A. Turner. The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review. *J Med Internet Res*, 16(3):e79, Mar 2014.

[7] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, and J. C. and. Forecasting a Moving Target: Ensemble Models for ILI Case Count Predictions. In *SDM*, 2014.

[8] C. Chew and G. Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5(11):e14118, 11 2010.

[9] N. A. Christakis and J. H. Fowler. Social network sensors for early detection of contagious outbreaks. *PloS one*, 5(9):e12948, 2010.

[10] J. U. Espino, W. R. Hogan, and M. M. Wagner. Telephone triage: a timely data source for surveillance of influenza-like diseases. In *AMIA Annual Symposium Proceedings*, page 215, 2003.

[11] G. Eysenbach. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings*, page 244, 2006.

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

[13] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PloS one*, 4(2):e4378, 2009.

[14] N. Kanhabua and W. Nejdl. Understanding the Diversity of Tweets in the Time of Outbreaks. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, pages 1335–1342, 2013.

[15] P. Kostkova. A Roadmap to Integrated Digital Public Health Surveillance: The Vision and the Challenges. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, pages 687–694, 2013.

[16] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In *HLT-NAACL*, pages 789–795, 2013.

[17] V. Lampos, T. De Bie, and N. Cristianini. Flu detector-tracking epidemics on Twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602. 2010.

[18] K. Lee, A. Agrawal, and A. Choudhary. Real-Time Digital Flu Surveillance using Twitter Data. In *Proceedings of the SDM Workshop on Data Mining for Medicine and Healthcare*, pages 19–27, 2013.

[19] K. Lee, A. Agrawal, and A. Choudhary. Real-time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1474–1477, 2013.

[20] J. Li and C. Cardie. Early Stage Influenza Detection from Twitter. *arXiv preprint arXiv:1309.7340*, 2013.

[21] S. Magruder. Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL technical digest*, 24(4):349–53, 2003.

[22] M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM*, 2011.

[23] F. Pervaiz, M. Pervaiz, N. Abdur Rehman, and U. Saif. FluBreaks: Early Epidemic Detection from Google Flu Trends. *J Med Internet Res*, 14(5):e125, Oct 2012.

[24] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[25] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

[26] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, 4, Dec. 2013.

[27] A. Signorini, A. M. Segre, and P. M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 6(5):e19467, 05 2011.

[28] R. Sugumaran and J. Voss. Real-time Spatio-temporal Analysis of West Nile Virus Using Twitter Data. In *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*, pages 39:1–39:2, 2012.

[29] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein. Monitoring Influenza Epidemics in China with Search Query from Baidu. *PLoS ONE*, 8:e64323, 05 2013.