Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction

Hayate ISO, Shoko WAKAMIYA, Eiji ARAMAKI
Nara Institute of Science and Technology
{iso.hayate.id3,wakamiya,aramaki}@is.naist.jp

Abstract

Because of the increasing popularity of social media, much information has been shared on the internet, enabling social media users to understand various real world events. Particularly, social media-based infectious disease surveillance has attracted increasing attention. In this work, we specifically examine influenza: a common topic of communication on social media. The fundamental theory of this work is that several words, such as symptom words (*fever*, *headache*, etc.), appear in advance of flu epidemic occurrence. Consequently, past word occurrence can contribute to estimation of the number of current patients. To employ such forecasting words, one can first estimate the optimal time lag for each word based on their cross correlation. Then one can build a linear model consisting of word frequencies at different time points for nowcasting and for forecasting influenza epidemics. Experimentally obtained results (using 7.7 million tweets of August 2012 – January 2016), the proposed model achieved the best nowcasting performance to date (correlation ratio 0.93) and practically sufficient forecasting performance (correlation ratio 0.91 in 1-week future prediction, and correlation ratio 0.77 in 3-weeks future prediction). This report reveals the effectiveness of the word time shift to predict of future epidemics using Twitter.

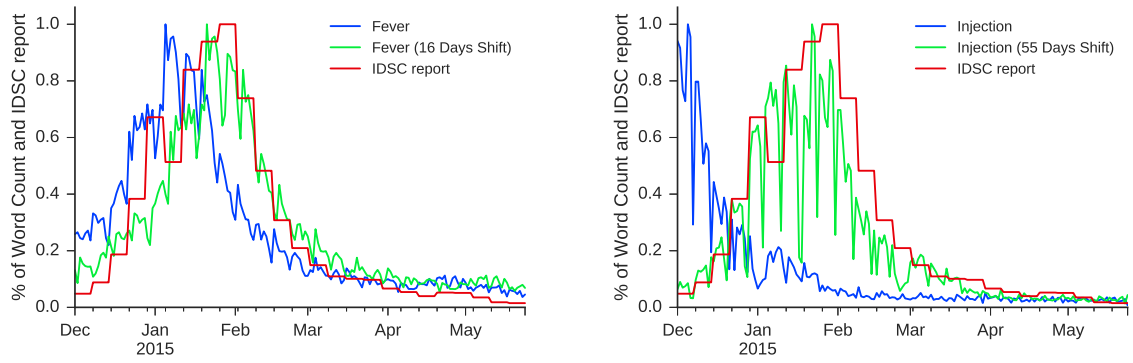
1 Introduction

The increased use of social media platforms has led to wide sharing of personal information. Especially Twitter, a micro-blogging platform that enables users to communicate by updating their status using 140 or fewer characters, has attracted great attention of researchers and service developers because Twitter can be a valuable personal information resource. The feasibility of such approaches, known as social sensors, has been demonstrated in various event detection systems such as earthquakes (Sakaki et al., 2010), outbreaks of disease (Chew and Eysenbach, 2010), and stock market fluctuations (Bollen et al., 2011). Among the applications mentioned above, this study particularly examines detection of seasonal influenza epidemics because the influenza detection is a popular application of Twitter. To date, more than 30 Twitter-based influenza detection and prediction systems have been developed worldwide (Charles-Smith et al., 2015).

Although the detailed functions of these systems differ, they share the underlying assumption that the flu spreading in the real world is immediately reflected to the tweets. Therefore, most systems have simply aggregated counts of daily flu-related tweets to obtain the current patient status (Aramaki et al., 2011; Collier et al., 2011; Chew and Eysenbach, 2010; Lampos and Cristianini, 2010; Culotta, 2013; Paul et al., 2014). Their typical materials are presented as shown below.

- *I got a* **flu** 🤒 *I can not go to school for the rest of the week*
- *I was diagnosed with a high* **fever**. *Maybe flu :(*

Although the former tweet is described by an actual influenza patient, the latter one merely expresses a suspicion of flu. From a practical (clinical) perspective, these differences have great importance because



(a) "*Fever.*" (b) "*Injection.*"

Figure 1: Motivating examples: The time lag of the frequency of a word enables one to obtain a good approximation to the number of patients. The blue line shows the word frequency. The green line shows the word frequency shifted time lag days. The red line shows the number of patients.

the latter is noise that impedes precise influenza surveillance. Therefore, earlier studies (Aramaki et al., 2011; Kanouchi et al., 2015; SUN et al., 2014) have devoted great efforts to removal of such noise (suspicion, negation, news wired, and so on).

This study employs such noisy tweets. We assume that a word, "*fever*" presents a clue to an upcoming influenza outbreak. Inferring that people are frequently afflicted by symptoms such as "*fever*" and "*headache*" immediately before the onset and diagnosis of influenza, we designate such words as **forecasting words**.

More concrete examples of forecasting words are presented in Figure 1a. The figure reveals that an approximately 16-day time lag exists between the frequency of "*fever*" (blue line) and the number of patients (red line). If this time lag was known in advance, one could obtain a good approximation of the number of patients (red line) by a 16-day time shift operation (green line). Similarly, flu prevention words such as "*shot*" and "*injection*" have previously been used to describe outbreaks.

- *I took a flu* **shot** *today* 💉
- *I don't wanna get a flu* **injection** *cuz it hurts me*

In the latter case as shown in Figure 1b, we can find much longer time lag (55 days) between tweets (frequency of "*injection*") and the reality (number of patients).

Presuming that each word has its own time lag, then the problems to be solved are two-fold: (1) estimating the optimal time lag for each forecasting word and (2) incorporating these time lags into the model.

For the first problem, the suitable time lag for each word is measured by calculating the cross correlation between the word frequency and the patient number. For the second problem, we construct a word frequency matrix that consists of a shifted word frequency timeline (Sec. 3). Next, a linear model called **nowcasting model** is constructed from the modified word matrix, for which the parameters are estimated using several regularization models, Lasso and Elastic Net (Sec. 4).

Moreover, the nowcasting model can be extended easily to a predictive model called a **forecasting model**. In the forecasting model (Δf days future), only forecasting words that have more than n day time lag are used (Sec. 5).

Nowcasting models can dramatically boost the current patient number estimation capability (correlation ratio 0.93; $+0.10$ point). Forecasting models have demonstrated successful prediction performance (the correlation ratio 0.91 in 1-week future prediction, and the correlation ratio 0.77 in 3-weeks future prediction). This performance goes beyond the practical baseline (over 0.75 correlation).

Our contributions are summarized as presented below.

- We discover that **forecasting words** have a time lag between the virtual world (number of tweets in Twitter) and the real world (number of patients).

- We propose a method to build time-shifted features using **cross correlation** measures.
- We realize **nowcasting model** and its extended one, **forecasting model**, based on the time shift with parameter estimation. This report is the first of the relevant literature describing a successful model enabling the prediction of future epidemics over the practical baseline.

We make code and data publicly available. [1]

2 Dataset

2.1 Influenza Corpus

We collected 7.7 million influenza related tweets, starting from August 2012 to January 2016, via Twitter API[2]. Then, we filtered noises (removed retweets including the word, *RT*, and tweets linked to other web pages including the word, *http* from the collected tweet data). In the case of just counting influenza-related tweets, we should only consider unique users to avoid to count more than ones the tweets of the same patients. However, we didn't filter out the users which posted influenza-related tweets multiple times because we provide the different word for the different role even if these tweets were posted by the same patients. For example, the word, "fever" for nowcasting, and the word, "injection" for forecasting. To analyze a word, we applied a Japanese morphological parser (JUMAN[3]) and obtained the stem forms. As a result, 27,588 words were extracted. Then, we investigated the word frequency per day to build a word matrix ($days \times words$) as shown in Figure 2a.

2.2 IDSC report

In Japan, the Infectious Disease Surveillance Center (IDSC) announces the number of influenza patients once a week during an influenza epidemic season (typically during November–May in Japan). In fact, IDSC reports tend to delay around a week likewise the U.S. Centers for Disease Control and Prevention (CDC) (Paul et al., 2014), but even if we consider such time delay, twitter stream attains the peak faster than the real world.

To use the IDSC reports for evaluation, we divided the data into the following three periods: 2012/12/01–2013/05/31 (Season 1), 2013/12/01–2014/05/31 (Season 2), and 2014/12/01–2015/05/24 (Season 3). We prepared a buffer time (60 day maximum time shift) immediately preceding the experimental periods to secure the time shift width.

3 Method

To estimate the current influenza epidemics (nowcast) and forecast the future ones, the number of influenza patients was derived from the following linear model.

$$\hat{y}^{(t)} = x_1^{(t-\hat{\tau}_1)}\hat{\beta}_1 + x_2^{(t-\hat{\tau}_2)}\hat{\beta}_2 + \cdots + x_{|V|}^{(t-\hat{\tau}_{|V|})}\hat{\beta}_{|V|}$$

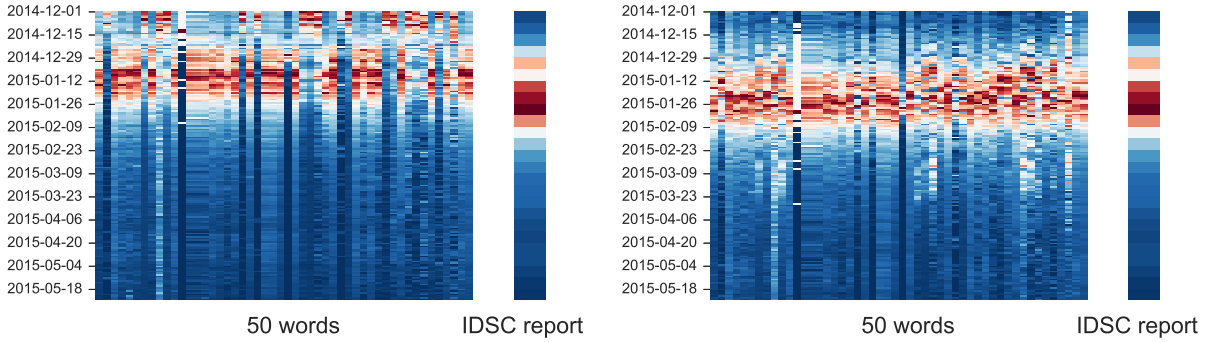
Therein, $\hat{y}^{(t)}$ shows the estimated number of influenza patients at time t, $x_v^{(t)}$ stands for the count of a word v at time t, and $\hat{\beta}$ represents a weight estimated in the training phase, $\hat{\tau}_v$ denotes a suitable time shift parameter for word v decided in the training phase, and $|V|$ denotes the size of vocabulary.

This section first provides methods to explore the most suitable time shift width $\hat{\tau}_v$ for each word v (Sec. 3.1). Then, the parameter estimation method is described (Sec. 3.2). Finally, the model of future prediction based on the original model is explained (Sec. 3.3).

[1] http://sociocom.jp/~iso/forecastword

[2] The tweet data dropout during June–October in 2013 and during June–October in 2014, because the Twitter API specifications were changed in those periods.

[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN



(a) Ordinary Word Matrix.

(b) Time Shifted Word Matrix.

Figure 2: Word matrix transformation. The Y-axis shows a timeline. The X-axis shows words with the IDSC reports (right side).

3.1 Time Shift Estimation

The first problem to be solved is finding the optimal time shift width that achieves the best fit to the target influenza timeline. Given the IDSC reports and wider range of tweets, **Cross Correlation** is used to search for the most suitable time shift width for each word frequency as

$$r_{\mathbf{x}_v,\mathbf{y}}(\tau) = \frac{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v^{(t-\tau)})(y^{(t)} - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_v^{(t-\tau)} - \bar{x}_v^{(t-\tau)})^2 \sum_{t=1}^{T}(y^{(t)} - \bar{y})^2}},$$

where τ is a time shift parameter (time shift width)[4]. The cross correlation $r_{\mathbf{x}_v,\mathbf{y}}(\tau)$ measures the similarity between (τ days) time shift variable $\mathbf{x}_v$ and objective $\mathbf{y}$. In this study, $x_v^{(t-\tau)}$ is the count of word v with time shift width τ days earlier from t and $\mathbf{y} = [y^{(1)}, \ldots, y^{(T)}]^\top$ is the number of patients from the IDSC reports. It is formulated as $\hat{\tau}_v = \underset{\tau}{\operatorname{argmax}}\, r_{\mathbf{x}_v,\mathbf{y}}$.

Next, we construct a matrix, $\mathbf{X} \in \mathbb{N}^{T \times V}$, where T stands for the timeline and V represents the vocabulary, according to the Algorithm 1.

Algorithm 1: Time-shifted word matrix for nowcasting.

Set the maximum shift parameter $\tau_{\max}$
for $v \leftarrow 1$ **to** $|V|$ **do**
  for $\tau \leftarrow 0$ **to** $\tau_{\max}$ **do**
    Calculate Cross Correlation $r_{x_v,y}(\tau)$
  end
  $\hat{\tau}_v = \underset{\tau \in \{0,\ldots,\tau_{\max}\}}{\operatorname{argmax}} r_{x_v,y}(\tau)$
  Shift the word vector to maximize Cross Correlation $\hat{\mathbf{x}}_\mathbf{v} \leftarrow [x_v^{(1-\hat{\tau}_v)}, x_v^{(2-\hat{\tau}_v)}, \ldots, x_v^{(T-\hat{\tau}_v)}]$
end
return *Shifted Word Matrix* $\mathbf{X} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{|V|}]$

The algorithm decides the optimal time shift width ($\hat{\tau}_{\mathbf{x}_v,\mathbf{y}}$) based on the cross correlation for each word. After time shifts for all words, a shifted word matrix $\mathbf{X}$ is constructed.

Figure 2a presents the initial (original) word matrix ($\tau = 0$ for all words) of 50 words (randomly selected). This matrix includes several low-correlated words, making several vertically irregular lines. In contrast, the time shift operation arranges the irregular words to match the IDSC reports, producing a beautiful horizontal line, as shown in Figure 2b.

3.2 Nowcasting

To construct the linear model (called **nowcasting model**), the parameter $\boldsymbol{\beta}$ is estimated as minimizing the squared error. For this study, the vocabulary size $|V|$ is of much larger order than sample size T

[4]The cross correlation is exactly the same as the Pearson's correlation when $\tau = 0$.

so that the ordinary least squares estimator is not unique. It heavily overfits the data. According to the previous study's manner, parameters with a penalty are estimated as shown below.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \mathbf{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})$$

In that equation, $\mathbf{P}(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is the penalty term.

In the case of $\mathbf{P}_{lasso}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \lambda\|\boldsymbol{\beta}\|_1$, the regularization method called the Least Absolute Shrinkage and Selection Operator (Lasso) is a well-known method for selecting and estimating the parameters simultaneously (Tibshirani, 1994). In earlier studies, Lasso was employed to model influenza epidemics by Lampos and Cristianini (2010). However, in the case of vocabulary size $|V|$, which is much larger order than sample size T, it has been observed empirically that the prediction performance of l_1-penalized regression, the Lasso is dominated by the l_2-penalized one.

Therefore, we employ the Elastic Net (Zou and Hastie, 2005), which combines the l_1-penalty and l_2-penalty $\mathbf{P}_{enet} = \lambda(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2)$, where α is called l_1 ratio. The Elastic Net was already employed for nowcasting influenza-like illness rates using search query log, not Twitter (Lampos et al., 2015). In the case of $\alpha = 1$, Elastic Net is exactly the same as Lasso and $\alpha = 0$, Ridge (l_2 regularization). Similarly to Lasso, the Elastic Net simultaneously does automatic variable selection and continuous shrinkage. It has a l-2 regularization advantage that selects groups of correlated variables. Elastic Net, as the generalized method of Lasso and Ridge, estimates with equal or better performance compared to both.

3.3 Forecasting

Our nowcasting model can be extended naturally to **forecasting model**. To predict the number of future patients Δf days after, we force to shift the word frequency at least Δf days. To do so, a setting of the nowcasting model in Algorithm 1 is just changed to $\tau_{\min} = \Delta f$, as shown in Algorithm 2. It enables forecasting of future epidemics, demonstrating a widely applicable methodology of the proposed approach.

Algorithm 2: Time-shifted word matrix for forecasting.

Set the maximum shift parameter $\underline{\boldsymbol{\tau}_{\mathbf{min}}}, \tau_{\max}$
for $v \leftarrow 1$ **to** $|V|$ **do**
  for $\tau \leftarrow \underline{\boldsymbol{\tau}_{\mathbf{min}}}$ **to** $\tau_{\max}$ **do**
    Calculate Cross Correlation $r_{x_v,y}(\tau)$
  end
  $\hat{\tau}_v = \underset{\tau \in \{\underline{\boldsymbol{\tau}_{\mathbf{min}}}, \ldots, \tau_{\max}\}}{\operatorname{argmax}} r_{x_v,y}(\tau)$
  Shift the word vector to maximize Cross Correlation $\hat{\mathbf{x}}_{\mathbf{v}} \leftarrow [x_v^{(1-\hat{\tau}_v)}, x_v^{(2-\hat{\tau}_v)}, \ldots, x_v^{(T-\hat{\tau}_v)}]$
end
return *Shifted Word Matrix* $\mathbf{X} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{|V|}]$

4 Experiment 1: Nowcasting

To assess the nowcasting performance, we used the actual influenza reports provided by the Japanese IDSC.

4.1 Comparable Methods

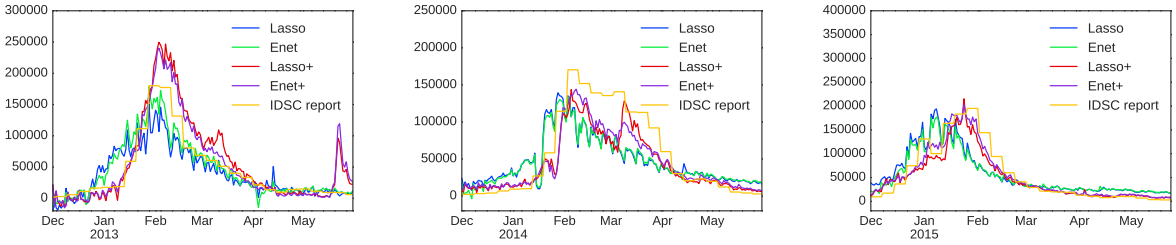
We compared four linear methods for nowcasting as shown below:

- `Lasso`: l_1-regularization method (Tibshirani, 1994; Lampos and Cristianini, 2010),
- `Lasso+`: Lasso and time shift combined method,
- `ENet`: Elastic-Net, which combines l_1-, l_2-regularization (Zou and Hastie, 2005),
- `ENet+`: Elastic-Net and time shift combined method.

All hyperparameters were tuned via five-fold cross validation in the training dataset.

Train	Season 2	Season 3	Season 1	Season 3	Season 1	Season 2	Avg.
Test	Season 1		Season 2		Season 3		
`Lasso`	0.854	0.916	0.768	0.894	0.770	0.753	0.826
`Enet`	0.900	**0.927**	0.809	**0.914**	0.792	0.805	0.884
`Lasso+`	**0.952**	0.907	0.951	0.888	0.955	**0.963**	**0.936**
`Enet+`	0.944	0.898	**0.960**	0.878	**0.967**	0.959	0.934

Table 1: Correlation between estimated values and the IDSC reports.



(a) Train in Season 2; Test in Season 1. (b) Train in Season 1; Test in Season 2. (c) Train in Season 1; Test in Season 3.

(d) Train in Season 3; Test in Season 1. (e) Train in Season 3; Test in Season 2. (f) Train in Season 2; Test in Season 3.

Figure 3: Timelines of estimated values obtained using the four methods for nowcasting.

4.2 Dataset and Evaluation Metric

The detailed dataset is described in Sec. 2. To construct the time-shifted word matrix, we set $\tau_{\max} = 60$. Our tweet corpus had a dropout period, so that we did not calculate the cross correlation with more than a 60-day shift. We employed each season's data as training data and others as test data.

The evaluation metric is based on correlation (Pearson correlation) between the estimated value and the value of the IDSC reports.

4.3 Result

Results of modeling accuracy are presented in Table 1. Correlations of our baselines, `Lasso` and `Enet`, were lower than those of previous studies. Results suggest that our dataset is more difficult than those used in earlier studies.

In contrast, time-shifted models (`Lasso+`, `Enet+`) demonstrated about 0.1 point improvement than their baseline models, indicating the contribution of time shift features.

It is noteworthy that `Lasso` type model and `Enet` type one did not differ so much. The whole trained model chose l_1 ratio parameter that is nearly equal to 1, so that the `Enet` type model became almost identical as `Lasso` type model.

Overestimation

Results showed that values in Figure 3a were overestimated in mid-May. One reason is that tweets related to news such as "**Scientists create hybrid flu that can go airborne**"[5] were popular in social media. Although tweets linked to web pages were removed during preprocessing, many tweets without links to web pages were posted by many people worried about the news. An example of such tweets is the following:

[5] `http://go.nature.com/29ATqc9`

- *What? In an attempt to make a vaccine for bird flu and swine flu had created a new strain of influenza virus? 😡 What are you doing? 😡😡😡*

In addition, the model trained in Season 2 included the word "*bird*" as one feature. This word's time shift was 15 days. Consequently, this peak occurred.

In most cases, these kinds of outlier words are not selected through model selection, but preprocessing will play an crucial role to prevent these kinds of outlier.

5 Experiment 2: Forecasting

We evaluate the forecasting performance described in Sec 3.3.

5.1 Comparable methods

`Lasso` and `Enet` have no features for predicting future values. Therefore, we use `Lasso+` and `Enet+` for forecasting. Additionally, we employ the following baseline model of `BaseLine`: $\hat{y}^{(t)}_{\text{test}} = y^{(t)}_{\text{train}}$ for comparison with our proposed models.

5.2 Dataset and Evaluation Metric

To evaluate the forecasting performance, we used the same dataset and evaluation metric as Experiment 1, except that we set the minimum time shift $\tau_{\min}$ from 1 day to 30 days.

5.3 Result

Results of forecasting accuracy are presented in Figure 4. In both models, the accuracy was superior to the baseline until around 3 weeks into the future. In addition, the accuracy for prediction one week into the future was almost identical to that in the case of $\tau_{\min} = 0$. That result might occur because the accuracy about one week future was nearly the same as that for the current state. In addition, there were many highly correlated features by shifting around 10 days into the future. Consequently, our model demonstrated equivalent performance up to 10 days into the future.

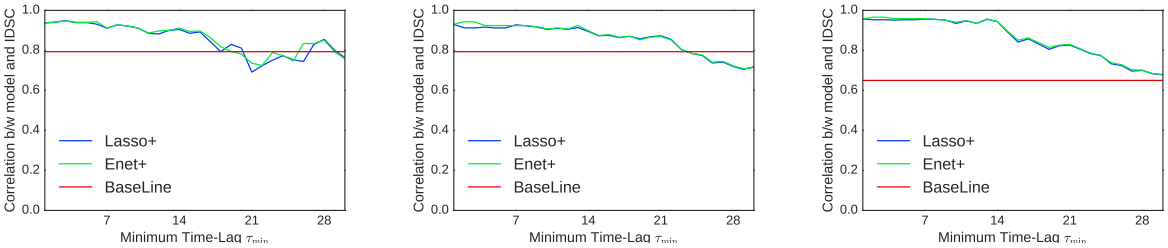
Furthermore, the forecasting performance decreased dramatically along with the increase of $\tau_{\min}$, as shown in Figure 4e. We discuss that point further in Sec. 6.

Figure 5 presents timeline plots of examples. From Figure 5a to Figure 5d are shown the values estimated by the forecasting models trained in Season 2 and tested in Season 1 for $\tau_{\min} \in \{7, 14, 21, 28\}$. The estimated values showed a consistently similar shape to that of the IDSC report. In Figure 5c, the same word, "*bird*", occurred as described in Sec. 4.3. In contrast, the weight for "*bird*" decreased in Figure 5d for that reason, the forecasting accuracy increased.

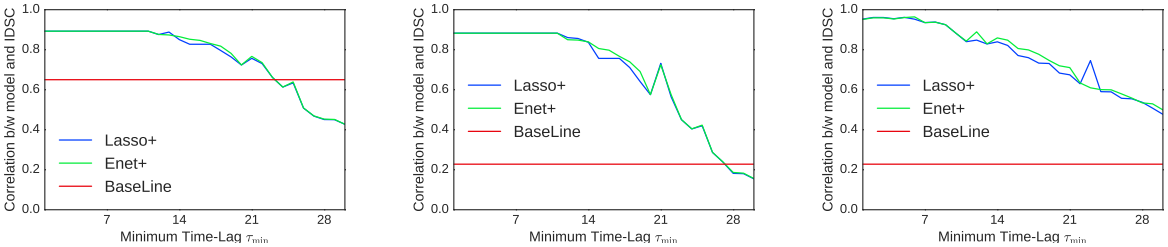
Then, from Figure 5e to Figure 5h show the values estimated by the forecasting models trained in Season 3 and tested in Season 2 for the same $\tau_{\min}$. Our models overestimated before outbreaks and underestimated after the peak of influenza epidemics. For $\tau_{\min} = 28$, this phenomenon was widely evident. We discuss that point further in Sec. 6.

6 Discussion

In general, the proposed approach (time shift operation) fitted the IDSC reports, demonstrating the basic feasibility. However, exceptions were apparent, as for the model trained in Season 3. One reason is that a gap exists in the suitable time shift widths between the train (Season 3) and the other (Seasons 1 and 2). `Lasso+` model trained in Season 3 selected the words, "*fever*" with $\hat{\tau}_{\textit{fever}} = 16$, "*vaccination*" with $\hat{\tau}_{\textit{vaccination}} = 55$, "*absent*" with $\hat{\tau}_{\textit{absent}} = 10$, and others as features. These words have high correlations only in Season 3, with poor correlation in other seasons. The most drastic example is "*vaccination*" with $\hat{\tau}_{\textit{vaccination}}$, (over 0.849 correlation in Season 3). This word is adversely affected by other seasons (0.313 correlation in Season 1 and 0.04 correlation in Season 2). The reason for the lost correlation was that $\hat{\tau}_{\textit{vaccination}}$ in Season 3 differed from that of other seasons. This phenomenon suggests that "*vaccination*" is just an annually cycling word. Neither the cycle of "*vaccination*" nor that of influenza is fixed, bringing us different time lags.

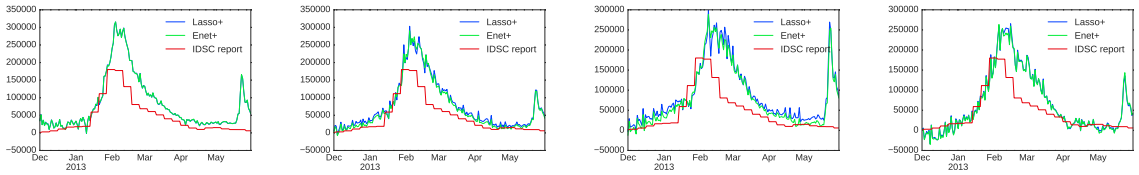


(a) Train in Season 2; Test in Season 1. (b) Train in Season 1; Test in Season 2. (c) Train in Season 1; Test in Season 3.

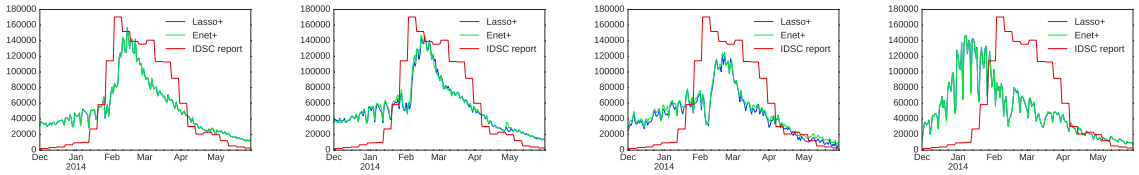


(d) Train in Season 3; Test in Season 1. (e) Train in Season 3; Test in Season 2. (f) Train in Season 2; Test in Season 3.

Figure 4: Correlation between estimated values using the two methods for forecasting.



(a) $\tau_{\min} = 7$ in Fig. 4a. (b) $\tau_{\min} = 14$ in Fig. 4a. (c) $\tau_{\min} = 21$ in Fig. 4a. (d) $\tau_{\min} = 28$ in Fig. 4a.



(e) $\tau_{\min} = 7$ in Fig. 4e. (f) $\tau_{\min} = 14$ in Fig. 4e. (g) $\tau_{\min} = 21$ in Fig. 4e. (h) $\tau_{\min} = 28$ in Fig. 4e.

Figure 5: Timelines of values estimated using the two methods for forecasting and the IDSC reports in each $\tau_{\min}$.

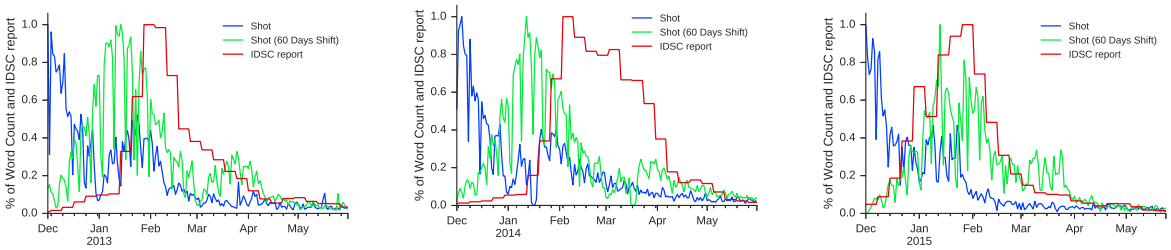
This inconsistency of time shifts also affected the forecasting performance directly. As shown in Figure 4e, the forecasting performance was decreased dramatically against the increase of $\tau_{\min}$. In spite of the word "*shot*" is the largest weighted feature in the case of $\tau_{\min} = 21$ and Train in Season 3, these word correlations were 0.310 in Season 1 and 0.03 in Season 2. Consequently, it caused a considerable decrease of the forecasting accuracy. In contrast, some words, such as "*fever*" and "*symptom*", showed consistently similar time shifts.

A technique to distinguish actual forecasting words such as "*fever*", and noises (simple year cycle words), "*vaccination*" is highly anticipated for use in the near future. If multiple-year training sets were available, one could filter out such noisy words.

Although some room for improvement remains, the basic feasibility of the proposed approach has been demonstrated. The time shift was effective for social media based surveillance. In addition, the model enables prediction.

7 Related Work

To date, numerous web based surveillance systems have been proposed, targeting the common cold (Kitagawa et al., 2015), drug side effects (Bian et al., 2012), cholera (Chunara et al., 2012), *E. Coli* (Diaz-Aviles et al., 2012), problem drinking (MA et al., 2012), smoking (Prier et al., 2011), campylobacteriosis (Chester et al., 2011), dengue fever (Gomide et al., 2011), and HIV/AIDS (Ku et al., 2010). Influenza has especially drawn much attention from earlier studies (Ginsberg et al., 2009; Polgreen et al., 2009;



(a) Season 1. (b) Season 2. (c) Season 3.

Figure 6: Frequencies of "*Shot*" in respective seasons.

Hulth et al., 2009; Corley et al., 2010) to current Twitter-based studies (Aramaki et al., 2011; Collier et al., 2011; Chew and Eysenbach, 2010; Lampos and Cristianini, 2010; Culotta, 2013).

Because of great variance in data resources and evaluation manner (region, year, only winter or all seasons), a precise comparison would be difficult and meaningless, Culotta (Culotta, 2013) and Ginsberg (Ginsberg et al., 2009) are apparently better than the others in US (correlation ratios $= 0.96$ and 0.94, respectively). Aramaki et al. (2011) achieved the best score for Japan (correlation ratio $= 0.89$). This study also examined Twitter data in Japan, and achieved competitive results for nowcasting. Another aspect of reviews of related studies is the manner of tweet counting. In earlier studies, a simple word counting, the direct number of tweets, is considered an index of the degree of disease epidemics. However, such a simple method is adversely affected by the huge numbers of noisy tweets. Currently, counting approaches of two types have been developed: (1) a classification approach (Kanouchi et al., 2015; SUN et al., 2014; Aramaki et al., 2011) aimed at extracting only tweets including patient information, and (2) a regression approach (Lamb et al., 2013; Culotta, 2010; Lampos and Cristianini, 2010; Paul and Dredze, 2011) that handles multiple words to build a precise regression model.

The proposed study fundamentally belongs among regression approaches, which explore optimal weight perimeters for each word. An important difference is that this study handles one more parameter for each word: time shift (days). To handle many parameters, we first ascertain the best time shift widths. Then we explore weight parameters using L1 or elastic net. It is noteworthy that this study does not employ any classification method, engaging a room to improve by incorporation with classification techniques.

8 Conclusions

This study proposed a novel social media based influenza surveillance system using forecasting words that appear in Twitter usage before main epidemics occur. First, for each word, the optimal time lag was explored, which maximized the cross correlation to influenza epidemics. Then, we shifted a matrix consisting of word frequencies at different time points by each optimal time lag. Using the time-shifted word matrix, this study produced and evaluated a nowcasting model and forecasting model designed to predict the number of influenza patients. In the experimentally obtained results, the proposed model achieved the best nowcasting performance to date (correlation ratio 0.93) and practically sufficient forecasting performance (correlation ratio 0.91 in the 1-week future prediction, and correlation ratio 0.77 in 3-week future prediction). This report is the first of the relevant literature describing a model that enables prediction of future epidemics. Furthermore, the model has much room for potential application to prediction of other events.

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1576.

Jiang Bian, Umit Topaloglu, and Fan Yu. 2012. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the International Workshop on Smart Health and Wellbeing (SHB)*, pages 25–32.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PloS one*, 10(10):e0139701.

Tammy L. Stuart Chester, Marsha Taylor, Jat Sandhu, Sara Forsting, Andrea Ellis, Rob Stirling, and Eleni Galanis. 2011. Use of a web forum and an online questionnaire in the detection and investigation of an outbreak. *Online J Public Health Inform*, 3(1):1–7.

C. Chew and G. Eysenbach. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5:e14118.

Rumi Chunara, Jason R. Andrews, and John S. Brownstein. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *Am J Trop Med Hyg.*, 86(1):39–45.

Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. 2011. Omg u got flu? analysis of shared health messages for bio-surveillance. *J Biomed Semant*, 2.

Courtney D. Corley, Diane J. Cook, AR Mikler, and Karan P. Singh. 2010. Text and structural data mining of influenza mentions in web and social media. *International Journal of Environmental Research and Public Health*.

Aron Culotta. 2010. Detecting influenza outbreaks by analysing twitter messages. *CoRR*, abs/1007.4748.

Aron Culotta. 2013. Lightweight methods to estimate influenza rates and alcohol sales volume from twitter messages. *Lang. Resour. Eval.*, 47(1):217–238.

Ernesto Diaz-Aviles, Avaré Stewart, Edward Velasco, Kerstin Denecke, and Wolfgang Nejdl. 2012. Towards personalized learning to rank for epidemic intelligence based on social media streams. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 495–496.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4.

Janaína Gomide, Adriano Veloso, Wagner Meira, Jr., Virgílio Almeida, Fabrício Benevenuto, Fernanda Ferraz, and Mauro Teixeira. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the International Web Science Conference (WebSci)*, pages 3:1–3:8.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4(2):e4378, 02.

Shin Kanouchi, Mamoru Komachi, Naoaki Okazaki, Eiji Aramaki, and Hiroshi Ishikawa. 2015. Who caught a cold ? - identifying the subject of a symptom. In *Proceedings of the Association for Computer Linguistics (ACL)*, pages 1660–1670.

Yoshiaki Kitagawa, Mamoru Komachi, Eiji Aramaki, Naoaki Okazaki, and Hiroshi Ishikawa. 2015. Disease event detection based on deep modality analysis. In *Proceedings of the Association for Computer Linguistics*, pages 28–34.

Yungchang Ku, Chaochang Chiu, Yulei Zhang, Li Fan, and H. Chen. 2010. Global disease surveillance using social media: Hiv/aids content intervention in web forums. In *Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on*, pages 170–170.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the social web. In *Proceedings of International Workshop on Cognitive Information Processing (CIP)*, pages 411–416.

Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. 2015. Advances in nowcasting influenza-like illness rates using search query logs. *Scientific reports*, 5.

Moreno MA, Christakis DA, Egan KG, Brockman LN, and Becker T. 2012. Associations between displayed alcohol references on facebook and problem drinking among college students. *Archives of Pediatrics and Adolescent Medicine*, 166(2):157–163.

M. J. Paul and M. Dredze. 2011. You are what you tweet: Analysing twitter for public health. In *Processing of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Michael J Paul, Mark Dredze, and David Broniatowski. 2014. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, and Robert A. Weinstein. 2009. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448.

Kyle W. Prier, Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. 2011. Identifying health-related topics on twitter: An exploration of tobacco-related tweets as a test topic. In *Proceedings of the International Conference on Social Computing, Behavioral-cultural Modeling and Prediction (SBP)*, pages 18–25.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 851–860.

Xiao SUN, Jiaqi YE, and Fuji REN. 2014. Real time early-stage influenza detection with emotion factors from sina microblog. In *Proceedings of the Workshop on South and Southeast Asian NLP in the International Conference on Computational Linguistics (COLING)*, pages 80–84.

Robert Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

H. Zou and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67:301–320.