Predicting Flu Epidemics Using Twitter and Historical Data

Giovanni Stilo[1], Paola Velardi[1], Alberto E. Tozzi[2], and Francesco Gesualdo[2]

[1] Dipartimento di Informatica Sapienza Università di Roma, Italy
{stilo,velardi}@di.uniroma1.it
[2] Bambino Gesù Children Hospital, Roma, Italy
{alberto.tozzi,f.gesualdo}@gmail.com

Abstract. Recently there has been a growing attention on the use of web and social data to improve traditional prediction models in politics, finance, marketing and health, but even though a correlation between observed phenomena and related social data has been demonstrated in many cases, yet the effectiveness of the latter for long-term or even mid-term predictions has not been shown. In epidemiological surveillance, the problem is compounded by the fact that infectious diseases models (such as susceptible-infected-recovered-susceptible, SIRS) are very sensitive to current conditions, such that small changes can produce remarkable differences in future outcomes. Unfortunately, current or nearly-current conditions keep changing as data are collected and updated by the epidemiological surveillance organizations. In this paper we show that the time series of Twitter messages reporting a combination of symptoms that match the influenza-like-illness (ILI) case definition represent a more stable and reliable information on "current conditions", to the point that they can replace, rather than simply integrate, official epidemiological data. We estimate the effectiveness of these data at predicting current and past flu seasons (17 seasons overall), in combination with official historical data on past seasons, obtaining an average correlation of 0.85 over a period of 17 weeks covering the flu season.

Keywords: Twitter mining, epidemiological surveillance, predictability of health-related phenomena.

1 Introduction

Prediction of social phenomena in politics, finance, marketing and health is traditionally based on historical data of the same type, i.e. on time series $S(t) = (s_{t-n}, \ldots s_{t-1}, s_t)$ taken up until and including the time *t* in which the prediction is produced. These data are used to train predictors based on linear or non-linear regressions, machine learning, or model-based methods [1], the latter being the hardest way to do prediction since they require deep insight into the observed phenomenon. In recent research [2-11], it has been shown that better predictions can be obtained when augmenting historical data with social data, such as the frequency time series *K(t)* of a keyword in web search data or in micro-blogs.

D. Ślęzak et al. (Eds.): BIH 2014, LNAI 8609, pp. 164–177, 2014.

Even though a correlation between observed phenomena and related social data has been demonstrated in many cases, yet the effectiveness of the latter for long-term or even mid-term predictions has not been shown. Furthermore, the usefulness of social observations is usually limited to recent values (k_t and k_{t-1}).

More related to the research described in this paper is the problem of predicting health-related phenomena, such as disease outbreaks. A seminal work in this area is [6], in which the level of influenza (influenza-like illness, ILI) in the U.S. is estimated using the relative frequency of search queries related to influenza-like illness. Similarly, in [7], the authors demonstrate that query search volumes associated to Dengue fever can predict the incidence of Dengue. Another recent study [8] analyses the problem of predicting the tendency of hand-foot-and-mouth disease (HFMD), clustering HFMD-related search queries, medical pages and news reports. In some cases, a correlation between search volumes and disease trends has been identified and, in 2008, a Google service, Flu Trends[1] (GFT), was developed to estimate and predict influenza activity by aggregating Google search query volumes. However, web search peaks can be completely unrelated to the incidence of a disease, as search behaviors change over time and discussions on traditional media may become reflected in search patterns. For example, GFT overestimated peak flu during the 2013 season[2], following breaking news on bird flu cases in China.

In [9,10] Twitter messages are used to predict flu trends. The advantage of Twitter (and micro-blogs in general) is that, unlike for user queries, a context is provided to distinguish cases in which a user is actually infected from those in which he/she is expressing fear of being infected. For example, Lamb et al. [9] separate tweets reporting infection (*flu*) from those expressing concerns and fear ("*a little worried about flu epidemic*!"), using a classifier trained with a number of specific linguistic features, like expressions of concern. The correlation of the related time series of Twitter messages with official ILI data published by the Center for Disease Control and Prevention[3] (CDC) is reported to be 0.98 in 2009 and 0.79 in 2011.

We note however that the studies mentioned so far do not actually "predict" disease trends, though they have been shown to correlate more or less well with available data on disease outbreaks provided by official epidemiological data. Rather, since both Twitter and GFT provide real-time data, methods based on these data are able to provide an "instant" forecast, while epidemiological data become available typically with one week of delay. A more interesting objective is to make a mid or long-term prediction, i.e. to be able to predict an influenza peak several weeks in advance. This objective is targeted in [11], where a model-based predictor is defined. The system uses an ensemble of SIRS[4] (susceptible-infected-recovered-susceptible) epidemic models to simulate the number of people infected with influenza in all major US cities. The authors use a data assimilation technique to adjust observable and non-observable

[1] http://www.google.org/flutrends/

[2] http://www.nature.com/news/when-google-got-flu-wrong-1.12413

[3] www.cdc.gov/

[4] en.wikipedia.org/wiki/Epidemic_model

model state variables. The predictor is based on historical data, along with real-time data on humidity, official CDC reports, and GFT estimates. The model has shown to accurately predict the influenza peak up to 9 weeks before in 2012. On week 52, prior to the influenza peak in all main cities, 63% of the city forecasts were accurate, where accuracy is computed in terms of precision at predicting the peak in a ±1window.

The forecasting model described in [11] is by far the most complex and accurate presented in literature. However, the actual improvement obtained thanks to the use of GFT is not clear and, as a matter of facts, the authors note than in a previous study they used only GFT data, but then decided to employ an alternative metric that only in part relies on GFT. In this paper, we do not aim to define a better predictor, rather, our objective is to mitigate the causes for which a very complex calibration model is needed, as described in [11]. Infectious diseases models (such as SIRS) are very sensitive to current conditions, such that small changes can produce remarkable differences in future outcomes. Unfortunately, current or nearly-current conditions, e.g. the estimated number of infected individuals at time t, t-1.., as provided by the official surveillance organizations (for example, CDC), keep changing as data are locally collected and updated. On the other side web-based indicators, such as GFT, are stable but less reliable, since they might be affected by other phenomena (fear of being infected, information needs, etc.) than the one being observed. Furthermore, the case definition for ILI patients provided by CDC, requires a combination of symptoms[5] that cannot be mirrored accurately by GFT, as noted also in [11].

In this paper we show that the time series of Twitter messages reporting a combination of symptoms that precisely match the ILI case (hereafter denoted as *ILI-Tweets*) represent a stable and reliable information on "current conditions", to the point that they can reliably replace official ILI data (ILI^{CDC}). Using ILI-Tweets, we can produce reliable mid-term forecasts with a simpler model than the one in [11].

The paper is organized as follows: in Section 2, we summarize our method to extract ILI-Tweets. We also show that ILI^{CDC} data, as published weekly by CDC, are quite variable especially in the shot run (e.g. the current and past two-three values), but, when they eventually stabilize, there is a remarkably high correlation with our ILI-Tweets: in other terms, the ILI^{CDC} curve tends to overlap with our ILI-Tweets curve as official data gets stable. In Section 3 we present our forecasting model, based on ILI-Tweets and historical ILI^{CDC} data on past seasons. Section 4 is dedicated to evaluation: we both validate our model on current (on-going) 2013 flu season and retrospectively, on the past 17 seasons for which ILI^{CDC} data are available, including outlier seasons. Section 5 concludes the paper.

2 Using Twitter Data for Syndromic Surveillance

This section shortly described the algorithm used to extract ILI-related messages from Twitter. Further details can be found in [12-14].

[5] `http://www.acha.org/ILI_Project/ILI_case_definition_CDC.pdf`

2.1 Tracking Twitter Messages Reporting ILI Symptoms

Twitter mining algorithms used in previous health-related studies have measured the occurrence of single pre-specified terms, consisting of either the name or synonyms of a clinical condition (e.g: *H1N1* or *swine flu*) or of words, arbitrarily chosen by the authors, related to the clinical syndrome itself (e.g. *flu, vaccine, tamiflu*) and/or to specific expression, e.g. fear of infection [9]. However, this kind of approach may suffer from major biases, which we will illustrate with an example. Consider the following striking difference in the usage of terms describing the same health conditions, the first by a clinician, the second by a patient: "*Clinicians should maintain a high index of suspicion for this diagnosis in patients presenting with influenza-like symptoms that progress quickly to respiratory distress and extensive pulmonary involvement.*"[6] "*For the past 3 days I have had a stuffy, runny nose, congested chest, fever, sore ears and throat and burning eyes. I've been taking cold and flu medication, and it doesn't help*"[7]. Clearly, the patient's symptoms should induce "a high index of suspicion", but the similarity between these symptom descriptions is not so obvious as to allow capture by an automated system, for two reasons: First, in blogs and forums, people are motivated by a communication need (frequently "one-to-one", between just two individuals), rather than by an information need, and therefore naïve language is often preferred to technical language. Thus, understanding the way people talk about medical terms (diseases, symptoms, and treatments) in "peer to peer" communications is crucial for an effective monitoring of health-related behaviors based on social data. Second, it is likely that, in their tweets, most users will describe a combination of symptoms rather than a diagnosis. An approach that takes into account only disease-related keywords can miss a large volume of messages in which users include a mix of signs and symptoms that may in reality be describing a clinical syndrome. With reference to the previous example, high co-occurrence rates of symptoms like *runny nose, congested chest, sore ears* etc. may be used to trigger an alarm in syndromic surveillance systems.

To cope with these issues, we adopted an entirely different approach. We first developed an algorithm to automatically learn a variety of expressions that people use to describe their health conditions, thus improving our ability to detect health-related "concepts" expressed in non-medical terms and, in the end, producing a larger body of evidence. We then implemented a Twitter monitoring instrument to finely analyze the presence and combinations of symptoms in tweets. We transformed five common syndrome definitions into a Boolean query, thereby basing our analysis on a combination of symptoms (each expanded with a set of correspondent naïve terms) rather than on a suspected or final diagnosis. For example, the Boolean query for influenza, matching at best the official CDC definition for ILI (see Section 1), is:

$$(1)\quad [(\mathit{fever}) \vee (\mathit{chills})) \vee (\mathit{malaise}) \vee (\mathit{headache}) \vee (\mathit{myalgia})] \wedge [(\mathit{cough}) \vee (\mathit{pharyngitis}) \vee (\mathit{dyspnea})]$$

This query is extended replacing the technical terms with the disjunction of its correspondent naïve terms retrieved by our algorithm, for example: *malaise→ malaise, unease, discomfort, weakness, feeling of sickness, feel sick, bodily discomfort, body*

[6] www.ncbi.nlm.nih.gov/pubmed/20085663

[7] ehealthforum.com

aches, body pain, pain in body. Query expansion with naïve terms considerably increases the number of matches, thus providing a statistically reliable body of evidence. An example of tweet matching the ILI query is: "*If this is the flu! I am going to be so pissed: fever, nausea, neck pain, sore throat, all this coughing.. its back to bed!"*. Furthermore, we geo-localize our matching tweets using a variety of methods, not described here for sake of space (the interested reader is referred to our publications). Therefore, we can produce a reliable estimate of ILI cases in U.S. and even more fine-grained geographical distributions, for selected regions.

Detection of naïve language and symptom-driven keyword analysis (rather than disease-driven) represent a major difference with previous methods for syndromic surveillance. First, knowledge of naïve language provides a considerably larger corpus of evidence. Then, second, knowledge of patients' language allows fine-grained queries to be performed on the Twitter corpus, separating, for example, patients with simple cold symptoms from those with an allergy, or a "true" ILI, thus solving a "noise" problem pointed out in [11], and not considered, e.g. in [9]. Third, our methodology (similarly to [9]) is very reliable in selecting only tweets of people that actually complain of being infected, rather than people worried by the possibility of being infected. In fact, people may say "*I'm scared about this flu*" but they are unlikely to say "*I'm afraid to get fever, nausea, body aches, and cough!*".

2.2 Creation of the ILI Tweets Dataset

We started collecting our ILI-related Twitter data since February, 2012. We used the available Twitter API[8] and a set of 78 disease-related naïve terms to track about 100% of the total traffic including these words. In peak periods (e.g. February 2013, January 2014), we collected over 3000 Tweets per day matching the ILI query (1). Additionally, we could monitor the higher or lower incidence of individual symptoms during a specific period under observation, e.g. *cough, pharyngitis* and *chills* have been the predominant symptoms complained of by influenza patients in past year (2012-13). The correlation of our data with official flu reports in the US during the past season 2012-13 was remarkably high (around 0.99%), and better, even, than Google Flu Trends (GFT) over the same period, however we started our collection after the peak period. For the current season, which we could monitor since the early beginning of the infection, the correlation is confirmed to be very high, as shown in Figure 1a. The correlation of ILI-Tweets was 0.965 against 0.947 of GFT. Notice also that both GFT and ILI-Tweets curves seems to be shifted one week in advance with reference to CDC. Figure 1b shows that, when shifting the ILI-Tweets curve one week to the right, there is an almost perfect overlapping with CDC. This is a phenomenon that we observed also during the past 2012-2013 season, and can be explained by the fact that patients' reaction on the web and on social media is instantaneous: they send a message as the symptoms occur. Instead, there might be a delay between the occurrence of the illness and the visit to a doctor, with subsequent registration of the case.

We also remark that our method is applicable to any frequent disease, not just flu, and is not prone to fluctuations in web search behaviors. These results makes our

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

ILI-Tweets data potentially more useful for ILI predictions than official CDC data, which are, instead, rather unstable as US States clean and submit additional ILI data from their healthcare providers. Figure 2 shows the ILI^{CDC} curves in subsequent CDC publications (e.g. Pub-46 means the CDC publication on week 47, including data up to week 46), and our ILI-Tweets curve, as on week 52. All curves are z-normalized[9] for comparison. Figure 2 shows that, as ILI^{CDC} values become stable, they get closer to our ILI-Tweets curve. Note also that the problem of data fluctuation is not considered in [11]; in this study, predictions ILI^{CDC} data, but at the end of the season. Even though the method is robust against fluctuations, we don't know if the quality of predictions would have been the same with varying ILI^{CDC} data.

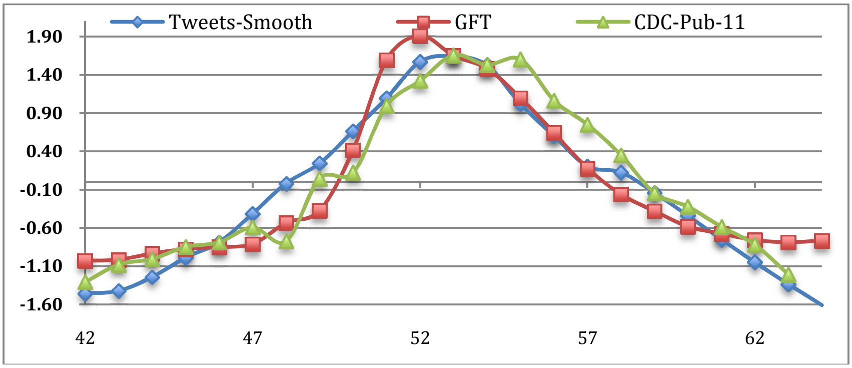

Fig. 1a. Final correlation during 2013-14 flu season, between CDC data (as on Pub-9), ILI Tweets (smoothed with Loess) and Google Flu (already smoothed). Correlation of GFT vrs CDC is 0.947, correlation' of ILI-Tweets vrs CDC is 0.965.

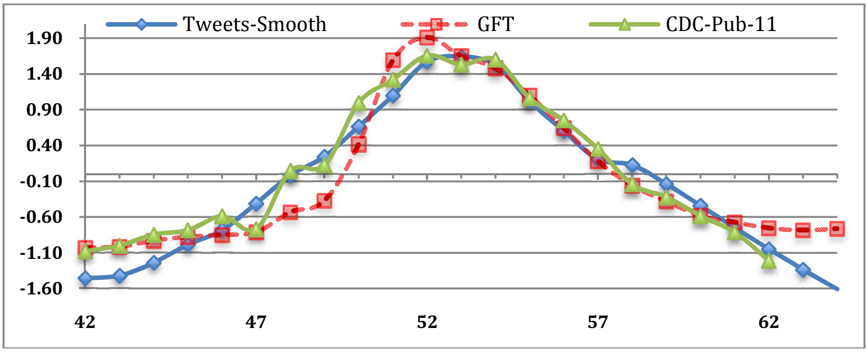

Fig. 1b. As for Fig. 1a, when shifting ILI-Tweets and GFT one week ahead

[9] http://stn.spotfire.com/spotfire_client_help/norm/norm_z_score.htm

3 Summary of Prediction Models

In this Section we describe our prediction model. More precisely, we consider the following available data sources:

1. $S(Y_{i-k})=(x_{40}^{i-k},x_{41}^{i-k},..x_{1}^{i-k},...x_{n}^{i-k})$, the time series of historical ILI data on past seasons *i-k*, where *n*=20 or 39[10];

2. $S(Y_i)=(x_{40}^{i},x_{41}^{i}...x_{m-1}^{i},?,?...?)$, the time series associated to the current year Y_i. We assume that in week *m* the official values from week 40 to week *m-1* are publicly available.

3. $T(Y_i)=(y_{40}^{i},y_{41}^{i},..y_{m}^{i},?,?..)$, the time series of Twitter messages associated to the current year Y_i, whose combination of reported symptoms match the ILI case. We assume that in week *m* the number of matching and associated tweets is known.

While both $S(Y_{i-k})$ $k\neq 0$ and $T(Y_i)$ are stable, $S(Y_i)$ values are unstable, as shown in Figure 2. Such fluctuations, though not substantial, are critical for early predictions, i.e. when *m* is small (<7-8 values, as discussed in Section 4). On the other side, early predictions are definitively more interesting than late predictions: this motivates our choice of using Twitter data rather than $S(Y_i)$ values. Our predictor is based on 2 alternative models:

1. Prototype: we derive a prototype curve, obtained trough z-normalization and alignment of past curves, centered on the seasonal peak. The prototype provides an "average" ILI profile, but is not temporally anchored;

2. Fusion: A temporally anchored curve, obtained trough z-normalization of past seasons, similarity ranking and fusion of past seasons, as compared with available values of the season to be predicted, represented by the Twitter curve.

The best model is automatically selected, as discussed later. The steps of our methodology are the following:

Step 1: z-normalization

All time series are normalized using the z-score. Figure 3 shows the result of this step, limited to past seasons. Note that there are a few seasons that are "outliers", i.e. they behave quite differently from the others, either because they have a very early peak (e.g. 09/10, 03/04) or because they have a very late peak (e.g. 08/09). Instead, it is not infrequent that a season has a double peak.

Step 2: Derivation of Prototype curve (Method 1)

As mentioned before, we define two prediction methods. Our first method is based on the derivation of a prototype curve. To obtain such prototype, curves in Figure 3 are shifted to align their seasonal peak. An average ILI profile is then derived, as shown in Figure 4.

[10] National surveillance data goes from week 40 to week 20 of the subsequent year until season 2001-2002, while n=39 since 2002-2003.

The prediction is based on sliding the curve $T(Y_i) = (y^i_{40}, y^i_{41},..y^i_m, ?, ?..)$on the Average ILI Profile (AIP), until the best match is found. The match is only based on values, not on temporal correspondence. For example, if the best match is along the crossed red points of Figure 4, the prediction from value *m+1* is based on the subsequent values of the AIP curve. This method is used when the current season is classified as an outlier in its early or late stage. Outlier seasons exhibit absolute values exceeding of an experimentally determined threshold θ the correspondent values in the same weeks of the previous two seasons.

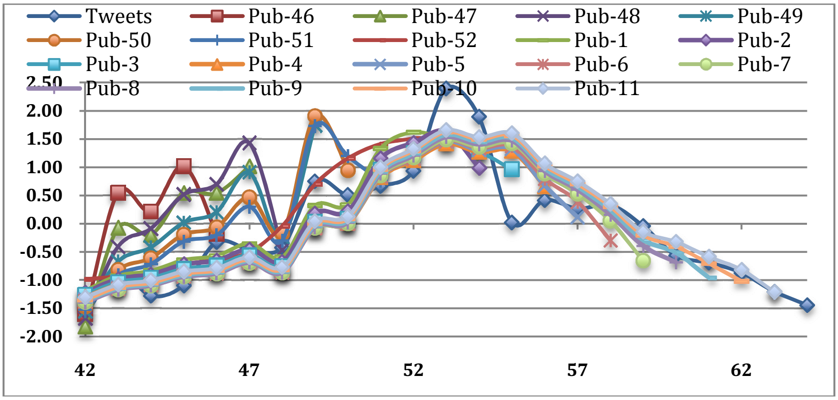

Fig. 2. z-normalized curves for subsequent ILI^{CDC} publications and (not smoothed) ILI Tweets, during early 2013-14 flu season

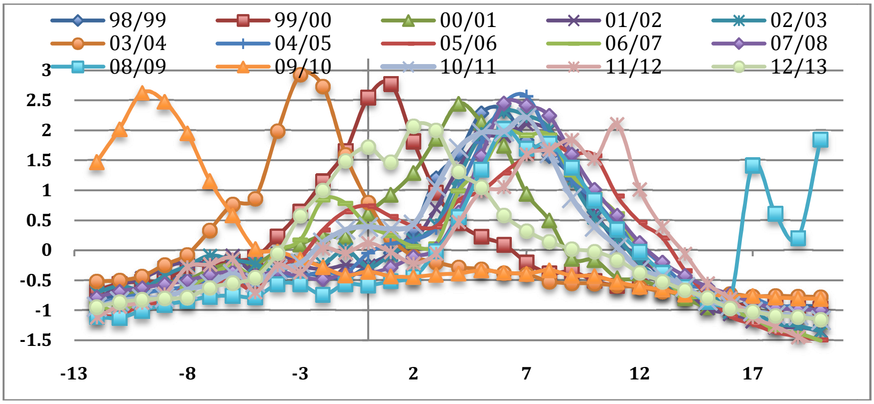

Fig. 3. z-normalized curves for past flu seasons

Step 3: Derivation of a Fusion curve (Method 2)

Our second prediction method is based on deriving a curve that is a fusion of past seasons, weighted according to similarity. The fusion curve is obtained by comparing the known (normalized) values of current season with corresponding values in past seasons, computing a similarity weight for each season and then generating a prediction curve based on combining these seasons. Known values in current season are represented by the known *m* values of $T(Y_i)$. Note that another advantage of using $T(Y_i)$ rather than $S(Y_i)$ is that the *m-th* value is available in week *m*, while ILI$^{\text{CDC}}$ data, besides being subject to fluctuations, become available with one week of delay. We now describe the sub-phases of method 2. In what follows, for simplicity we use the same notation for normalized and not normalized time series.

Step 3.1: derive the "early graphs"

For every $S(Y_{i-k}) = (x_{40}^{i-k}, x_{41}^{i-k}, ..x_{1}^{i-k}, ...x_{n}^{i-k})$ we consider only the first *m* values, which are also available (either real or estimated from Twitter) for year Y_i. We call these the "early graphs", denoted as $S^e(Y_{i-k})$ $k = 0,1,2...$

Step 3.2: Compute point-wise quadratic distance

For every value x_j^i (or y_m^i) in $S^e(Y_i)$ we compute its quadratic distance with reference to x_j^{i-k} and assign a score to the season with the closest value in week *j*.

Step 3.3 Select the most similar season(s)

We explored two scoring schemas:

1) Boolean: a season Y_{i-k} receives a + 1 for every best-matching week, regardless of the week;

2) Weighted: a season receives a score weighted by $1/\ln(m-j)$, e.g. scores are higher for matches in weeks j closest to current week *m*.

With the Boolean scoring method, we select all the season with a score ≥ 1 .With the Weighted method, we select all the seasons with a non-zero cumulative score. Let [Y*] be the set of seasons selected by any of the two scoring methods. A normalized prediction graph is created by averaging the values, in each week, of the selected seasons [Y*]. In both methods, the contribution of each selected season is smoothed by the inverse of cumulated weight.

Step 4. Select best predictors

In Step 2 and Step 3 we defined two prediction methods: one is based on an average ILI profile, the second is based on past most similar seasons. Experiments (see the evaluation Section) show that neither approach is entirely satisfactory: in general, very early predictions (based on very few values of the early graph) and outlier seasons (as those shown in Figure 3) are better predicted using the average profile method, while when a certain number of week values are available, the similarity curve performs best. To select the best predictor we used a classifier to automatically select the best predictors, given the number of available weeks and the absolute distance between absolute values of previous seasons.

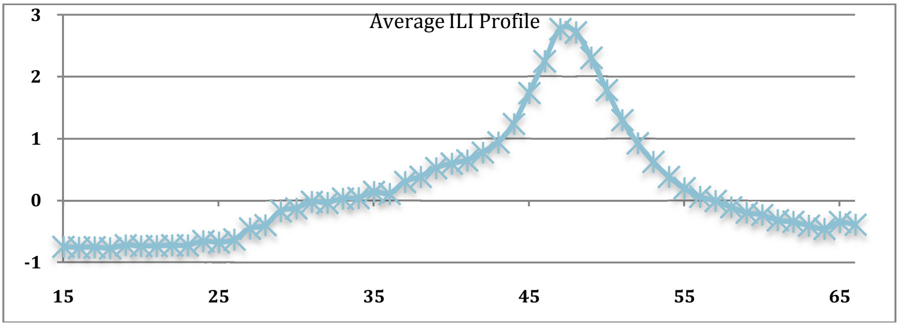

Fig. 4. Average ILI profile (z-normalized, not temporally anchored)

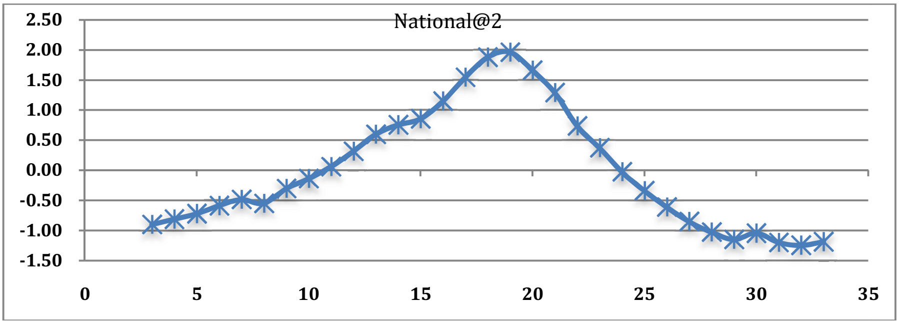

Fig. 5. Prediction of current flu season based on a fusion of most similar past seasons (using the Weighted scoring) with reference to an "early curve" derived from ILI-Tweets on week 2.

Step 5. Produce the final prediction

At the end of step 4 we obtain a prediction with z-normalized values. To produce the final prediction, we need to compute the absolute values for current season, based on selected most similar early graphs of past seasons. To re-weight the (predicted) new values, we use the following procedure. First, we compute the average number of total patients for the involved seasons ([Y*] and Y_i), in the period under analysis.

For example, if the season to be predicted is 13/14, and the selected most similar seasons are 02/03 and 98/99 (both with a score 2), and early graphs span from week 40 to 46, we denote these average values as $\mu_{13/14}, \mu_{02/03}, \mu_{98/99}$.

We compute two re-weighting factors as follows:

$$\delta_{02/03}^{13/14} = \frac{\mu_{13/14}}{\mu_{02/03}}, \delta_{98/99}^{13/14} = \frac{\mu_{13/14}}{\mu_{98/99}}$$

We then compute the season score factors $\xi_s^{13/14}$ as follows:

$$\xi_{02/03}^{13/14} = \frac{r_{02/03}}{\sum_{s\in\{02/03;98/99\}} r_s}, \xi_{02/03}^{13/14} = \frac{r_{98/99}}{\sum_{s\in\{02/03;98/99\}} r_s}$$

Finally, every predicted value for the new season is computed as follows

$$\text{i=}\{47\ldots53,1\ldots39\}\text{: } x_{13/14}^{i} = \sum_{s\in\left\{\frac{02}{03},\frac{98}{99}\right\}} x_s^i * \delta_s^{\frac{13}{14}} * \xi_s^{13/14}$$

Note that a similar re-weighting procedure is initially applied to estimate the value y_k^i of flu cases in the early graph for Y_i, given the number of Twitter messages complaining flu in week *k*. Furthermore, if the selected similar seasons have different scores, their contribution to the computation is smoothed by their score.

Finally, in Figure 5 we show our prediction for year 13/14. The prediction is obtained using the Fusion method (method 2) with Weighted scoring, however the two scoring techniques were experimentally found to have no striking difference in performance. According to our real-time estimate, the peak is predicted in weeks 5-6. The season starts on week 48 and ends on week 16, for a total of 22 weeks.

4 Evaluation Based on Past Seasons

Since the current season is still in progress and we started collecting our ILI Tweets data on February 1st, 2013, we do not have a fully available season to test our method. In previous Section, we have already remarked that, from week 43 to week 2 of season 13/14, our correlation with official data is 0.94, however the actual peak is not yet known at the time we are writing.

To evaluate the quality of our approach, we then performed the following experiment: under the reasonable hypothesis of statistically independent seasons, we remove seasons one at the time, and we then try to predict them retrospectively. For example, to predict season 02/03 we use data from all the other seasons, and we then compare our prediction with the ground truth. Table 1 shows the correlation results obtained when using the method based on Fusion with Weighted scoring on the white rows, and Prototype on the orange rows. Note that selection of the method is automatically determined as described in step 4 of Section 3, however, season 08/09 (in yellow) was not recognized as an outlier, due to the late secondary peak (see Figure 2), therefore results are based on the Fusion method. In the Table 1, rows are the years to be predicted, and columns are the "m", e.g. "45" means that we predict week 45 for year Y_j based on weeks 40-45. Cells (*j,k*) are the Pearson's correlation between the predicted and actual curves for year j, from week 40 up to week k. The last row of Table 1 is the average correlation of all our predictions with reference to the actual values of the seasons to be predicted.

As already remarked, we do not actually have the $T(Y_j)$ values for years previous than 2013 (since, as previously noted, we started collecting flu-related tweets on February 2013). However, under the hypothesis that our method exhibits a stable and strong correlation with the flu seasons (as shown for partial but large fractions of seasons 12/13 [13-15] and 13/14), the effect of knowing $T(Y_j)$ is simulated simply by considering the correspondent known values for $S(Y_i)$.

Note also that, even though in this experiment the early curves for a year to be predicted are represented by the ILICDC values rather than by the Twitter estimate of these values (as it should be in our method), the correlation values in Table 1 are not necessarily optimistic, since, even if we could use stable ILICDC values for a year to be predicted (which, as already noted, is not the case) rather than un-official ILI-Tweets, still

ILI-Tweets keep an advantage over ILICDC data, since they provide an extra "real-time" value for week m. By comparing the values in cells (j,k) with (j,k+1) it is seen that one extra value almost always has a positive effect on performance.

Table 1 shows for example that, in week 48 (which may be considered a relatively early prediction, usually 4-8 weeks before the peak), the average obtained Pearson's correlation is 0,80. As expected, the correlation grows with the length of the available early curve, up to 0,94 on week 18. The average total correlation, i.e. the average of values in the last row, is 0,85.

Table 2 shows the average precision in predicting the peak over all historical data: the columns are the weeks in which the prediction is made, and precision for a season is computed as in [11]: 1 if prediction is accurate within ± 1 week of the observed ILI peak, else it is 0. Even though the comparison is based on uneven data (17 seasons US-wide in our case, season 12/13 and individual predictions for 108 cities[11], in [11]), we

Table 1. Retrospective prediction (using mixed approach)

Year / week	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	Avg
97/98	0,77	0,89	0,89	0,92	0,86	0,60	0,87	0,83	0,87	0,84	0,81	0,89	0,92	0,92	0,98	0,98	0,98	0,87
98/99	0,75	0,78	0,94	0,96	0,92	0,86	0,86	0,97	0,98	0,98	0,98	0,97	0,98	0,98	0,97	0,95	0,97	0,93
99/00	0,48	0,44	0,44	0,78	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,88
00/01	0,84	0,71	0,52	0,84	0,69	0,62	0,93	0,88	0,89	0,88	0,88	0,92	0,93	0,97	0,98	0,99	0,98	0,85
01/02	0,82	0,54	0,88	0,97	0,97	0,98	0,95	0,92	0,94	0,97	0,96	0,98	0,97	0,97	0,96	0,97	0,97	0,92
02/03	0,91	0,56	0,91	0,88	0,78	0,95	0,97	0,97	0,93	0,98	0,96	0,97	0,98	0,96	0,97	0,98	0,96	0,92
03/04	0,10	0,68	0,94	0,68	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,95	0,87
04/05	0,89	0,46	0,95	0,93	0,86	0,98	0,92	0,77	0,88	0,83	0,76	0,91	0,91	0,96	0,95	0,97	0,98	0,88
05/06	0,82	0,87	0,74	0,62	0,90	0,89	0,88	0,86	0,78	0,85	0,82	0,87	0,91	0,87	0,90	0,96	0,97	0,85
06/07	0,90	0,85	0,70	0,83	0,93	0,97	0,95	0,83	0,83	0,81	0,87	0,90	0,96	0,97	0,98	0,94	0,98	0,89
07/08	0,63	-0,19	0,73	0,51	0,81	0,90	0,97	0,79	0,86	0,88	0,91	0,95	0,90	0,93	0,89	0,91	0,92	0,78
08/09	0,44	0,54	0,50	0,64	0,63	0,38	0,35	0,29	0,45	0,53	0,54	0,52	0,58	0,53	0,63	0,61	0,59	0,52
09/10	0,66	0,66	0,76	0,84	0,88	0,94	0,96	0,94	0,95	0,95	0,93	0,97	0,96	0,98	0,98	0,99	1,00	0,90
10/11	0,99	0,87	0,93	0,88	0,93	0,98	0,95	0,96	0,96	0,85	0,91	0,92	0,87	0,88	0,96	0,94	0,97	0,93
11/12	0,69	0,85	0,77	0,84	0,78	0,77	0,68	0,84	0,80	0,78	0,73	0,84	0,86	0,85	0,85	0,88	0,89	0,81
12/13	0,60	0,45	0,85	0,69	0,89	0,86	0,78	0,88	0,88	0,93	0,91	0,92	0,97	0,99	0,99	0,99	0,99	0,86
Person Mean	0,70	0,62	0,78	0,80	0,86	0,85	0,87	0,85	0,87	0,87	0,87	0,90	0,91	0,92	0,93	0,94	0,94	0,85

Table 2. Precision at ± 1 peak prediction (retrospective, all seasons)

Week	43	44	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7
Prec.	0,563	0,563	0,563	0,563	0,625	0,625	0,5	0,5	0,563	0,563	0,688	0,75	0,75	0,75	0,938	0,913	0,813

[11] Our data are hardly comparable with [11], since in that paper predictions are made separately for different cities. Unfortunately, even though we can more finely geo-localize our ILI Tweets, data at the city level are too few for a reliable prediction, because of the well known difficulty to obtain systematic and reliable location indicators in tweets.

note that in [11] (Table 2) the total % accuracy on week 52 is 0.631 against 0.563 in Table 2, on week 1 is 0.638 against 0.688, on week 5 is 0.739 against 0.938, and both predictions have a small drop on week 6 (0.733 and 0.913).

In [11] the authors also compute a total U.S. prediction accuracy obtained as an average of 10 regions, using the Health and Human Service (HHS) region scale: in this experiment, which is more comparable with our data, the accuracy of forecasts in year 12/13 ranged from 0.595 on week 52 to a high 0.90 in week 5, after the seasonal peak (the season last year had a rather anticipated peak). In the same year, our peak prediction on week 52 was 3 weeks after the actual peak, while since week 2 the prediction was correct (0 distance). Furthermore, the correlation of our Fusion curve with the actual ILI curve was 0.88 in week 52 and 0.97 on week 5 (see Table 1, last row). We finally note that correlation is a better performance measure than accuracy in peak prediction since, as previously shown in Figure 2 and 3, many seasons have more than one peak.

5 Conclusions

The major strength of the predictor presented in this paper is the reliability of values that represent the "current state" of the system, as demonstrated by Figures 1 (a and b) and 2 and in our previously published work [13-15].

We are not aware of previous studies that recognized (and solved) the problem of fluctuation of real-time $\mathrm{ILI}^{\mathrm{CDC}}$ data publications, since all papers, including [11], do not actually predict in real-time, but only at the end of the season, when $\mathrm{ILI}^{\mathrm{CDC}}$ data are eventually stable and reliable. Furthermore, ILI-Tweets collected according to our methodology overcome all the limitations of GFT data, also highlighted in [11], i.e.:

1. Media news may inflate GFT ILI estimates;
2. GFT ILI does not accurately model the CDC case definition for ILI, which requires a precise combination of symptoms (captured by ILI-Tweets). GFT estimates may then be affected by other types of respiratory diseases.

For truly real-time prediction, the problem of unreliable knowledge on "current conditions" is critical, since infectious diseases models are very sensitive to fluctuations, such that small changes can produce remarkable differences in future outcomes. If accurate data are available in real-time (as for our ILI-Tweets), even simpler predictors based on historical data may obtain good performance both in terms of correlation and peak prediction.

References

1. Yu, S., Kak, S.: A Survey of Prediction using Social Media (2012), http://arxiv.org/ftp/arxiv/papers/1203/1203.1647.pdf (retrieved)
2. Asur, S., Huberman, B.H.: Predicting the future with social media, http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf (retrieved)

3. Radinsky, K., Horvitz, E.: Mining the web to predict future events. In: WSDM 2013, pp. 255–264 (2013)
4. Ruiz, E.J., Hristidis, V., Castillo, C., Gionis, A., Jaimes, A.: Correlating financial time series with micro-blogging activity. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 513–522. ACM, New York (2012)
5. Carrifere-Swallow, Y., Labbfe, F.: Nowcasting with Google Trends in an Emerging Market. Central Bank of Chile, Working Papers n. 588 (2010)
6. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., et al.: Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014 (2009)
7. Althouse, B.M., Ng, Y.Y., Cummings, D.A.T.: Prediction of Dengue Incidence Using Search Query Surveillance. PLoS Negl. Trop. Dis. 5(8) (2011)
8. Xu, D., Liu, Y., Zhang, M., Ma, S., Ciu, A., Ru, L.: Predicting Epidemic Tendency through Search Behaviour Analysis. In: Proc. of 22nd IJCAI (2011)
9. Lamb, A., Paul, M.J., Dredze, M.: Separating Fact from Fear: Tracking Flu Infections on Twitter. In: NAACL (2013)
10. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S., Liu, B.: Predicting Flu Trends using Twitter Data. In: CPNS 2011 (2011)
11. Shaman, J., Karspeck, A., Yang, W., Tamerius, J., Lipsitch, M.: Real-time influenza forecasts during the 2012-2013 season. Nature Communications 4, 2837 (2013), doi:10.1038/ncomms3837
12. Gesualdo, F., Stilo, G., Agricola, E., Gonfiantini, M.V., Pandolfi, E., Velardi, P., Tozzi, A.E.: Influenza-like illness surveillance on Twitter through automated learning of naïve language. PloS One Public Library of Science One, Journal (2013)
13. Velardi, P., Stilo, G., Tozzi, A.E., Gesualdo, F.: Twitter mining for fine-grained syndromic surveillance. Artificial Intelligence in Medicine, Special Issue on Text Mining and Information Analysis (in press, 2014)
14. Stilo, G., De Vincenzi, M., Tozzi, A.E., Velardi, P.: Automated Learning of Everyday Patients' Language for Medical Blog Analytics. In: Proceedings of the Recent Advances in Natural Language Processing (RANLP 2013) Hissar, Hissar (Bulgaria), September 9-11 (2013)