# After the Boom No One Tweets: Microblog-based Influenza Detection Incorporating Indirect Information

Shoko Wakamiya Nara Institute of Science and Technology, Nara, Japan wakamiya@is.naist.jp Yukiko Kawai Kyoto Sangyo University Kyoto, Japan kawai@cc.kyoto-su.ac.jp Eiji Aramaki Nara Institute of Science and Technology, Nara, Japan aramaki@is.naist.jp

# ABSTRACT

The recent rise in popularity and scale of social media services (SNSs) has created the growing necessity for SNS-based information extraction systems. A popular application of SNS data is health surveillance for predicting the outbreak of epidemics by detecting diseases from text messages posted on SNS platforms. Such applications share the same logic: they incorporate SNS users as social sensors. Such social sensor approaches also share a common problem: SNS-based surveillance can be reliable if sufficiently numerous users are active, but small inactive populations produce inconsistent results. To overcome this problem, this paper presents a novel approach using indirect information covering both urban areas and rural areas within the posts. This system uses not only direct information, but also indirect information that mentions other places. Indirect information is less reliable (too noisy or too old) than direct information. Therefore, we do not use the described data directly. Instead, we consider that indirect information inhibits direct information. For example, when indirect information appears often, we consider that everyone already has a known disease, leading to a small amount of direct information. In the experiment using three years' long collection of tweets (7 million influenza-related tweets in Japanese) described herein, this proposed approach improved the detection performance not only in rural cities, but also in urban cities, thereby demonstrating the feasibility of this approach.

# **CCS Concepts**

Information systems→Information systems applications;
Computing methodologies→Natural language processing;
Applied computing→Life and medical sciences;

#### Keywords

Influenza surveillance, location mention, twitter, social network, spatial analysis

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org. EDB, October 17-19, 2016, Jeju Island, Republic of Korea © 2016 ACM. ISBN 978-1-4503-4754-9/16/10...\$15.00 DOI: http://dx.doi.org/10.1145/3007818.3007822



Figure 1: Population bias in Twitter-based Influenza surveillance. Most Twitter users are in urban cities (such as Tokyo and Osaka). Other cities are adversely affected by a shortage of data, which biases influenza detection there.

# 1. INTRODUCTION

The increased use of social media platforms entails more widely shared personal information. Twitter, a micro-blogging platform that enables users to communicate by updating their status using 140 or fewer characters, has attracted much attention of researchers and service developers because Twitter can be a valuable personal information resource. These approaches share the important premise that Twitter users can be human sensors for event detection [1]. The feasibility of such approaches has been demonstrated on various occasions, such as earthquakes, outbreaks of disease, and stock market fluctuations.

This study particularly examines such applications for detecting disease epidemics, by taking advantage of the swiftness of the information transmission of Twitter, which outperforms other traditional methods of medical reporting means. Numerous Twitter-based disease detection and prediction systems have been developed worldwide. Such systems have also been demonstrating some weaknesses. One important deficit is the imbalance of the population distribution because most SNS users reside in urban areas and because analysts have difficulty getting a sufficient amount of data from rural areas. Especially, Japan's user population is strongly concentrated in a few central cities, such as Tokyo and Osaka. Other users live outside the area, in less populated regions in Japan. This population bias causes difficulties in the consistency of performance. Fig. 1 shows the geographic distribution of our dataset (7,666,201 influenza-related tweets during 2012-2015). Rural areas have fewer young people than cities do. Therefore, even fewer SNS users are available. More detailed data are presented in Fig. 2. Except for a few high-



Figure 2: Numbers of influenza patients (bar graphs) and tweets related to influenza (line) in each prefecture. 47 prefectures are ordered based on prefecture codes.



Figure 3: Most social sensor-based approaches consider people as sensors (right and center). Although social sensors exploit only direct information, the proposed activity uses indirect information (right).

population cities, most areas have fewer tweets. Some such areas have numerous influenza patients.

To overcome this problem, we use information from broader targets than in earlier studies. One solution is using indirect information that has been discarded in previous studies. Examples of such indirect information are described below.

#### (1) My grandma *in Kyoto* is in bed with flu.

#### (2) NEWS: classes in Osaka have been closed because of the flu

The fundamental idea is presented in Fig. 3. Although tweets are concentrated in urban areas, indirect information covers wider areas. Indirect information is unreliable (sometimes too noisy or too old). In example (1), the time '*the grandma*' caught flu is unknown. In example (2), the flu had already spread to the area. Because of the difficulties presented above, earlier studies have not used such indirect information to any great degree.

An example of the tweet timeline is presented in Fig. 4. Direct information (black line) shows a similar timeline to the gold standard timeline (red area). However, after the peak of epidemics, the amount of direct information decreases, leading to underestimation errors. In contrast, indirect information (red line) shows complex phenomena: it has many peaks, especially before peaks. Apparently, indirect information is difficult to use.

This study uses a different approach that specifically examines the relation between indirect information and the human motivation to tweet. We consider after the boom of influenza, the topic of influenza has become out-of-fashion, inhibiting people's motivation to tweet about the flu. In other words, this study assumes that *people prefer reporting new information*, and that they are *insensitive to already-propagated information*.



Figure 4: Amounts of direct and indirect information in a tweet timeline in Hokkaido in 2013. Direct information (black line). Indirect information (red line). Gold standard timeline (red area). The x-axis shows the date. The y-axis indicates the number of tweets and the number of patients (normalized by the max value in the SEASON).

Table 1: Area resolution of surveillance

	Target (# of areas)	Data size (million tweets)
Aramaki [16]	Japan (1 area)	300
Achrekar [27]	US (10 areas)	1.9 *
Culotta [28]	US (1 area)	0.5
Kanouch [29]	Japan (1 area)	300
De Quincy [30]	Europe (1 area)	0.14
Doan [31]	US (1 area)	24 *
Szomszor [32]	Europe (1 area)	3

\* indicates the number of users (million users)

Another difficulty is detection of the degree of the propagated information. This study specifically examines the amount of indirect information because it indicates a person in different places also knows of the event. Consequently, this study examines a hypothesis: *the degree of propagation (popularity) is correlated with the amount of indirect information.* 

In the early stage of a season, most people report influenza precisely. Because the indirect information is propagated widely, most people know of the influenza epidemic, and become insensitive to the event. We designate such inactivated people as

ATT	Duration	2012/08/02-2016/01/03
ALL	# of tweets (Size)	7,666,201 (2.275 GB)
SEASON2012	Duration	2012/11/01-2013/05/31
	# of tweets (Size)	1,959,610 (729.4 MB)
SEASON2013	Duration	2013/11/01-2014/05/31
	# of tweets (Size)	501,542 (143.7 MB)*
SEASON2014	Duration	2014/11/01-2015/05/31
	# of tweets (Size)	2.736.685 (808.2 MB)

Table 2: Data description.

\*We were unable to collect sufficient tweets in SEASON2013 because of the specification change of Twitter API.

*trapped sensors.* This study investigates the degree to which this model improves the event detection performance.

This study uses a Twitter platform based on Japanese language in particular. However, the results are not expected to depend on specific platform or language because no platform and languagespecific technique is used.

# 2. RELATED WORK

# 2.1 Social Sensors for Health-related Events

Social media are used to detect various events such as earthquakes [1-3], political elections [4-6], and stock prices in a market [7]. Among the various applications, health-related event detection has been drawing much attention from researchers in areas such as air pollution [8], online doctor reviews [9], West Nile virus [10], cholera [11], *E. coli* outbreak [12], Dengue fever outbreak [13], and influenza [10, 14–32]. One review of the literature has reported that half of SNS-based surveillance are related to influenza (15 of 33 papers) [26]. That is true because influenza has been a major worldwide public health concern. Especially, unexpected influenza pandemics, which have been experienced three times already in the 20th century (e.g. "Spanish flu"), are global issues.

Twitter is the most frequently used social medium for influenza detection [10, 14–32]. Studies have consistently demonstrated high correlation between the number of influenza patients and the actual influenza-related tweets. However, most studies target only country-level detection. Furthermore, surveillance of detailed areas is rarely challenged (Table 1). One reason is the volume shortage of tweets in small areas. Therefore, it remains unknown whether a small rural area can achieve the same high-performance or not. One point of this study is to investigate performance in areas with small populations.

#### 2.2 Location Estimation

Location estimation including estimation of one's place of residence is an important issue of this study. Although the simplest and most reliable is to use GPS information, many difficulties can arise. For instance, many users turn off this function to maintain privacy of their information. As a result, location estimation from the SNS original text is necessary. Related studies have identified two difficulties in location estimation of SNS texts: (1) the first is detecting a location name in tweet messages; (2) the second is to disambiguate the location names.

For the first, a collection of location names is necessary. Usually, Wikipedia has been used as the basis of a location name dictionary. We also used a location name dictionary obtained from Japanese Wikipedia.

Table 3: Proposed scope.

Tweet	P/N	D/I
BBC News: Okinawa is influenza pandemic	Р	Ι
Okinawa suffers a major outbreak of influenza	Р	D
RT: My mother got flu today	Р	Ι
I got influenza shot today	Ν	D
Dr said influenza will be late in this season	N	Ι

P/N denotes positive (P) or negative (N); D/I denotes Direct information (D) or Indirect information (I). Note that we used retweet (RT), too, in the same manner as normal tweets (non-RT tweets).



Figure 5: Influenza surveillance system modules.

To resolve the second difficulty, several researchers have examined location name disambiguation methods [33]. Locationindicative words from tweet data are found by calculating the information gain ratios. That earlier research effort shows that words improved the estimation performance of the user location. They concluded that the procedure requires little memory: it is fast. Moreover, lexicographers can use it to extract location-indicative words. A probabilistic framework was developed to quantify the spatial variation manifested in search queries [34], which brings them to spatial probabilistic distribution models. One study [35] estimated geographic regions from unstructured, non-georeferenced text by computing a probability distribution over the Earth's surface. Another study [36] estimated a city-level user location based purely on a content of tweets, which might include reply tweet information, without the use of any external information, such as a gazetteer or IP information. Two unsupervised methods [37] have been proposed based on notions of Non-Localness and Geometric-Localness to prune noisy data from tweets. One report [38] described language models of locations using coordinates extracted from geotagged Twitter data.

Although this study uses geocoding services provided by Google, incorporating such techniques can support future studies.

# **3. MATERIAL**

Data used for this study are of two types: Twitter data for the proposed system (Section 3.1), and the timeline report of the number of influenza patients, which are gold standard data (Section 3.2). The dataset is available at the data archive site. The data

include the number of tweets with the gold standard data. We have also developed visualization and a data distribution site<sup>1</sup>.

### 3.1 Tweet Data

Our data are a collection of influenza-related tweets for three years. We have been collecting influenza-related tweets written in Japanese via the Twitter streaming API for five years (from 2012/08/02 to 2016/01/03). All tweets contain influenza-related Japanese keywords "I-N-FU-RU (flu in Japanese)" or "I-N-FU-RU-E-N-ZA (influenza in Japanese)". These data include noise tweets that include a keyword, but do not index an influenza patient. Examples of such noise tweets are mentions of "influenza vaccination." This study calls such noisy tweets *influenza negative tweets*. To filter out such *influenza negative tweets*, we have trained the SVM-based classifier using the training set. The training set consists of 5,000 tweets extracted randomly from the corpus. Human annotators annotated these data.

Because influenza epidemics appear in the winter, we split the data as shown below.

- SEASON2012: 2012/11/01-2013/05/31
- SEASON2013: 2013/11/01-2014/05/31
- SEASON2014: 2014/11/01-2015/05/31

Details of the test data are presented in Table 2.

# 3.2 Gold Standard Data

The gold standard data are weekly reports from the Infectious Disease Surveillance Center (IDSC). The report presents the number of influenza patients for every Japanese prefecture (47 areas). This test set enables week-based evaluation in 47 areas.

#### 4. INFLUENZA SURVEILLANCE SYSTEM

The system comprises three modules as shown in Fig. 5: a positive/negative (P/N) classification module (Section 4.1), a location detection module (Section 4.2) and data aggregation module (Section 4.3). For the aggregation, we use two methods: a LINEAR model using three types of location information (Section 4.3.1) and a TRAP model (Section 4.3.2).

#### 4.1 NLP Module: Positive or Negative

Using the tweet corpus described in Section 3, we built a classifier that judges whether a given tweet is positive or negative. This task is a sentence binary classification (such as spam e-mail filtering). We used a SVM-based classifier under the bag-of-words (BOW) representation. We split a Japanese sentence into a sequence of words using a Japanese morphological analyzer, MeCab (ver.0.98) with IPADic (ver.2.7.0)<sup>2</sup>. A polynomial kernel (d=2) is used. The other parameters are the same as described in an earlier report [16]. Several earlier studies [29] have not used retweet (RT) texts, but they are used for the present study.

To build the training set, a human annotator assigned one of two labels: positive or negative. In this labeling, we regarded a tweet that meets the following two conditions as positive data.

#### • Condition 1: Patient

One or more people who have influenza are likely to be present around the tweet person. Here, we regard "around" as a distance in the same city. For cases in which the distance is unknown, we regard it as negative. Because of this annotation policy, the re-tweet type message is negative.

# • Condition 2: Tense

<sup>1</sup> http://mednlp.jp/influ/

<sup>2</sup> http://taku910.github.io/mecab/

The tense should be the present tense (current) or recent past. Here, we define the "recent past" as the prior one-day period: "yesterday".

Using the constructed influenza corpus (Table 3), we trained the SVM-based classifier. The corpus comprises pairs of sentences and a label (positive or negative). For more precise information related to the corpus, previous reports are helpful [16].

# 4.2 Location Detection Module (Direct or Indirect)

#### We use three methods for location detection.

#### 4.2.1 GPS Information (GPS)-Direct Information

A tweet contains GPS data if a Twitter user allows use of the location function. However, most users turn off this function for the privacy reasons. Currently, the ratio of tweets with GPS information is only 0.46% (=35,635/7,666,201) in our dataset.

# 4.2.2 Profile Information (PROF)-Direct Information

Several Twitter users describe the address in profile. We regard the Tweet person as near the profile address. The ratio of tweets with profile location is 26.2% (=2,010,605/7,666,201). This information is used in the study [16]. To disambiguate the location names, we use a Geocoding service<sup>3</sup> provided by Google Maps<sup>4</sup>. Specifically, we sent queries about Twitter user's locale to Google Maps and obtained returned results in JSON format. We wrote a simple parser in Python to parse these returned results to get information about the country.

#### 4.2.3 Indirect Information (IND)

Several tweets contain the location name in the contents, such as "My friend in OSAKA caught flu." This study handles such indirect information. To detect the location name in contents, we use a location name list, consisting of prefecture names and famous landmarks. The ratio of tweets with indirect information is 4.7% (=362,349/7,666,201).

The conflict of three geographic data is resolved as shown in Fig. 5. We use the location if GPS information is available. Otherwise, if a user profile information includes address data, then we use that information. The address data are geocoded by the geocoding service API provided by Google. Otherwise if the tweet's content contains location name (prefecture names), we consider it as the indirect information in the prefecture.

#### 4.3 Aggregation Module (LINEAR or TRAP)

A difficulty hindering the combination of different resources is how to combine them. This study investigates the two methods: (1) a simple aggregation (linear model) and (2) a trap model, which is the proposed method.

#### 4.3.1 LINEAR Model

A simple method to use indirect information is how to aggregate different types of information. In this model, we weight the direct information as more important than the indirect information.

We formalize the number of patients  $I_{LINEAR}(a, t)$  in area *a* at day *t* as follows:

#### $I_{LINEAR}(a,t)$

 $= w_{GPS} \cdot GPS(a, t) + w_{PROF} \cdot PROF(a, t) + w_{IND} \sum_{b \in A} IND(a, b, t)$ (1)

<sup>4</sup> http://maps.google.com

<sup>&</sup>lt;sup>3</sup> http://www.geocoding.jp/

where GPS(a,t) is the number of tweets with GPS information. PROF(a,t) is the number of tweets with profile information. IND(a,b,t) is the number of tweets with indirect information.  $w_{GPS}$ ,  $w_{PROF}$  and  $w_{IND}$  are weight parameters.

# 4.3.2 TRAP Model

- This model includes the following two assumptions.
- People prefer a new event, and are insensitive to an alreadypropagated event.
- The degree of propagation (popularity) is correlated with the amount of indirect information.

The first assumption derives from human nature: people hesitate to inform others of an already known fact. For example, if the Twitter stream is full of repeated influenza information, then such a situation dampens enthusiasm to tweet similar information.

The second assumption comes from the features of Twitter. Most indirect information consists of RT or news information, which tends to delay the direct information. The volume of this type of information corresponds to the volume of people who never tweet.

Based on those two assumptions, in the early stage of a season, most social sensors are "activated" to report the influenza precisely (Fig. 6(a)). Because the indirect information spreads widely, most people become "inactivated" to the event (Fig. 6(b)). We designate such inactivated people as *trapped sensors*. Under these circumstances, even the number of influenza tweets is small. The number of patients might be larger than the tweet volume because a *trapped sensor* might disregard the influenza.

We formalize the number of patients  $I_{TRAP}(a, t)$  in area *a* at day *t* using a popularity function, pop(a,t), as

$$I_{TRAP}(a,t) = \frac{I_{LINEAR}(a,t)}{w_{USERS} \cdot N_a - w_{TRAP} \cdot \log(pop(a,t)+1)}$$
(2)  
$$pop(a,t) = \sum_{c=1}^{t} IND(a,c)$$

where  $I_{LINEAR}(a, t)$  is the linear model (see equation (1) in Section 4.3.1), variable  $N_a$  is set based on the number of the potential active tweeting users defined by the number of tweets.  $w_{USERS}$  and  $w_{TRAP}$  are weight parameters. A function, pop(a,t), represents the degree of popularity of a crowd's attention to the influenza.

# 5. EXPERIMENTS

These experiments used the Japanese inflectional disease data. **5.1 Test Data and Gold Standard Data** 

For the experiments, we used the corpus described in Section 3.1 as a test set. As gold standard data, we used data reported by the Infectious Disease Surveillance Center (IDSC). The report has been released once a week. To conduct a daily basis evaluation, the average value was used for interpolation.

Because influenza is a seasonal disease, we split the tweet data into three seasons based on the IDSC's data as introduced in Section 3.2.

#### 5.2 Methods

We compared four methods as described below.

• **TRAP**: TRAP is the proposed model, which detects disease epidemics by considering the balance between direct information (GPS information and profile information) and indirect information (referred location). In these experiments, we set  $N_a$  to a value based on the number of potential active tweeting users for equation (2) in Section 4.3.2. Then we set the weight parameters  $w_{USERS}$  and  $w_{TRAP}$  to 0.05 and 0.2, respectively, based on results of preliminary experiments.

Table 4: Experimental methods

Method	NLP	GPS	PROF	IND
TRAP+NLP	1	1	1	1
TRAP		1	1	1
LINEAR+NLP	1	1	1	1
LINEAR		1	1	1
EMNLP2011 [16]	1	1	1	
(BASELINE+PROF+NLP)	•	•	•	
BASELINE+PROF		1	1	
BASELINE +NLP	1	1		
BASELINE		1		

\*NLP is the positive/negative classifier, GPS is GPS information, LOC is profile information, and IND is indirect information.



Fugue 6: Concept image of TRAP model. (a) People actively report the influenza before epidemics. (b) However, most people lose interest to share the direct information after epidemics, because lots of indirect information have been already existed. In the proposed model, we call such people Trapped Sensors.

- **LINEAR:** LINEAR is a model that uses GPS information, profile information, and indirect location information together. In these experiments, weight parameters *w*<sub>GPS</sub>, *w*<sub>PROP</sub> and *w*<sub>IND</sub> in equation (1) in Section 4.3 are set to 1.0.
- BASELINE+PROF: This is a baseline method presented in [16]. This approach uses GPS information and profile location. *I*<sub>BASE+PROF</sub>(*a*,*t*)=GPS(*a*,*t*)+PROF(*a*,*t*).
- **BASELINE:** This is the simple baseline, using only GPS information.

 $I_{BASE}(a,t) = GPS(a,t).$ 

In addition to evaluation of the effectiveness of positive/negative classification (NLP technique), we also conducted with/without test. This test brings us the 8 (= $4\times2$ ) as shown in Table 4.

#### **5.3 Evaluation Metric**

The evaluation metric is based on correlation (Pearson correlation coefficient) between the gold standard values and the estimated values. The correlation coefficient *Correl(.)* for a given data array consisting of the gold standard data and the data by a comparable method, i.e. { $(GOLD_{a,b}TRAP_{a,t})$ }  $(a \in A, t \in SEASON$ 



Figure 7: Temporal changes of positive influenza tweets for three SEASONs in 6 prefectures, Japan. The y-axis shows the ratio of tweets; the x-axis is time from the beginning of SEASON2012 to the end of SEASON2014. The red line represents by our system. The light red bar shows the gold standard data. The black dotted line shows results by baseline.



Figure 8: Relation of the number of tweets (blue bar) and correlation coefficient of TRAP+NLP (red line) and EMNLP2011 (dotted black line) for each area. Areas are ordered by populations based on the number of tweets. The x-axis shows the area. The y-axis indicates the correlation coefficient (left side) and the number of tweets (right side). In most of all areas, the proposed approach (TRAP+NLP) shows a higher correlation ratio than the previous system BASELINE+PROF, which is proposed in EMNLP2011.

2012) is computed as a data array consisting of values of the gold standard data and values computed using TRAP model in an area *a* on each day *t* in SEASON2012.

Gold standard data are based on the number of patients. The system outputs the values based on the number of tweets. Although the scales mutually differ, they are not biased under the correlation-based evaluation. We regard high correlation as |r| > 0.7, medium correlation as  $0.4 < |r| \le 0.7$ , and low correlation as  $|r| \le 0.4$ .

Evaluation is done for four durations: (1) SEASON2102, (2) SEASON2013, (3) SEASON2014, and (4) SEASON-TOTAL (all (1)-(3)).

# 5.4 Results

Table 5 presents the results. Table 5(a) and Table 5(b) respectively present the correlation coefficients of methods with and without NLP for the gold standard data. Specifically, we discuss these results in terms of contributions of NLP-based classification (Section 5.4.1), profile location (Section 5.4.2), LINEAR model (Section 5.4.3), and TRAP model (Section 5.4.4).

# 5.4.1 Contribution of NLP-based Classification (TRAP vs. TRAP+NLP)

To evaluate the contribution of NLP for positive and negative classification, we compare results of TRAP in Table 5(b) and TRAP+NLP in Table 5(a). Although both methods are strongly correlated with the gold standard data, mostly TRAP+NLP (r=0.70 in SEASON-TOTAL) is higher than TRAP (r=0.64). This result demonstrates the contribution of NLP.

Although almost half of the tweets are removed using NLP classification, it works well for estimating influenza epidemics. It might indicate that the NLP classification in this domain (influenza or not) is easy, so it must be improved.

Not only for TRAP+NLP but also all other methods with NLP (BASELINE+NLP, EMNLP2011, and LINEAR+NLP) achieved better detection performance using the NLP classifier.

# 5.4.2 Contribution of Profile Information

(BASELINE+NLP vs. BASELINE+PROF+NLP)

To evaluate the contribution of profile information, we compare BASELINE+NLP and EMNLP2011 (BASELINE+ PROF+NLP). As shown in Table 5(a), the correlation coefficient of EMNLP2011 (r=0.69 in SEASON-TOTAL) is much higher than that of the

Table 5: Values of correlation coefficient r of methods with/without NLP. Moderate correlation (|r| > 0.7) is shown in bold. Highest correlation coefficient in each target area and each SEASON is underlined.

(a) With NLP					
Target	Method	SEASON 2012	SEASON 2013	SEASON 2014	SEASON TOTAL
	TRAP+NLP	0.76	0.70	0.69	0.70
A 11	LINEAR+NLP	0.70	0.55	0.53	0.50
All areas	EMNLP2011	0.74	0.68	0.67	0.69
	BASELINE+NLP	0.33	0.37	0.48	0.36
High	TRAP+NLP	0.80	0.77	0.72	0.75
population	LINEAR+NLP	0.78	0.65	0.64	0.64
areas (Top 10)	EMNLP2011	0.80	<b>0.77</b>	0.71	0.75
	BASELINE+NLP	0.55	0.60	0.63	0.53
Low population areas (Top 10)	TRAP+NLP	<u>0.75</u>	0.66	0.71	0.69
	LINEAR+NLP	0.62	0.46	0.48	0.43
	EMNLP2011	0.70	0.61	0.65	0.64
	BASELINE+NLP	0.21	0.26	0.35	0.25

$\begin{array}{c c c c c c c c c c c c c c c c c c c $	(b) Without NLP					
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	Target	Method	SEASON	SEASON	SEASON	SEASON
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$			2012	2013	2014	TOTAL
All areas     LINEAR     0.65     0.48     0.53     0.48       BASELINE+PROF     0.69     0.59     0.66     0.64       BASELINE     0.29     0.34     0.48     0.35       High population areas (Top 10)     TRAP     0.75     0.69     0.70     0.70       BASELINE     0.72     0.60     0.63     0.61       BASELINE+PROF     0.75     0.69     0.70     0.70       INEAR     0.75     0.69     0.70     0.70       BASELINE+PROF     0.75     0.69     0.63     0.61       LINEAR     0.72     0.60     0.63     0.50       Low     TRAP     0.71     0.61     0.53     0.57       population     INEAR     0.58     0.41     0.46     0.40       areas     BASELINE+PROF     0.65     0.52     0.65     0.59		TRAP	0.72	0.63	0.64	0.64
$\begin{array}{c c c c c c c c c c c c c c c c c c c $	A 11	LINEAR	0.65	0.48	0.53	0.48
BASELINE     0.29     0.34     0.48     0.35       High population areas (Top 10)     TRAP     0.75     0.69     0.70     0.70       BASELINE     0.72     0.60     0.63     0.61       BASELINE+PROF     0.75     0.69     0.70     0.70       BASELINE+PROF     0.75     0.69     0.70     0.70       LINEAR     0.44     0.56     0.63     0.50       Low     TRAP     0.71     0.61     0.53     0.57       population     LINEAR     0.58     0.41     0.46     0.40       areas     BASELINE+PROF     0.65     0.52     0.65     0.59	All areas	BASELINE+PROF	0.69	0.59	0.66	0.64
$\begin{array}{c c c c c c c c c c c c c c c c c c c $		BASELINE	0.29	0.34	0.48	0.35
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	High population areas (Top 10)	TRAP	0.75	0.69	0.70	0.70
$\begin{array}{c} \mbox{areas} \\ (Top 10) & \hline BASELINE + PROF \ 0.75 & 0.69 & 0.70 & 0.70 \\ \hline BASELINE & 0.44 & 0.56 & 0.63 & 0.50 \\ \hline Low & \hline TRAP & 0.71 & 0.61 & 0.53 & 0.57 \\ \hline population & \hline LINEAR & 0.58 & 0.41 & 0.46 & 0.40 \\ \hline BASELINE + PROF & 0.65 & 0.52 & 0.65 & 0.59 \\ \hline \end{array}$		LINEAR	0.72	0.60	0.63	0.61
$\begin{array}{c c c c c c c c c c c c c c c c c c c $		BASELINE+PROF	0.75	0.69	0.70	0.70
Low     TRAP     0.71     0.61     0.53     0.57       population     LINEAR     0.58     0.41     0.46     0.40       areas     BASELINE+PROF     0.65     0.52     0.65     0.59		BASELINE	0.44	0.56	0.63	0.50
population     LINEAR     0.58     0.41     0.46     0.40       areas     BASELINE+PROF     0.65     0.52     0.65     0.59	Low population areas (Top 10)	TRAP	0.71	0.61	0.53	0.57
areas BASELINE+PROF 0.65 0.52 0.65 0.59		LINEAR	0.58	0.41	0.46	0.40
$(\mathbf{T}, \mathbf{n}, 10)$		BASELINE+PROF	0.65	0.52	0.65	0.59
(10p 10) BASELINE 0.20 0.23 0.35 0.25		BASELINE	0.20	0.23	0.35	0.25

BASELINE+NLP (r=0.36) through all SEASON. This fact suggests that the profile information is highly related to improve the performance in detection of influenza epidemics.

As described above, both NLP classification and profile information improved the performance to detect influenza epidemics. This result shows the combination of these techniques (EMNLP2011) achieved higher performance.

5.4.3 Contribution of Indirect Information in LINEAR Model (BASELINE+PROF+NLP vs. LINEAR+NLP)

To evaluate the contribution of indirect information in LINEAR model, we compare the performance of EMNLP2011 (BASELINE+PROF+NLP) and LINEAR+NLP. Although the performance of both methods is medium, the correlation coefficient of LINEAR+NLP (r=0.50 in SEASON-TOTAL) is lower than EMNLP2011 (r=0.69) through all SEASONs as shown in Table 5(a). This point indicates the difficulty of detecting influenza epidemics solely by adding indirect information in a naive manner. 5.4.4 Contribution of Indirect Information in TRAP

# Model (BASELINE+PROF+NLP vs. TRAP+NLP)

To evaluate the proposed model, the TRAP model, we compare the respective performances of TRAP+NLP and EMNLP2011 (BASELINE+PROF+NLP), which were better than LINEAR+NLP in Section 5.4.3.

Actually, TRAP+NLP shows the highest correlation coefficient among methods (underlined in Table 5), indicating that it achieved the best performance of influenza epidemic detection for the gold standard data, which in turn suggests that TRAP model methods effectively contribute to exploitation of both direct and indirect information from social sensors for detecting disease epidemics accurately.

# 6. **DISCUSSION**

# 6.1 After the Boom No One Tweets

The fact that TRAP model outperforms the LINEAR model indicates that if influenza becomes a hot topic, people do not talk about it, which suggests the aspect of human nature by which people become bored quickly with news. From a psychological viewpoint, similar phenomena were so far proposed. Most showed rapid propagation of rumors (especially bad news) and its short life [39–41]. Among various social media services, Twitter is an extremely "fast" media. Thereby, the life of news might be shorter than other existing news. In other words, people might hesitate to tweet an already-known fact.

This report is the first of a study that handles such human nature using a statistical model. This model has sufficient room for application to additional studies. For example, we simply regard the simulation of the referred tweet as news. Better methods using other media, such news web site information, are reasonable. The manner of estimation of the potential tweet users can also improve by consideration of more realistic data.

#### 6.2 Effectiveness of Each Module

From the experimentally obtained results presented in Section 5, we were able to observe three findings as described below.

- (1) Effectiveness of NLP (Section 6.2.1)
- (2) Effectiveness of location detection (Section 6.2.2)

(3) Effectiveness of aggregation by TRAP model (Section 6.2.3) As described in Section 1, one novelty of this study is high-resolution geographic analysis. In this section, we discuss the effectiveness for each prefecture. Fig. 7 portrays temporal changes of the gold standard data (light red bar) and results of TRAP+NLP (red line) and LINEAR+NLP (black dotted line) for three SEASONs in 47 prefectures, Japan.

# 6.2.1 Effectiveness of NLP-based Classification

We discuss the effectiveness of NLP-based classification by comparing the performance of the methods with NLP in Top-10 high population areas in Table 5(a) with that of the methods without NLP in Top-10 low population areas in Table 5(b). The rank of population of areas is presented in Fig. 8.

In urban areas such as Tokyo and Osaka, the TRAP model performance was sufficiently high. In fact, the correlation coefficient of TRAP was equal or higher than 0.7. Regarding other results, all correlation coefficient values were higher than 0.5, reflecting medium correlation.

However, in more rural areas such as Shimane and Toyama, no drastic improvement was observed using NLP. Especially, little difference in performance was found between BASELINE+NLP and BASELINE. However, we were able to say that NLP never worsened the performance motivating the use of NLP.

#### 6.2.2 Effectiveness of Location Detection

The proposed method uses location information of three types: not only GPS information and profile information used in previous studies but also referred location. We discussed the effects of exploiting the referred location (as indirect information) as well as GPS information and profile information (as direct information). From Table 5(a), we observed that the indirect information might be not so important in high population areas such as Tokyo and Osaka. For example, EMNLP2011 realized high correlation (r>0.7)

in urban areas on average. In such areas, even BASELINE+NLP only using GPS information has medium correlation.

In contrast, using indirect information is effective in rural areas. Although EMNLP2011 was determined just medium correlation ( $r \le 0.7$ ) through all SEASONs, TRAP+NLP showed high correlation in SEASON2012 and SEASON2014, as shown in Table 5(a). Results of SEASON2013 might be affected by the lack of tweet data, as shown in Table 2.

This result might be caused by a common pattern by which much direct information was available in urban areas. In contrast, because a sufficient amount of direct information is not available from rural areas, there is some lack of exploiting indirect information.

#### 6.2.3 Effectiveness of Aggregation by TRAP Model

We discuss the effectiveness of the TRAP model by comparing correlation coefficients of the Top-10 high population areas and the one in Top-10 low population areas in Table 5(a).

In urban areas, the performance of two methods related to the TRAP model (TRAP+NLP and TRAP) was the highest among the others. Correlation coefficients of two methods related to LINEAR model (LINEAR+NLP and LINEAR) were less than 0.7, except in SEASON2012. For example, as for TOKYO (AREA13) and OSAKA (AREA27) in Fig. 7, TRAP+NLP matches the gold standard data well. In contrast, LINEAR+NLP has some gaps. These results confirm the TRAP model effectiveness for tweets of urban areas.

In rural areas, the performance of the methods related to TRAP model (TRAP+NLP and TRAP) was also the highest. Most of the correlation coefficients were higher than 0.6. Especially, the performance of TRAP+NLP in rural areas was higher than that of LINEAR+NLP in urban areas on average. For example, for SHIMANE (AREA32) and TOYAMA (AREA18) in Fig. 7, the results of both TRAP+NLP and LINEAR+NLP in SEASON2012 match with the gold standard well. However, the results in other SEASONs partially have gaps. The results of LINEAR+NLP are affected by the small numbers of tweets. For such areas, we improve the performance by adjusting the weight parameters adequately.

Overall, we were confirmed the effectiveness of aggregation using the TRAP model, which does not use three location information in the same way, but which distinguishes referred location as indirect information and use in a different way.

# 6.3 Relation between Volume of Tweets and Performance

The relation between population and the detection performance presents an important finding. Fig. 8 presents the relation between each area's population (blue bar) and performance (lines). The population is the number of tweets. The performance is the correlation coefficient. This figure compares TRAP+NLP (red line) with EMNLP2011 (dotted black line).

Results show that the performance of TRAP+NLP was higher than that of EMNLP2011 in urban areas. Specifically, the top 17 high population areas (from TOKYO (AREA13) to IBARAKI (AREA8)) exhibited high correlation (r>0.7). In these areas, more than 400 tweets were emitted.

However, other areas have large variance of performance. Although both methods sometimes stagnate the same level performance, TRAP+NLP mostly outperforms EMNLP2011. In AOMORI (AREA2), NAGANO (AREA17), OITA (AREA44), NAGASAKI (AREA42) and YAMANASHI (AREA16), the TRAP model achieved higher performance (r>0.7) than EMNLP2011(r≤0.7). For example, one typical example is AOMORI of SEASON2012 and SEASON2013. The graph of AOMORI in Fig. 7 shows that TRAP+NLP was able to detect a high level of continuous epidemic in SEASON2013, indicating the effectiveness of the TRAP model. However, as described previously, sometimes it was unable to detect tweets after an epidemic, which remains as a subject of future work.

Although the TRAP model achieved higher performance than EMNLP2011, the performance was of a medium level ( $0.4 < r \le 0.7$ ) in NIIGATA (AREA15), FUKUI (AREA 20), TOCHIGI (AREA 9), MIE (AREA24), IWATE (AREA 3), KAGOSHIMA (AREA 46) and ten other areas. For example, the graph of FUKUI in Fig. 7 shows that TRAP+NLP was unable to detect the sequential influenza epidemics in SEASON2012. There were gaps in other SEASONs. Therefore, the average performance through all SEASONs was medium. In only one (KUMAMOTO (AREA 43)), TRAP exhibited poorer performance than EMNLP2011 in SEASON2013 (see KUMAMOTO in Fig. 8). One of the reasons is medical treatment failure in Kumamoto in the SEASON. Even that was a domestic news, but tons of news on the failure appeared in Twitter stream, causing the bias.

Results show strong advantages of TRAP+NLP in high population areas. More importantly, TRAP+NLP never shows worse performance, except in one area. These findings are expected to contribute to similar SNS-based surveillance.

# 6.4 Toward Improvement of Detection Performance and Realization of Epidemic Prediction

To improve the detection performance of disease epidemics, it is important to implement functions that enable consideration of various effects related to geographic relations among areas: adjacency (neighborhood or not), accessibility (easy to access or not), isolation (island or not) and so on.

This study was conducted to elucidate the current situation of disease epidemics. To predict the spread of disease, we must develop the proposed method through integration with various prediction models such as SIR and FUNNEL presented in a report of an earlier study [42]. Therefore, we can identify outbreaks of infectious diseases with high accuracy before a wider outbreak.

#### 7. CONCLUSIONS

This paper proposed a novel approach that uses the locations described. This system used not only direct information, but also indirect information that mention other places. We assumed a model by which the indirect information inhibits direct information. In this experiment performed for high-resolution areas (prefecture level), this proposed approach exhibited improved the detection performance not only in rural cities, but also in urban cities, which demonstrated the effectiveness of the proposed method consisting of a TRAP model and NLP classification. This report is the first of a study that assesses handling of human reactions to a new event. This model offers sufficient room for additional study.

Future work will examine worldwide influenza surveillance. Furthermore, we plan to apply this method to other epidemic surveillances and to establish a novel method by integrating various models for their accurate prediction.

#### 8. ACKNOWLEDGMENTS

This research was supported in part by JSPS KAKENHI Grant Numbers 16K16057, 16H01722, 15K00162.

# 9. REFERENCES

 Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: realtime event detection by social sensors. The 19th International Conference on World Wide Web (WWW) 2010. p. 851-860.

- [2] Earle PS, Bowden DC, Guy M. Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys – Italy. 2011;54(6):708-715. doi: 10.4401/ag-5364.
- [3] Earle P. Earthquake Twitter. Nat Geosci. 2010;3(4):221-2. doi: 10.1038/ngeo832.
- [4] Kagan V, Stevens A, Subrahmanian VS. Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. IEEE Intell. Syst. 2015;30(1):2-5.
- [5] Mahmood T, Iqbal T, Amin F, Lohanna W, Mustafa A. Mining Twitter Big Data to Predict 2013 Pakistan Election Winner. 2013 16th International Multi Topic Conference (INMIC). 2013:49-54.
- [6] Fink C, Bos N, Perrone A, Liu E, Kopecky J. Twitter, Public Opinion, and the 2011 Nigerian Presidential Election. 2013 ASE/IEEE International Conference on Social Computing (Socialcom). 2013:311-20. doi: 10.1109/SocialCom.2013.50.
- [7] Ranco G, Aleksovski D, Caldarelli G, Grcar M, Mozetic I. The Effects of Twitter Sentiment on Stock Price Returns. PLoS One. 2015;10(9). doi: ARTN e013844110.1371/journal.pone.0138441.
- [8] Wang S, Paul MJ, Dredze M. Social media as a sensor of air quality and public response in China. Journal of Medical Internet research. 2015;17(3):e22. doi: 10.2196/jmir.3875.
- [9] Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A largescale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc. 2014;21(6):1098-103. doi: 10.1136/amiajnl-2014-002711.
- [10] Sugumaran R, Voss J. Real-time spatiotemporal analysis of West Nile virus using Twitter data. Proceedings of the Third International Conference on Computing for Geospatial Research and Applications2012. p. 1-2.
- [11] Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. The American Journal of Tropical Medicine and Hygiene. 2012;86(1):39-45. doi: 10.4269/ajtmh.2012.11-0597.
- [12] Diaz-Aviles E, Stewart A, Velasco E, Denecke K, Nejdl W. Towards personalized learning to rank for epidemic intelligence based on social media streams. Proceedings of the 21st International Conference on World Wide Web 2012. p. 495-6.
- [13] Gomide J, Veloso A, Wagner Meira J, Almeida V, Benevenuto F, Ferraz F et al. Dengue surveillance based on a computational model of spatiotemporal locality of Twitter. ACM Web Science Conference (WebSci) 2011. p. 1-8.
- [14] Paul MJ, Dredze M, Broniatowski D. Twitter improves influenza forecasting. PLoS Curr. 2014; 6. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.
- [15] Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One. 2013;8(12):e83672. doi: 10.1371/journal.pone.0083672.
- [16] Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. EMNLP2011. pp. 1568-1576.
- [17] Kanhabua N, Romano S, Stewart A, Nejdl W. Supporting temporal analytics for health-related events in microblogs. Proceedings of the 21st ACM International Conference on Information and Knowledge Management 2012. pp. 2686-2688.
- [18] Lamb A, Paul M, Dredze M. Separating fact from fear: Tracking flu infections on Twitter. Annual Conference of the North American Chapter of the Association for Computational Linguistics 2013.
- [19] Parker J, Wei Y, Yates A, Frieder O, Goharian N. A framework for detecting public health trends with Twitter. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2013. pp. 556-563.
- [20] Pawelek KA, Oeldorf-Hirsch A, Rong L. Modeling the impact of Twitter on influenza epidemics. Math Biosci. Eng. 2014;11(6):1337-56. doi: 10.3934/mbe.2014.11.1337.
- [21] Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. Journal of Medical Internet Research. 2014;16(10):e236. doi: 10.2196/jmir.3416.
- [22] Gesualdo F, Stilo G, Agricola E, Gonfiantini MV, Pandolfi E, Velardi P et al. Influenza-like illness surveillance on Twitter through

automated learning of naive language. PLoS One. 2013;8(12):e82489. doi: 10.1371/journal.pone.0082489.

- [23] Kim EK, Seok JH, Oh JS, Lee HW, Kim KH. Use of hangeul Twitter to track and predict human influenza infection. PLoS One. 2013;8(7):e69305. doi: 10.1371/journal.pone.0069305.
- [24] Signorini A, Segre AM, Polgreen PM. 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. 2011;6(5):e19467. doi: 10.1371/journal.pone.0019467.
- [25] Morita M, Maskawa S, Aramaki E, editors. Comparison between Social Media and Search Activity as Online Human Sensors for Detection of Influenza. The Fifth International Symposium on Languages in Biology and Medicine (LBM 2013); 2013.
- [26] Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One. 2015;10(10). doi: ARTN e013970110.1371/journal.pone.0139701.
- [27] Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B. Twitter improves seasonal influenza prediction. SciTePress. 2012:61-70.
- [28] Culotta A, editor Towards detecting influenza epidemics by analyzing Twitter messages. The First Workshop on Social Media Analytics (SOMA); 2010.
- [29] Kanouchi S, Komachi M, Okazaki N, Aramaki E, Ishikawa H. Who caught a cold? – Identifying the subject of a symptom. Association for Computational Linguistics 2015. pp. 1660-1670.
- [30] Quincey D, Kostkova P. Early warning and outbreak detection using social networking websites: The potential of Twitter: Springer Berlin Heidelberg; 2009. pp. 21-24.
- [31] Doan S, Ohno-Machado L, Collier N. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. Healthcare Informatics, Imaging and Systems Biology (HISB); pp. 62-71, 2012.
- [32] Szomszor M, Kostkova P, Quincey Ed, editors. #Swineflu: Twitter Predicts Swine Flu Outbreak in 2009. eHealth 2010; 2009.
- [33] Han B, Cook P, Baldwin T, editors. Geolocation prediction in social media data by finding location indicative words. COLING, 2012.
- [34] Backstrom L, Sun E, Marlow C. Find me if you can: Improving geographical prediction with social and spatial proximity. The 19th International Conference on World Wide Web2010. p. 61-70.
- [35] Adams B, Janowicz. K, editors. On the geo-indicativeness of nongeoreferenced text. The International AAAI Conference on Web and Social Media (ICWSM) 2012.
- [36] Chandra S, Khan L, Muhaya FB, editors. Estimating Twitter User Location Using Social Interactions: A Content Based Approach. Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom); 2011.
- [37] Chang H-w, Lee D, Eltaher M, Lee J, editors. @phillies tweeting from philly? Predicting Twitter user locations with spatial word usage. the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012); 2012.
- [38] Kinsella S, Murdock V, O'Hare N, editors. "I'm Eating a Sandwich in Glasgow': Modeling locations with tweets. The Third International Workshop on Search and Mining User-generated Contents (SMUC11); 2011.
- [39] Singh A, Singh YN. Nonlinear Spread of Rumor and Inoculation Strategies in the Nodes with Degree Dependent Tie Strength in Complex Networks. Acta Phys. Pol. B. 2013;44(1):5-28. doi: 10.5506/APhysPolB.44.5.
- [40] Kesten H, Sidoravicius V. The spread of a rumor or infection in a moving population. Ann Probab. 2005;33(6):2402-62. doi: 10.1214/009117905000000413.
- [41] Ozturk P, Li HY, Sakamoto Y. Combating Rumor Spread on Social Media: The Effectiveness of Refutation and Warning. P Ann Hicss. 2015:2406-14. doi: 10.1109/Hicss.2015.288.
- [42] Matsubara Y, Sakurai Y, Panhuis WGv, Faloutsos C. FUNNEL: automatic mining of spatially coevolving epidemics. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp. 105-114, 2014.